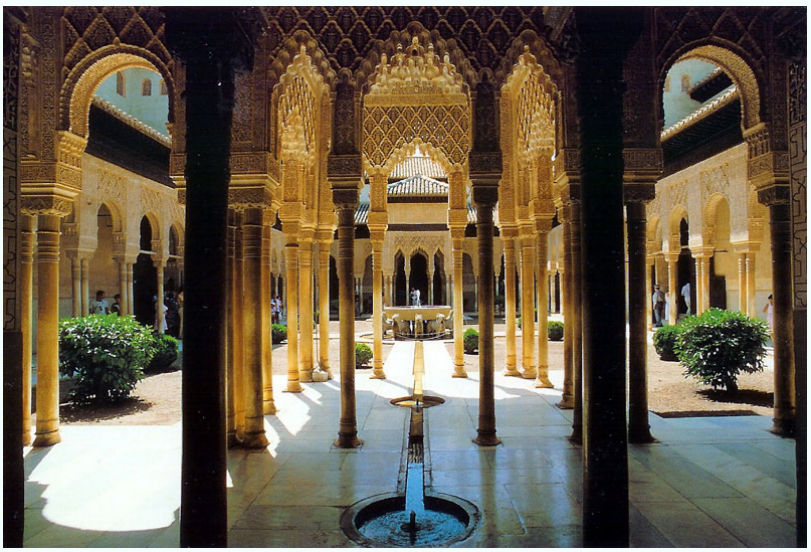
Terminology and Artificial Intelligence

TIA 2015

Granada, Spain

5-6 November, 2015
Proceedings of the Conference



Chairs:
Pamela Faber (University of Granada)
Thierry Poibeau (Lattice-CNRS)

Preface

Welcome to the 2015 edition of Terminology and Artificial Intelligence (TIA 2015). This year the conference will be held on November 4-6 in the historical city of Granada, Spain. It will be the first edition of TIA outside France.

We have received a wide range of submissions. Each submission was assigned to at least three reviewers and acceptance decisions were based on the reviews and the expertise of the programme committee. We finally accepted 16 long papers, 5 short papers and 4 posters and demonstrations. Authors of long and short papers also had the possibility to present their systems during a demo session.

This year we had the privilege of coordinating the conference. This has been a very enriching experience and we had the chance to work with a very efficient group of people who did an amazing job to make everything run smoothly. The local organization team was led by Pamela Faber and included Pilar León Araúz, Beatriz Sánchez Cárdenas, Arianne Reimerink, Silvia Montero Martínez, Miguel Vega Expósito, and Ana Belén Pelegrina Ortiz, all from the University of Granada.

Finally, we would like to thank the authors of submitted and accepted papers, and all the attendees to the conference, who will be the main actors from November 4 to November 6, 2015. We are convinced that we will experience a fantastic conference, scientifically exciting and full of fond memories, in the unique environment of Granada.

October 5, 2015

Pamela Faber
Thierry Poibeau

Table of Contents

Invited Papers

Long papers

Program Committee

Chairs

Pamela Faber	Universidad de Granada, Spain
Thierry Poibeau	Lattice-CNRS, France

Committee

Guadalupe Aguado de Cea	Universidad Politécnica de Madrid, Spain
Amparo Alcina	Universitat Jaume I, Spain
Sophia Ananiadou	University of Manchester, UK
Francisco Arcas Túnez	Universidad Catlica de Murcia, Spain
Nathalie Aussenac-Gilles	IRIT-CNRS, France
Caroline Barrière	Centre de recherche informatique de Montréal (CRIM), Canada
Núria Bel	Universitat Pompeu Fabra, Spain
Kevin Bretonnel Cohen	University of Colorado School of Medicine, USA
Paul Buitelaar	Insight Centre for Data Analytics, National University of Ireland, Ireland
Farid Cerbah	Dassault Aviation, France
Jean Charlet	AP-HP & INSERM UMRS 1142, France
Christian Chiarcos	Goethe Universität Frankfurt am Main, Germany
Philipp Cimiano	Bielefeld University, Germany
Nigel Collier	University of Cambridge, UK
Anne Condamines	CNRS and Université Toulouse Jean Jaurès, France
Béatrice Daille	Université de Nantes, France
Valérie Delavigne	Université Paris 3-Sorbonne nouvelle, France
Pascaline Dury	université Lumière Lyon 2, France
Fidelia Ibekwe-San Juan	Aix-Marseille Université, France
Marie-Christine Jaulent	INSERM UMRS 1142, France
Kyo Kageura	University of Tokyo, Japan
Olivia Kwong	The Chinese University of Hong Kong, China
Marie-Claude L'Homme	OLST, Université de Montréal, Canada
Alessandro Lenci	University of Pisa, Italy
Pilar León Araúz	University of Granada, Spain
Mercè Lorente Casafont	Universitat Pompeu Fabra, Spain
Maria J. Martin-Bautista	University of Granada, Spain
Silvia Montero-Martínez	University of Granada, Spain
Elena Montiel-Ponsoda	Universidad Politécnica de Madrid, Spain
Adeline Nazarenko	Université Paris 13 – Sorbonne Paris Cité, France
Carlos Periñán-Pascual	Universidad Politécnica de Valencia, Spain
Aurélie Picton	FTI-TIM, University of Geneva, Switzerland

Fabio Rinaldi	University of Zurich, Switzerland
Mari Carmen Suárez-Figueroa	Universidad Politécnica de Madrid, Spain
Sylvie Szulman	Université Paris 13, France
Pascale Sébillot	IRISA/INSA Rennes, France
Koichi Takeuchi	Okayama University, Japan
Rita Temmerman	Centre for Terminology, VUB, Belgium
Yannick Toussaint	LORIA, France
Mathieu Valette	ERTIM-INALCO, France
Lonneke van der Plas	University of Malta, Malta
Aline Villavicencio	Universidade Federal do Rio Grande do Sul, Brazil
Špela Vintar	University of Ljubljana, Slovenia
Sue Ellen Wright	Kent State University, USA
Pierre Zweigenbaum	LIMSI-CNRS, France

Invited Papers

Knowledge Extraction with Saffron: A Framework and Research Program

Paul Buitelaar
Unit for Natural Language Processing
Insight Centre for Data Analytics, National University of Ireland, Galway
IDA Business Park, Lower Dangan
Galway, Ireland
paul.buitelaar@insight-centre.org

Knowledge extraction from text is a longstanding challenge and ambition in Natural Language Processing and AI in general. Success and failure in this area depends however on definitions and use cases of 'knowledge', which I will address in this talk. Saffron[1] is a framework for knowledge extraction from text that has been developed over several years and was tested in a wide range of use cases. In the talk I will present the basic architecture and functionality of Saffron and its use in several applications. In the second part of the talk I will address some of the shortcomings of Saffron, which are the subject of our current research program.

[1] http://saffron.insight-centre.org/

Multilingualism and Conceptual Modelling

Ricardo Mairal
Facultad de Filologa
Dpto. de Filologas Modernas y sus Lingüísticas
Paseo de la Senda del Rey, 7
UNED
Madrid 28040
rmairal@flog.uned.es

One of the leading motivations behind the multilingual semantic web is to make resources accessible digitally in an online global multilingual context. Consequently, it is fundamental for knowledge bases to find a way to manage multilingualism and thus be equipped with those procedures for its conceptual modelling. In this context, the goal of this paper is to discuss how common-sense knowledge and cultural knowledge are modelled in a multilingual framework. More particularly, multilingualism and conceptual modelling are dealt with from the perspective of FunGramKB, a lexico-conceptual knowledge base for natural language understanding. This project argues for a clear division between the linguistic and the conceptual dimensions of knowledge. While the conceptual layer is organized into three modules, which result from a strong commitment towards capturing semantic knowledge (Ontology), procedural knowledge (Cognicon) and episodic knowledge (Onomasticon), the linguistic level includes a lexicon and a grammaticon (a syntactic repository of constructional schemata). Cultural mismatches are discussed and formally represented at both the conceptual and the linguistic levels of FunGramKB.

Long Papers

Using a distributional neighbourhood graph to enrich semantic frames in the field of the environment

Gabriel Bernier-Colborne **Marie-Claude L'Homme**
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (QC) Canada, H3C 3J7
{gabriel.bernier-colborne|mc.lhomme}@umontreal.ca

Abstract

This paper presents a semi-automatic method for identifying terms that evoke semantic frames (Fillmore, 1982). The method is tested as a means of identifying lexical units that can be added to existing frames or to new, related frames, using a large corpus on the environment. It is hypothesized that a method based on distributional semantics, which exploits the assumption that words that appear in similar contexts have similar meanings, can help unveil lexical units that evoke the same frame or related frames. The method employs a distributional neighbourhood graph, in which each word is connected to its nearest neighbours according to a distributional semantic model. Results show that most lexical units identified using this method can in fact be assigned to frames related to the field of the environment.

1 Introduction[1]

Recent work has shown that Frame Semantics (Fillmore, 1982; Fillmore and Baker, 2010) is an extremely useful framework to account for the lexical structure of specialized fields of knowledge (Dolbey et al., 2006; Faber et al., 2006; Schmidt, 2009; L'Homme et al., 2014). It is especially attractive in terminology since it provides an apparatus to connect linguistic properties of terms to a more abstract conceptual representation level.

[1]The work reported in this paper is carried out within a larger project entitled "Understanding the environment linguistically and textually", whose objective is to develop methods for characterizing the contents of texts on two different levels: 1. textual (using methods and techniques derived from corpus linguistics and text mining); and 2. linguistic (based on lexical semantic models).

Frame Semantics has proved especially useful to represent predicative units (verbs such as *deforest*, *recycle*, *warm*; predicative nouns such as *impact*, *pollution*, *salinization*; adjectives such as *clean*, *green*, *sustainable*), units that are often ignored in terminological resources. L'Homme et al. (2014) showed that the framework and more specifically the methodology devised within the FrameNet Project (Ruppenhofer et al., 2010) could be used to represent various lexico-semantic properties of predicative terms (in English and in French). L'Homme and Robichaud (2014) showed that frames could be connected via a series of relations and contribute to help us understand how terms are used to express environmental knowledge. However, as will be seen below, the work that led to the definition of frames and relations between frames mentioned above was done manually and turns out to be quite time-consuming. In this paper, we explore the potential of a semi-automatic, graph-based method to discover frame-relevant lexical units based on corpus evidence.

This paper is structured as follows. Section 2 explains how semantic frames help reveal part of the lexical structure of a specialized field of knowledge. Section 3 describes the graph-based method used to identify frame-relevant lexical units. Section 4 discusses how the model used in the manual evaluation of this method was selected. Section 5 presents the evaluation methodology and the results of the evaluation.

2 Frame Semantics applied to the field of the environment

In a specialized field such as the environment, many concepts correspond to processes, events

and properties which are typically expressed linguistically by predicative terms (verbs, predicative nouns and adjectives). However, traditional terminological models (and even less traditional ones, such as ontologies) are not properly equipped to describe the terms that denote these concepts and account for their specific linguistic properties, namely the fact that they require arguments (*X changes Y*; *impact of X on Y*). Frame Semantics (Fillmore, 1982; Fillmore and Baker, 2010) presents itself as a suitable alternative to these models since it is designed to connect linguistic properties to an abstract conceptual structure. In addition, it is well equipped to represent predicative lexical units and their argument structure.

2.1 Discovering frames in the field of the environment

L'Homme et al. (2014) describe a method to discover semantic frames based on an existing terminological resource called DiCoEnviro[2], that contains English and French terms related to the field of the environment. Each entry in DiCoEnviro is devoted to a lexical unit (LU), i.e. a lexical item that conveys a specific meaning, and states the argument structure of the LU, as in the following examples:

- warm$_{1a}$, vi: climate$_{[Patient]}$ warms
- warm$_{1b}$, vt: gas$_{[Agent]}$ or change$_{[Cause]}$ warms climate$_{[Patient]}$
- warm, adj.: warm climate$_{[Patient]}$

Argument structures state the number of obligatory participants, and two different systems are used to label them: the first one accounts for the semantic roles of arguments (**Agent**, **Patient**, **Cause**); the second one gives a typical term, i.e. a term that is representative of what can appear in that position.

Many entries – especially entries that describe predicative terms – come with annotated contexts that show how arguments[3] are realized in sentences extracted from an environmental corpus. For example, annotated contexts for *warm$_{1b}$* are shown in Table 1.

[2]See `http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi`.

[3]Non-obligatory participants are also annotated, as shown in the last sentence in Table 1, in which the phrase *since 1750*, which expresses Time, is annotated.

*The primary radiative effect of CO2 and water vapour$_{[CAUSE]}$ is to **WARM** the surface climate$_{[PATIENT]}$ but cool the stratosphere.*
*As increases in other greenhouse gases$_{[CAUSE]}$ **WARM** the atmosphere and surface$_{[PATIENT]}$, the amount of water vapour also increases, amplifying the initial warming effect of the other greenhouse gases.*
*The simulations of this assessment report (for example, Figure 5) indicate that the estimated net effect of these perturbations$_{[CAUSE]}$ is to **HAVE WARMED** the global climate$_{[PATIENT]}$ since 1750$_{[TIME]}$.*

Table 1: Annotated contexts for *warm$_{1b}$*

Argument structures and annotations were used to discover frames using two different methods. A semantic frame is a knowledge structure that represents specific situations (e.g. a teaching situation, a selling situation, a driving situation). A frame includes participants (called frame elements or FEs), some of which are obligatory (core FEs) and some of which are optional (non-core FEs). For instance, the Operate_vehicle frame describes a situation in which a Vehicle is set in motion by a Driver and includes the following core FEs: **Area**, **Driver**, **Goal**, **Path**, **Source**, and **Vehicle**. Lexical units such as *cycle*, *cruise*, *drive*, *pedal*, and *ride* evoke this frame (FrameNet, 2015). In this previous work, it was assumed that terms that share similarities with regard to their argument structures (number and semantic roles of arguments) and that share similarities with regard to the non-obligatory participants annotated in contexts are likely to evoke the same frame.

The first method consisted in comparing the argument structures and non-obligatory participants of terms already encoded in the terminological resource. This method shows that the verbs *cool$_{1a}$*, *warm$_{1a}$* and the nouns *cooling$_1$* and *warming$_1$* share many features. They all have a single argument (a **Patient**) and share some non-obligatory participants (**Degree**, **Duration**, **Location**).

The second method – which was applied only to the English terms – consisted in comparing the contents of the terminological resource to that of

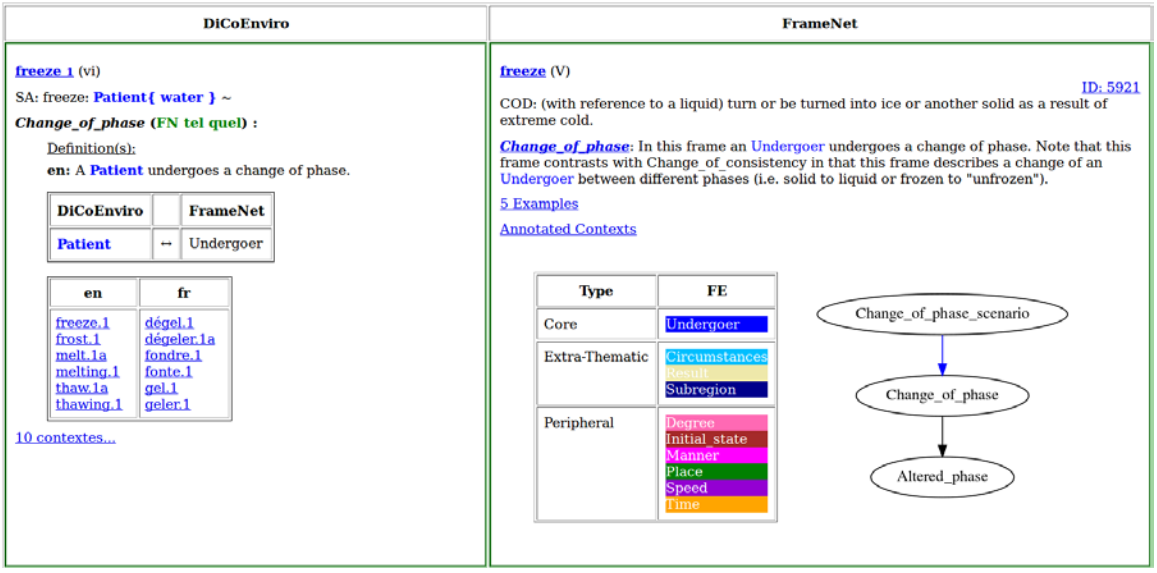


Figure 1: Comparison between the terminological database and FrameNet

FrameNet.[4] Relevant data were extracted from the FrameNet database for terms that were recorded in the terminological resource, as shown in Figure 1. This figure shows an example in which a correspondence between FrameNet and the terminological database could be established. However, in many instances, matches could not be made as nicely. In various cases, specific frames needed to be defined for the environmental terms (for instance, a new frame was created to capture adjectives such as *clean*, *environmental* and *green*, whose meaning can be loosely described as "that does not harm the environment"). In other cases, existing frames in FrameNet needed to be adapted to the data extracted from the terminological database for different reasons (slightly more specific meanings, different number of arguments, etc.).

2.2 A "framed" representation of the terminology of the environment

It soon became obvious that some of the frames identified based on the methods described in Section 2.1 could be linked. For instance, all processes related to changes affecting the environment appeared to be somehow related.

Again, using FrameNet (2015) as a reference, relations were established between some of the frames defined for environmental terms. Two relations not found in FrameNet (2015) were added (*Is opposed to* and *Is a property of*). This work led to the development of a resource called the Framed DiCoEnviro,[5] in which users can navigate through frames and relations between frames, and access the terms that evoke these frames along with their annotations. Figure 2 shows some of the relations identified between the frame Change_of_temperature (COT) (that contains verbs such as $cool_{1a}$, $warm_{1a}$ and the nouns $cooling_1$ and $warming_1$) and other frames.

3 Method for discovering related LUs

The methods described above allowed us to define a first subset of frames that are relevant for the field of the environment, link part of these frames and assign lexical units (LUs) to them. Based on this preliminary data, we explored the potential of a semi-automatic method to enrich our resource by adding new LUs to existing frames or discovering new frames. This method exploited distributional information obtained from a much larger corpus than the one used in the work described above.

The method we tested to discover related LUs is based on the neighbourhood graph induced by a distributional model of semantics. Distributional semantic models are commonly used to estimate semantic similarity, the underlying hypothesis be-

[4]The FrameNet team releases an XML version of the database (Baker and Hong, 2010).

[5]See `http://olst.ling.umontreal.ca/dicoenviro/framed/index.php` (in development).

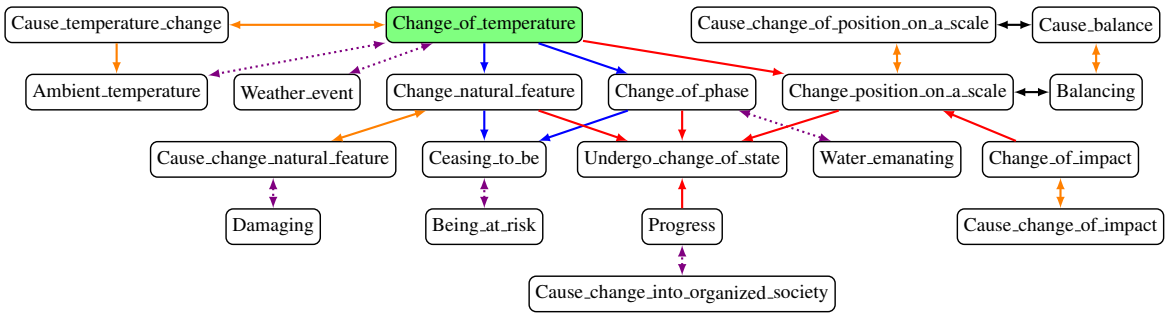


Figure 2: Change_of_temperature and related frames

ing that words that appear in similar contexts tend to be semantically related (Harris, 1954). The usual method of querying a distributional model is simply to compute, given a particular word, a sorted list of similar words. This method has several drawbacks, as has been pointed out recently by Gyllensten and Sahlgren (2015), who use a relative neighbourhood graph to query distributional models in a way that accounts for the fact that the query can have multiple senses. The method used here is similar in that it exploits a distributional neighbourhood graph. This allows us to take a list of terms and visualize their semantic neighbourhood, in order to identify related terms that can be encoded as frame-evoking LUs, either in existing frames or in new ones.

Various kinds of graphs could be used to compute and visualize the distributional neighbourhood of a particular word or set of words. We use a k-nearest-neighbour (k-NN) graph, two examples of which are the symmetric k-NN graph and the mutual k-NN graph (Maier et al., 2007). In a symmetric k-NN graph, two words w_i and w_j are connected if w_i is among the k nearest neighbours (NNs) of w_j or if w_j is among the k NNs of w_i. In a mutual k-NN graph, the two words are connected only if both conditions are true: w_i is among the k NNs of w_j and w_j is among the k NNs of w_i. In this work, we chose to use a mutual k-NN graph[6], the intuition behind this decision being that if two words are mutual NNs, there is a better chance that they actually do have similar meanings. This principle has been exploited elsewhere (Ferret, 2012; Claveau et al., 2014).

The graph construction procedure can be summarized as follows. Given a distributional semantic model, we compute the pairwise similarity between all words. For each word, we compute its k NNs by sorting all other words in decreasing order of similarity to that word and keeping the k most similar. Then, for each word w_i and each neighbour w_j in the k NNs of w_i, we add an edge in the graph between w_i and w_j if w_i is also among the k NNs of w_j. The resulting graph can be used to visualize the distributional neighbourhood of a term or set of terms.

4 Model selection

Any model that allows us to estimate the semantic similarity of two words can be used to build a semantic neighbourhood graph such as the one described in Section 3. We tested two different distributional semantic models for this purpose. Both models have several parameters which must be set and which can have a significant impact on the accuracy of the model in a given application. We therefore used an automatic evaluation procedure to tune the models' parameters and select a model for manual evaluation.

4.1 Corpus and reference data

The corpus used to build the models is the PANACEA Environment English monolingual corpus (Catalog Reference ELRA-W0063), a corpus containing 28071 web pages related to the environment (approximately 50 million tokens). The corpus was compiled automatically using a focused web crawler developed within the PANACEA project, and is freely distributed by ELDA for research purposes.[7] The corpus was

[6]We also tested the symmetric k-NN graph, but we only report results obtained with the mutual graph. We achieved higher F-scores using the mutual graph.

[7]See `http://catalog.elra.info/product_info.php?products_id=1184`.

converted from XML to raw text and lemmatized using TreeTagger (Schmid, 1994).

Reference data were extracted from the Framed DiCoEnviro.[8] The reference data are sets of LUs that evoke the same semantic frame. The list of English LUs was extracted from each of the frames included in the Change_of_temperature (COT) scenario[9] (cf. Figure 2). Two LUs (*thawing* and *thinning*) were excluded because they were not in the vocabulary used to construct the models, which contains the 10,000 most frequent lemmatized words in the corpus, excluding stop words. We obtained 13 sets containing a total of 53 LUs, each frame containing between 2 and 7 LUs. The number of unique LUs is 45, several LUs evoking more than one frame.[10]

4.2 Models tested

Two different distributional semantic models were tested. The first is a bag-of-words (BOW) model (Schütze, 1992; Lund et al., 1995), which is based on a word-word cooccurrence matrix computed using a sliding context window. The second is word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b), a neural language model that has been used in many NLP applications in the past few years. Word2vec (W2V) learns distributed word representations that can be used in the same way as BOW vectors to estimate semantic similarity.

The models' parameters were tuned by testing various combinations of parameter values, building neighbourhood graphs from each resulting model, and computing evaluation metrics on these graphs based on the reference data described in Section 4.1.

Some of the main choices that must be made when training a model using word2vec pertain to the architecture of the model (continuous skip-gram or continuous bag-of-words), the training algorithm (hierarchical softmax or negative sampling), the use of subsampling of frequent words, the size (dimensionality) of the word vectors and the size of the context window. We tested various values for each of these parameters, including the recommended values[11] when available. A total of 160 models were tested. In the case of the BOW model, important parameters[12] include the type, shape and size of the context window, the weighting scheme applied to the cooccurrence frequencies, and the use of dimensionality reduction. Again, we tested different values for these parameters. Each model was tested with and without dimensionality reduction, for which we used singular value decomposition (SVD). A total of 320 BOW models were built and evaluated (160 unreduced and 160 reduced using SVD).

For both models, we used the cosine similarity to estimate the similarity between words.

4.3 Evaluation metrics for model selection

For each model tested, we constructed multiple k-NN graphs, using different values of k. For each of these graphs, we computed evaluation metrics using the reference data described in Section 4.1. We used precision and recall to check to what extent LUs belonging to the same frame were connected in the graph. These metrics are computed for each of the 45 unique LUs in the reference data. Let w_i be an LU, $R(w_i)$ the set of related LUs that evoke at least one of the frames evoked by w_i, and $NN(w_i)$ the set of words that are adjacent to w_i in the graph. Furthermore, let TP_i (true positives) be the number of words in $NN(w_i)$ that are one of the related LUs in $R(w_i)$, FP_i (false positives) the number of words in $NN(w_i)$ that are not in $R(w_i)$ and FN_i (false negatives) the number of words in $R(w_i)$ that are not in $NN(w_i)$. The evaluation metrics are then calculated as usual:

$$\text{precision}_i = \frac{TP_i}{TP_i + FP_i}$$

$$\text{recall}_i = \frac{TP_i}{TP_i + FN_i}$$

$$\text{F-score}_i = \frac{2 \times \text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i}$$

[8] Data extracted on 2015-05-22. Data has been added since then, as the resource is in development.

[9] Frames related to the scenario only through a *See also* relation were excluded.

[10] Polysemous LUs evoke different frames. For instance, warm_{1a} (intransitive verb) evokes the Change_of_temperature frame; warm_{1b} (transitive verb) evokes the Cause_temperature_change frame; and warm_2 (adjective) evokes the Ambient_temperature frame.

[11] See `https://code.google.com/p/word2vec/#Performance`.

[12] Several studies have assessed the influence of this model's parameters. The relative importance of several parameters was quantified using analysis of variance by Lapesa et al. (2014).

The average precision, recall and F-score for a particular graph are then computed by taking the mean scores over all LUs in the reference data.

4.4 Results

Table 2 shows how precision, recall and F-score vary with respect to k. As the density of the graph increases, recall increases and precision decreases, the average F-score peaking around $k = 10$. The table also shows the number of nodes in the subgraph corresponding to the 45 LUs and their adjacent nodes in the graph. Based on these results, we selected 10 as an appropriate value of k.

k	nb nodes	precision	recall	F1
5	125	0.2120	0.2125	0.1915
10	206	0.1681	0.3005	**0.1971**
15	284	0.1429	0.3560	0.1858
20	359	0.1253	0.3999	0.1730
25	431	0.1108	0.4339	0.1594

Table 2: Evaluation metrics and number of nodes in the subgraph wrt k (averaged over all models)

Table 3 shows the average and maximum scores of each model (BOW, BOW reduced using SVD, and W2V) with $k = 10$. These results suggest that the BOW model performs best for this application.

Model	Avg prec. (max)	Avg rec. (max)	Avg F1 (max)
BOW	**0.1960** (**0.2775**)	**0.3153** (**0.4268**)	**0.2184** (**0.3016**)
SVD	0.1567 (0.2007)	0.2987 (0.3830)	0.1903 (0.2412)
W2V	0.1517 (0.2245)	0.2875 (0.4206)	0.1826 (0.2727)

Table 3: Evaluation metrics wrt model (with $k = 10$)

By analyzing how precision and recall varied with respect to the BOW model's parameters, we determined the optimal parameter values for this application. For example, the optimal window size was determined to be 3 words. The corresponding graph was then evaluated manually.

5 Evaluation

Once the model had been selected, the corresponding neighbourhood graph was evaluated manually. The evaluation was carried out by one of the co-authors of this paper, who is responsible for the development of the Framed DiCoEnviro. The 45 unique LUs in the reference data had 137 unique neighbours (adjacent nodes in the graph). These 137 words were evaluated manually in order to determine to what extent the graph can serve to discover frame-evoking LUs that can be added to the database.

The evaluation was carried out one frame at a time by observing the subgraph corresponding to that frame's LUs and their neighbours (adjacent nodes in the neighbourhood graph). For example, the subgraph for the frame Cause_change_of_impact is shown in Figure 3. In each subgraph, the LUs already encoded in that frame were highlighted in green, and those encoded in other frames in the COT scenario were highlighted in blue. One or more numbers were appended to the label of each LU to indicate which frame(s) it evokes.

For each word that was not already encoded as an LU in the COT scenario (i.e. for each white node), the evaluator was asked to choose one of the following categories:

1. The word should be encoded as an LU in the COT scenario
   (a) in an existing frame;
   (b) in a new frame.
2. The word should be encoded as an LU in another scenario
   (a) in an existing frame that is related to the COT scenario (by a *See also* relation);
   (b) in an existing frame that is not related to the COT scenario;
   (c) in a new frame.
3. The word should not be encoded as an LU in the database, but it is the realization of a core FE of one of the frames in the COT scenario.
4. The word should not be encoded as an LU in the database, nor is it the realization of a core FE of one of the frames in the COT scenario.

Table 4 shows the results of this evaluation. As these results show, most lexical items identified by the method (105 out of 137) can be encoded in a relevant frame in the field of the environment and

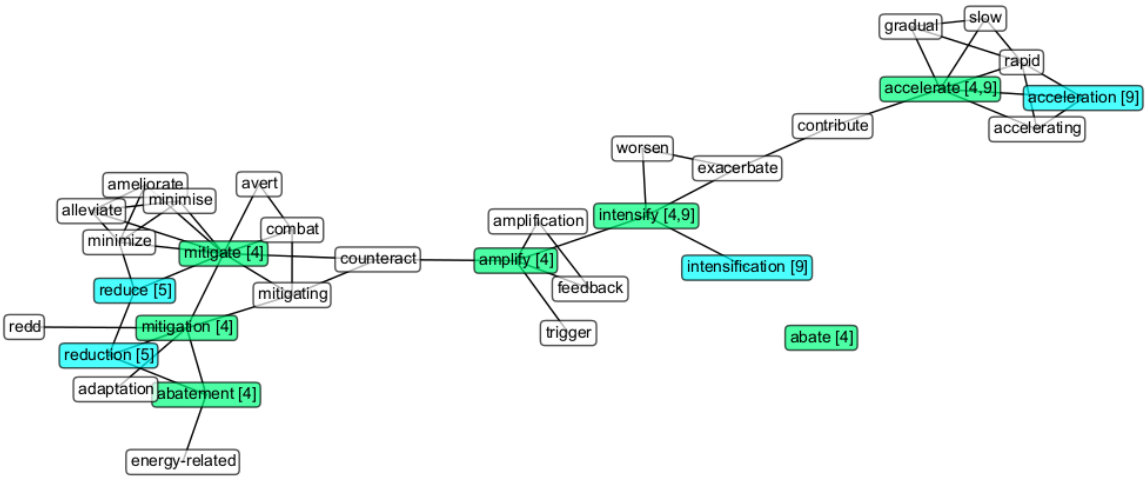


Figure 3: Subgraph for the frame Cause_change_of_impact

should be added to our resource. Among these, 88 would be frame-evoking LUs (categories 1 and 2) and 17 would be encoded as realizations of FEs (category 3). Interestingly, 48 lexical items are related to the COT scenario (categories 1a and 1b). The method allowed us to identify: 1. new frame-evoking LUs (such as *amplification*, *drop*, and *scarcity*) that had not been encoded in existing frames (category 1a); 2. LUs (such as *alteration* and *eliminate*) that evoke frames that had not been created (category 1b); and 3. variants (such as *cooler* for *cool* and *stabilise* for *stabilize*).

Category	Nb cases
1a	39
1b	9
1 (total)	**48**
2a	7
2b	3
2c	30
2 (total)	**40**
3	**17**
4	**32**
Total	**137**

Table 4: Summary of results.

The method also identified 40 items that would be encoded in environmentally relevant frames, but in a different scenario (category 2). It is worth pointing out that among these, 7 items correspond to LUs that would evoke a frame that is linked to the COT scenario (category 2a).

Finally, although 32 lexical items identified by the method would not be encoded in the resource and are thus considered false positives from the point of view of our application, further explanations are required. Some lexical items could evoke more general frames. For instance, *rapid* and *slow* would appear in the same frame if the general lexicon were considered. Other items identified are acronyms. *GW*, for instance, is the acronym for *global warming*. Technically, it could be defined as an LU evoking the COT frame, but multi-word terms and acronyms are not considered in the resource.

6 Concluding remarks

All in all the results obtained are quite interesting and show that the method can be used to assist lexicographers when defining frames and their lexical content, as the distributional neighbourhood of frame-evoking LUs often contain LUs that evoke the same frame or related frames. Distributional neighbourhood graphs provide information about the content of a specialized corpus that would be impossible to extract manually from such a large corpus. They are a very useful complement to other corpus tools, such as term extractors and concordancers, as they help lexicographers save time and locate relevant lexical units (near synonyms, variants) that they would otherwise miss.

In future work, we plan to integrate this methodology to assist lexicographers when defining new frames related to the field of the envi-

ronment. It could be particularly useful to obtain a view on corpora that deal with new or more specific topics and unveil the lexical units used to convey the knowledge related to these topics. It would also be interesting to test the potential of the method in other fields of knowledge. Extensions of this work could also involve using a graph-based clustering method to discover sets of lexical units that evoke the same frame without using existing frames.

Acknowledgments

This work was supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada.

References

Collin Baker and Jisup Hong. 2010. Release 1.5 of the FrameNet data. International Computer Science Institute. Berkeley.

Vincent Claveau, Ewa Kijak, and Olivier Ferret. 2014. Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. In *Actes de la 21e conférence sur le traitement automatique des langues naturelles (TALN)*, p. 220–231.

Andrew Dolbey, Michael Ellsworth, and Jan Scheffczyk. 2006. BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies. *Proceedings of KR-MED 2006: Biomedical Ontology in Action*, Baltimore, Maryland.

Pamela Faber *et al.* 2006. Process-oriented terminology management in the domain of Coastal Engineering. *Terminology* 12(2): 189–213.

Olivier Ferret. 2012. Combining bootstrapping and feature selection for improving a distributional thesaurus. In *Proceeding of the 20th European Conference on Artificial Intelligence (ECAI)*, p. 336–341.

Charles J. Fillmore. 1982. Frame Semantics. In *Linguistics in the Morning Calm*, p. 111–137. Seoul: Hanshin Publishing Co.

Charles J. Fillmore and Collin Baker. 2010. A frames approach to semantic analysis. In B. Heine and H. Narrog (ed.), *The Oxford Handbook of Linguistic Analysis*, p. 313–339. Oxford: Oxford University Press.

FrameNet. 2015. `https://framenet.icsi.berkeley.edu/fndrupal/`. Accessed: 2015-09-24.

Amaru Cuba Gyllensten and Magnus Sahlgren. 2015. Navigating the semantic horizon using relative neighborhood graphs. *CoRR*, abs/1501.02670.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2–3): 146–162.

Gabriella Lapesa, Stefan Evert, and Sabine Schulte im Walde. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, p. 160–170.

Marie-Claude L'Homme and Benoît Robichaud. 2014. Frames and terminology: Representing predicative units in the field of the environment. In Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex-IV), p. 186–197.

Marie-Claude L'Homme, Benoît Robichaud, and Carlos Subirats. 2014. Discovering frames in specialized domains. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 1364–1371.

Kevin Lund, Curt Burgess, and Ruth Ann Atchley. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, p. 660–665.

Markus Maier, Matthias Hein, and Ulrike Von Luxburg. 2007. Cluster identification in nearest-neighbor graphs. In *Algorithmic Learning Theory*, p. 196–210. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, p. 3111–3119.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended theory and practice. `http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf`. Accessed: 2015-09-24.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Thomas Schmidt. 2009. The Kicktionary: A Multilingual Lexical Resource of Football Language. In H.C. Boas (ed.), *Multilingual FrameNets in Computational Lexicography. Methods and Applications*, p. 101–134. Berlin/NewYork: Mouton de Gruyter.

Hinrich Schütze. 1992. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing (Supercomputing'92)*, p. 787–796.

Ontologies for terminological purposes: the EndoTerm project

Sara Carvalho
NOVA CLUNL
Department of Linguistics
Faculty of Social Sciences and Humanities
Universidade NOVA de Lisboa
Avenida de Berna, 26-C
1069-061 Lisboa Portugal
and
School of Technology and Management
University of Aveiro
R. Com. Pinho e Freitas, 28
3750-127 Águeda Portugal
sara.carvalho@ua.pt

Christophe Roche
Condillac Research Group
LISTIC
Université de Savoie
Mont Blanc
Campus Scientifique
73376 Le Bourget du Lac
France
christophe.roche@univ-savoie.fr

Rute Costa
NOVA CLUNL
Department of Linguistics
Faculty of Social Sciences and Humanities
Universidade NOVA de Lisboa
Avenida de Berna, 26-C
1069-061 Lisboa Portugal
rute.costa@fcsh.unl.pt

Abstract

In today's digital society, characterized by the Semantic Web and by Linked Data, ontologies, in the sense of Knowledge Engineering, have paved the way for new perspectives for Terminology, namely in what concerns the operationalization of terminological products. The collaborative work involving Terminology and ontologies has led to the emergence of new theoretical perspectives, one of which being Onto-terminology. This approach aims to reconcile Terminology's linguistic and conceptual dimensions whilst maintaining their fundamental differences and, in addition, enables the construction of a computer-readable representation of a given conceptualization. Bearing this in mind, this paper presents the EndoTerm project, a multilingual resource within the medical domain – with <Endometriosis> as the core concept – that comprises both verbal and non-verbal representations and that can be computationally represented and manipulated. The presentation of micro-concept systems based on these verbal and non-verbal representations will support a reflection upon the role of the latter in terminology work.

1 Introduction

Today's digital society has paved the way for new perspectives and opportunities for Terminology. In a context characterized by the Semantic Web[1] and by Linked Data[2], the need for the operationalization of terminologies, i.e. a computational representation of their concept system, has become increasingly important. In this respect, ontologies, in the sense of Knowledge Engineering (KE) – "a formal, explicit specification of a shared conceptualization"[3] –, constitute, according to Roche (2015: 129), "one of the most promising paths towards operationalizing terminologies". Granting a key status to ontology in terminology work implies, nevertheless, rethinking Terminology's theoretical and methodological principles and acknowledging the existence of a double dimension – linguistic and conceptual – that may enhance Terminology's role as a scientific discipline in its own right.

In the recent years, this joint work involving Terminology and ontology has led to the development of numerous resources in various areas of knowledge, one of them being Medicine. The current challenges concerning the way medical information and knowledge are produced, used, stored and shared require efficient and reliable

[1] Berners-Lee *et al*. (2001); Shadbolt *et al*. (2006).
[2] Berners-Lee (2006); Bizer *et al*. (2009)

[3] Even though Gruber's definition – "explicit specification of a conceptualization" (1993: 199) – prevails in the literature as the most widely quoted, for the purpose of this paper, ontology in the sense of KE will be regarded bearing in mind Studer *et al*.'s proposal (1998) quoted above, as it introduces three critical features: the fact that this specification should be explicit, i.e. the type of concepts used and the constraints on their use are explicitly defined, and formal, i.e. machine-readable; and that the conceptualization should be shared, i.e. an ontology should capture knowledge that is consensual among a given community. These authors have merged Gruber's definition and the one put forward by Borst (1997) – "a formal specification of a shared conceptualization". For further information, see Guarino *et al*. (2009)

solutions, in a society that demands immediate and multi-platform access to all digital content.

eHealth, defined by the World Health Organization (WHO) as "the cost-effective and secure use of information and communications technologies in support of health and health-related fields, including health-care services, health surveillance, health literature, and health education, knowledge and research"[4], has been considered a top priority by national and international institutions worldwide, with several action plans and programs focusing on expert collaboration, patient empowerment and interoperability[5].

In order to be achieved, these and other goals may greatly benefit from the input provided by an approach combining the operationalization potential of ontologies with Terminology's vital contribution to specialized knowledge as regards its representation, organization and dissemination.

In short, this paper aims to reflect on the role of ontologies in supporting the creation of concept systems for terminological purposes, particularly in the subject field of Medicine. Within Medicine, special attention will be given to Obstetrics and Gynecology, namely to the concept of <Endometriosis>[6], a chronic, inflammatory disease of gynecological nature that is yet relatively unknown, even among the expert community.

This paper will be structured as follows: section 2 will focus on the theoretical background, specifically in what concerns Terminology's double dimension perspective and the notion of Ontoterminology. Section 3 will be dedicated to the role of ontologies and/or terminological systems in the biomedical domain. Section 4 will provide a brief overview of the EndoTerm project, presenting a case study around the concept of <Laparoendoscopic single-site surgery>, a type of surgery currently being used within the context of endometriosis. Based on verbal and non-verbal representation, as well as on the input of subject field experts, a set of conceptual maps will be put forward. The final section will consist of some concluding remarks.

2 Terminology and ontology

2.1 Terminology's double dimension

This approach, which encompasses both a linguistic and conceptual dimension that are interrelated, has been more recently described by Roche (2012, 2015), Costa (2013) and Santos & Costa (2015). According to Roche (2015: 136), Terminology is "both a science of objects and a science of terms". For Costa (2013), it is precisely this double dimension, and the study of the relationship between one and the other that makes Terminology assume its role as an autonomous scientific subject.

This double dimension approach implies, therefore, that both the experts' conceptualization of a given subject field and the discourses produced by them must be taken into account. The cornerstone of this approach lies in the complementarity of these two fundamentally different dimensions. Understanding the relationship between the two dimensions is crucial in terminology work, as it will contribute to define a methodology that will not compromise the main goal of a terminological project as it is understood in this paper, which is to represent, organize and share the knowledge from a domain, based on the way it is conceptualized by a community of experts.

Consequently, it is believed that experts are indispensable to terminology work, working collaboratively with the terminologist in the different steps of the project, in order to identify the key concepts of the subject field, as well as the way they relate to each other and how they are represented (cf. Costa *et al*., 2012)

Nonetheless, and bearing in mind what was described in the introductory section, it is of paramount importance that the terminological products may, at some point, be operationalized, i.e. have a computational representation, and thus a more effective impact on the everyday life of the different target groups within the various subject fields.

The rising interest in the aforementioned conceptual and linguistic dimension, as well as in the subsequent synergies involving Terminology

[4] http://www.who.int/healthacademy/media/WHA58-28-en.pdf (30.07.2015)

[5] As an example, the successful implementation of interoperable Electronic Health Records (EHR) and ePrescription systems is one of the pivotal elements of the eHealth Action Plan 2012-2020, developed by the European Commission and available at: https://ec.europa.eu/digital-agenda/en/news/ehealth-action-plan-2012-2020-innovative-healthcare-21st-century (30.07.2015)

[6] Throughout this paper, concepts will be capitalized and written between single chevrons, whereas terms will be presented in lower case and between double quotation marks (Cf. Roche, 2015)

and ontologies has led to the emergence of new theoretical perspectives[7], one of which being Ontoterminology.

2.2 Ontoterminology: a new approach to Terminology?

Proposed by Roche *et al*. (2009), Ontoterminology aims to reconcile Terminology's linguistic and conceptual dimensions while maintaining their fundamental differences. Defined as a "terminology whose conceptual system is a formal ontology" (Roche *et al*., 2009: 325), this approach considers the conceptualization of a given subject field as the starting point of any terminological project, thus corroborating ISO 704's view that "producing a terminology requires an understanding of the conceptualization that underpins human knowledge in a subject area" (2009: 3).

As mentioned in 2.1, the expert plays an essential role throughout the process. However, Roche (2007) believes there may be risks inherent to the extraction of ontologies directly from texts, since very often, and due to inconsistencies, ellipses, metaphors and other phenomena, the lexical networks extracted from texts may not match the conceptual systems created with the help of the experts – hence, the discourse about knowledge should not be confused with knowledge itself: "Saying is not Modelling" (2007).

This is not to say, though, that natural language should be excluded from terminology work. In fact, "to conceptualize one must verbalize" (Roche, 2015: 149). Resorting to specialized texts is indeed relevant[8], although it must be taken into account that texts do not contain concepts *per se*, but the linguistic usages of the terms that designate them. All in all, specialized texts constitute an invaluable resource to the terminologist, especially in their first contact with a given domain, and the experts can – and should – play a critical role in advising the terminologist as to the texts that are deemed representative and/or mandatory in a given area.

Access to both linguistic and extra-linguistic knowledge is essential to any terminological project, provided the text selection is supported by rigorous criteria and the methodology/-ies used are consistent with the type of resource being created, its purpose(s), target group(s) and respective needs[9].

Instead of making them incompatible, the Ontoterminology approach aims to integrate the linguistic and the conceptual dimensions whilst preserving their core identities. This is visible in Roche's (2012) extension of the classical semantic triangle by Ogden and Richards (1923), called the "double semiotic triangle" (Figure 1).

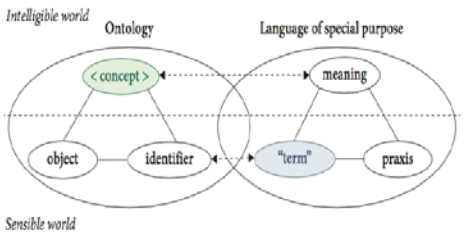


Figure 1: Double semiotic triangle (Roche, 2012)

In this diagram, it becomes clear that even though both dimensions are present in ontoterminological projects, they rely on two distinct semiotic systems and should therefore not be confused. By separating signified (meaning) and signifier (term) – related to Linguistics and natural language – from the concept and its name (identifier) – part of a formal system, Ontoterminology acknowledges a distinction between the definition of the term, written in natural language, and the definition of the concept, written in a formal language[10].

This distinction can be particularly important in subject fields where concepts can be both represented and defined in a non-verbal way[11]. Med-

[7] "Termontography" has been developed by the CVC in Brussels within the scope of the FF Poirot European Project and seeks to integrate ontologies in terminology work by combining Ontology Engineering, Terminography and Corpus Linguistics (Kerremans *et al*. (2004); Kerremans & Temmerman (2004); Temmerman & Kerremans (2003)). Despite the fact they do not share the same goals and are based on a different theoretical and methodological framework, comparing these approaches is not the purpose of this paper.

[8] And, in some cases, even indispensable, especially in the legal field, where texts are the pillar of expert knowledge and communication (see Costa *et al*. 2011, 2013).

[9] Santos & Costa (2015) advocate a mixed methodology in terminological work (onomasiological and semasiological), although they argue that the order "is not arbitrary" (p. 176). For knowledge representation purposes, a concept-based approach may constitute a more adequate starting point.

[10] The formal language supporting concept definitions should allow these to be objective (not depending on an individual interpretation), consistent and constructive (allowing the conceptualization to be computationally manipulated) (Roche, 2015).

[11] A more thorough analysis on the role of the non-verbal in terminology and knowledge representation may be found, for instance, in Galinski & Picht (1997); Picht (1999, 2011);

icine is one of such domains: Figure 2 depicts the female reproductive system of a woman suffering from endometriosis, and it includes the extent and location of the disease in terms of lesions and adhesions.

PERITONEUM			
Deep Endo	–	>3cm	· 6
CULDESAC			
Complete Obliteration			· 40
R. OVARY			
Deep Endo	–	1-3cm	· 16
Dense Adhesions	–	<1/3	· 4
L. TUBE			
Dense Adhesions	–	>2/3	· 16
L. OVARY			
Deep Endo	–	1-3cm	· 16
Dense Adhesions	–	>2/3	· 16
TOTAL POINTS			114

Figure 2: Stage-IV endometriosis (*Classification of Endometriosis* by the American Society for Reproductive Medicine, 1997)

Far from seeing a mere illustration, a subject field expert would immediately recognize a case of Stage-IV (severe) endometriosis. Rather than being regarded as signs from a Saussurean perspective, the terms that can be identified here ("peritoneum", "culdesac", "deep endo", "complete obliteration", "dense adhesions", etc.) should be perceived as signs in the sense of William of Ockham, for whom a sign is "tout ce qui, étant appréhendé, fait connaître quelque chose d'autre" (cf. 1988: 7)[12].

The potential of the ontoterminological approach, supported by the acknowledgement of Terminology's double dimension, provides an opportunity to make a contribution to the subject field of Medicine, in particular to <Endometriosis>, allowing the creation of EndoTerm, a multilingual resource that comprises both verbal and non-verbal representation and that can be computationally represented and manipulated[13].

3 Terminological resources in Medicine

As mentioned above, Medicine is currently undergoing significant changes in what concerns the production, use, storage and dissemination of medical information and, subsequently, medical knowledge. Nowadays, it is somewhat difficult to conceive – at least in some parts of the world – the practice of medicine without computerized medical records, prescriptions, examinations or even procedures, especially with the advent of robotic surgery.

Due to the increasing needs and challenges that have characterized this area over the last few decades, a new discipline has emerged, in the confluence of Information Science, Computer Science and Healthcare: Health Informatics has been defined as "the interdisciplinary study of the design, development, adoption and application of IT-based innovations in healthcare services delivery, management and planning[14]."

In order to facilitate the computer-based processing and exchange of medical or clinical information among all the stakeholders, that information is represented and organized via a number of terminological products, often grouped under the notion of "terminological system", with several typologies having been proposed throughout the years (see Table 1).

	Keizer *et al*. (2000)	ISO 17115 (2007)	EN 12264 (2005)	Duclos *et al*. (2014)
classification	✓	✓	✓	✓
coding system	✗	✓	✗	✓
coding scheme	✗	✓	✓	✗
nomenclature	✓	✗	✓	✓
ontology	✓	✗	✗	✓
taxonomy	✗	✗	✗	✓
terminology	✓	✓	✓	✓
thesaurus	✓	✗	✓	✓
vocabulary	✓	✗	✓	✓

Table 1: Typologies of terminological systems.

Used by the ISO/TC215 "Health Informatics", this umbrella term is characterized as a "set of designations within the domain of health care with, when appropriate, any associated rules, relationships and definitions" (ISO 1828: 2012). Albeit relevant, this definition does not fit the purposes of this paper and the project it aims to present, as it does not address the conceptual di-

Madsen (forthcoming); Roche (forthcoming); Prieto-Velasco (forthcoming).

[12] It is clear that, on the one hand, we do find terms in discourse that give rise to the construction of meaning – a *signifié* in the Saussurean sense, i.e. they acquire value in discourse. On the other hand, and as signs, terms also have the capacity to exist outside of discourse (Ockham's perspective), pointing towards the concept and thus providing access into the specialized domain.

[13] A first glimpse of which can be see in Section 4.

[14] U.S. National Library of Medicine. Available in: https://www.nlm.nih.gov/hsrinfo/informatics.html

mension of terminological resources and, hence, their ongoing evolution from "simple code-name-hierarchy arrangements, into rich, knowledge-based ontologies of medical concepts", as noted by Cimino (2001)[15].

Concept orientation has been presented in the literature as one of the key principles underlying the creation of today's (bio)medical terminological resources (see, for example, Chute *et al.* (1996); Coiera (2003); Duclos *et al*. (2014); etc.), and was, in fact, one of the twelve requirements, also known as desiderata, that Cimino (1998) believed should support all terminological systems within the medical context in the 21st century[16].

In recent years, many (bio)medical terminological resources have been designed or redesigned, in order to incorporate ontology-based elements, such as formal concept definitions, which, in turn, will enable both the operationalization and the aspired interoperability in this field. Yet each resource serves a specific purpose, which, in turn, determines their epistemological principles, core structure, the organization of the various concepts, as well as the language(s) of expression.

One of the initial stages of the EndoTerm project included extensive research of a set of representative (bio)medical resources (e.g. International Classification of Diseases (ICD), Medical Subject Headings (MeSH), Human Disease Ontology (DOID), Unified Medical Language System (UMLS)), to be used as a starting point in the creation of a thorough concept map of the domain in question. One of the following subsections will contain an example of one of these resources and its respective results concerning <Endometriosis>, namely the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT). Firstly, however, it is important to contextualize the concept of departure within our research project.

[15] It should be mentioned, though, that the boundaries among these different types of resources have become more and more blurred, in such a way that the term "ontology" is often being used indistinctly to refer to all of them. Grabar *et al.* (2012: 376-377) list several examples from the (bio)medical domain that illustrate "the lack of precise distinction among semantic resources in the literature".

[16] Check Cimino (1998, 2006) for further information on the Desiderata.

3.1 Endometriosis: facts and figures

Endometriosis is defined as "the presence of endometrial-like tissue outside the uterus, which induces a chronic, inflammatory reaction" (Kennedy *et al*., 2005). The exact prevalence of the disease is unknown, but it is believed to affect an estimated 176 million women of reproductive age worldwide (Adamson *et al.,* 2010). While its etiology is uncertain, it is likely to be multifactorial, including genetic, immunological, endocrinological and environmental influences.

Women with endometriosis typically have a range of pain-related symptoms, such as dysmenorrhea, dyspareunia, dyschezia, dysuria, non-cyclical pelvic pain, as well as chronic fatigue (Dunselman *et al*., 2014). A recent study conducted in 10 countries throughout the world has reported an overall diagnostic delay of 6.7 years (Nnoaham *et al*., 2011). Moreover, the World Endometriosis Research Foundation (WERF) EndoCost study (Simoens *et al*. (2012) has shown that the costs arising from women with endometriosis treated in referral centers are substantial (an average annual total cost per woman of €9579), an economic burden that is at least comparable to the costs of other chronic diseases, such as diabetes, Crohn's disease, or rheumatoid arthritis.

Surgical procedures play a key role in the diagnosis and treatment of the disease and are often depicted in the form of videos, which is why they were chosen as the focus of the case study to be presented in Section 4[17].

3.2 Endometriosis in SNOMED-CT

The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) is a comprehensive, multilingual healthcare terminology, resulting from the merge of the Systematized Nomenclature of Pathology (SNOP), published by the College of American Pathologists, and the Clinical Terms Version 3 (former Read Codes), designed by the UK's National Health Service[18]. When implemented in an application, and due to the Description Logic foundation of this tool, SNOMED-CT enables the representation of clin-

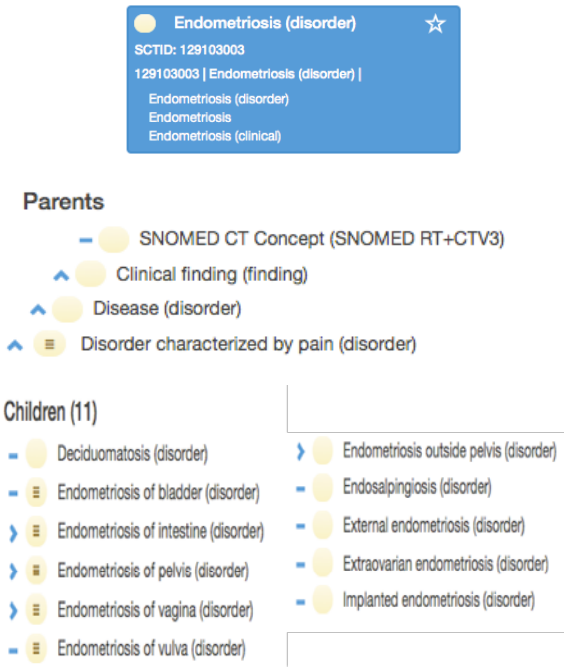
[17] Laparoscopy plus histology of resected endometriosis is, in fact, considered the "gold standard" in the diagnosis of this condition (Dunselman *et al*., 2014).

[18] It is currently owned and distributed by the International Health Terminology Standards Development Organisation (IHTSDO).

ical content in electronic health formats (e.g. EHR) in a consistent, reliable and computer-readable way[19].

The building blocks of this resource are the: i) **concepts**, representing clinical meanings and organized into hierarchies, ranging from general to specific (with 19 top-level concepts); ii) **descriptions**, which link appropriate human-readable terms to concepts; and iii) **relationships**, connecting concepts to other related concepts[20]. Each one of these three components has their own unique numeric identifier. Figure 3 illustrates the results obtained for <Endometriosis> in SNOMED-CT.

In the blue box on the top left corner, it can be seen that <Endometriosis (disorder)> is the concept name, and it coincides with the so-called Fully Specified Name (FSN), whereas "Endometriosis (clinical)" is the preferred synonym and "Endometriosis" the acceptable synonym. The Parents and Children elements refer to the "supertypes" and the "subtypes" of the concept in question, linked via |Is a| relationships.



[19]For more information, see: http://www.ihtsdo.org/snomed-ct; https://elearning.ihtsdotools.org

[20] The relationships in SNOMED-CT express defining characteristics of a concept and they can be divided into: a) subtype hierarchy relationships (Is a); or b) attribute relationships, which have a particular value provided by another concept, i.e. procedure concepts are linked, for instance, to certain sites.

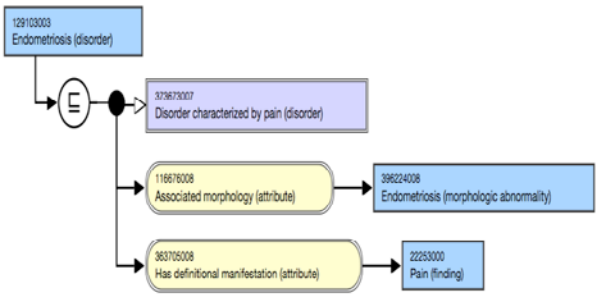


Figure 3: <Endometriosis> in SNOMED-CT (adapted from http://browser.ihtsdotools.org)

It should be noted that the subtype concepts are mainly related to the different organs or body parts where the disease can be located (e.g. bladder, intestine, etc.). There are further subdivisions in some of the Children that have not been included due to space constraints.

The final diagram represents the two types of concept relationships associated to <Endometriosis (disorder)>, distinguished by colours and types of arrows. The purple concept is an upper-level SNOMED-CT concept, linked to the initial concept by a |Is a| relationship. The yellow bubbles display an attribute relationship [has Associated morphology] and [has definitional manifestation] between the initial concept and <Endometriosis (morphological abnormality)> and <Pain (finding)>, respectively.

4 Terminology and Knowledge Organization

4.1 The EndoTerm Project

As previously mentioned, the EndoTerm project aims at the creation of a multilingual[21] terminological resource based around the concept of <Endometriosis>. This will be destined to future experts and to experts of other, related domains, mainly for training purposes. One of the objectives is to integrate the resource in an e-learning platform.

Since there are very few specialized texts about this disease in European Portuguese (EP), as most experts publish in English, it is believed a contribution could also be made to enrich the domain terminology in EP and, simultaneously, to improve linguistic quality criteria, which, in

[21] In English, European Portuguese and French. German might be included at a later stage of the project.

the future, might be applied to other projects involving information retrieval.

Although the inclusion of both verbal and non-verbal elements had already been foreseen in the project, due to the importance of the latter in this particular subject field, the group of experts that have been collaborating in this endeavour suggested the analysis of a type of resource that is becoming more and more important within the medical community: the video article[22].

By combining verbal (narration from the expert(s), slides with text, etc.) and non-verbal elements (2D or 3D images, animations, surgery footage), video articles constitute a noteworthy resource to take into account in the light of Terminology's double dimension. As a new type of scholarly communication that seems to be here to stay, its inclusion in a specialized corpus in a medical terminology project may become inevitable, which will, in turn, pose interesting theoretical and methodological challenges.

4.2 <LESS surgery>: the case study

The case study presented in this paper is based on a video article entitled "Single port laparoscopy"[23], which portrays a gynecological procedure – in this case, a hysterectomy, commonly seen as a last resort in cases of severe endometriosis – using a relatively recent type of surgery called single port laparoscopy.

The further study of the concept <Single port laparoscopy> pointed towards a lack of terminological consensus among the expert community. In fact, more than 20 acronyms used to designate this concept have been identified in the literature[24].

In order to solve this problem, a multidisciplinary medical consortium[25] gathered in 2008 and decided that the term "laparoendoscopic single-site surgery" (also known as LESS surgery) most accurately depicted the surgical procedure in question.

Based on information provided by textual sources, some of which cited below, by the aforementioned video article and others on the same topic, as well as by the feedback from two senior expert gynecologists who are also surgeons, a concept modeling proposal based on <LESS surgery> was created using a software environment for concept system building called OTe (Ontoterminology engine) Soft, supported by ontoterminological principles (see Section 2.2.).

Designed by the Condillac research team[26], this tool has a clear concept orientation, even though the user can also incorporate terms and, thus, the linguistic dimension. OTe Soft is structured around concepts, perceived as knowledge of a plurality of things that "help organize reality by grouping similar objects through what they have in common (Roche, 2015) (e.g. <Laparoscope>). One or more terms may be assigned to each concept, in various languages: i) natural (e.g. "laparoscope" (EN); "laparoscope" (FR); etc.; or ii) formal (e.g. programming language). In addition, a concept may be qualified by attributes, which have a given value, and be assigned one or more instances, also called "things", i.e. representations of elements in reality (Check Figure 6).

Concepts are linked to each other via concept relations: subsumption (is_a) (generic) and composition (part_of) (partitive) are presented by default. However, the tool allows the user to create new concept relations, as long as the logical principles are maintained (e.g. two concepts cannot be linked by the *instance of* relation)[27]. These relations are represented by different colours, in order to facilitate the graph's visual readability. The final "product" is called model, or semantic network, which can be exported in various formats (json, RDFS or OWL).

The following figures (4, 5 and 6) present examples of micro-concept maps built around the concept of <LESS surgery>: due to possible visual constraints, only partial views are shown here. The first micro-map (Figure 4) aims to po-

[22] For a more detailed description of this new type of resource, see Carvalho *et al.* (forthcoming).

[23] Available at: http://www.fertstert.org/article/S0015-0282%2812%2900387-1/fulltext

[24] See Box *et al*. (2008), Gill *et al*. (2010), Autorino *et al*. (2011), Rao *et al*. (2011), Sarkissian & Irwin (2013), Mori (2014), Naitoh (2014).

[25] Called the Laparoendoscopic Single-Site Surgery Consortium for Assessment and Research (LESSCAR), that published a consensus statement with the main conclusions of that meeting (Gill *et al*., 2010).

[26] www.condillac.org

[27] One of the challenges of creating a concept-modeling proposal lies, in fact, in defining other types of concept relations that do not fall under the generic or partitive categories. The ISO standards (1087-1:2000 and 704: 2009) lack diversity and systematisation, by classifying all the remaining relations as "non-hierarchical" (cf. Nuopponen 2011, 2014).

sition <LESS surgery> within the broader concept of <Surgery>.

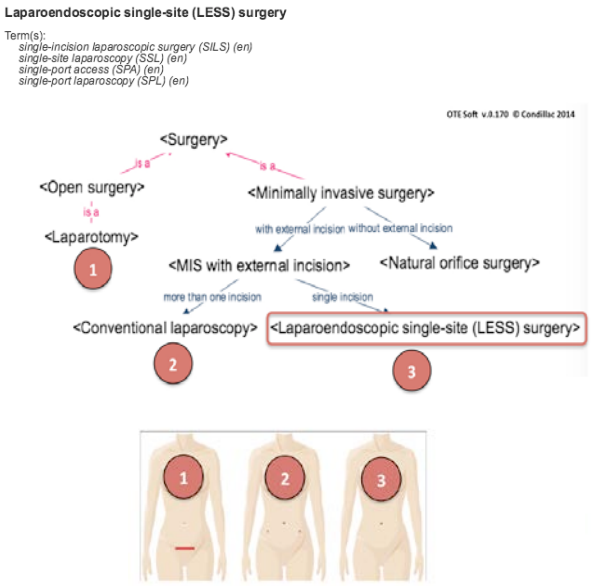


Figure 4: The concept of <LESS surgery>

There is a first subdivision presenting <Open surgery> and <Minimally invasive surgery> as subordinate concepts of <Surgery>. In the latter subtype, the subsequent hierarchy-based modeling was constructed through specific differentiation, bearing in mind the Aristotelian definition of *genus* + *differentia*[28]: i) with/without external incision; ii) with one incision/with more than one incision. Besides other advantages, such as the operationalization potential mentioned before, this concept modeling strategy constitutes a valuable starting point for the terminologist in the construction of natural language definitions.

On the upper left side, the linguistic dimension is also visible, and it includes the terms associated to the <LESS surgery> concept. In this case, it was decided to list some of the synonyms of the concept identified in the literature: Single-Incision Laparoscopic Surgery – SILS; Single-Site Laparoscopy (SSL); Single-Port Access (SPA); Single-Port Laparoscopy (SPL). Although the image does not show that, the user has the possibility of navigating through the concept network via concepts, terms, or relations. The three images in Figure 4 were added afterwards, as the current version of the OTe Soft tool does not yet allow the user to upload external resources (e.g. images, videos, diagrams, etc.)[29].

Figure 5 explores the types of umbilical incisions that may occur in a LESS surgery, being that the single incision in the umbilicus (navel) is regarded by the expert community as the essential characteristic of the concept, i.e. the characteristic which makes the concept what it is and constitutes its essence (cf. ISO 1087-1: 2000). In this figure, the metaphoric use attributed to the <Omega incision> should also be emphasized.

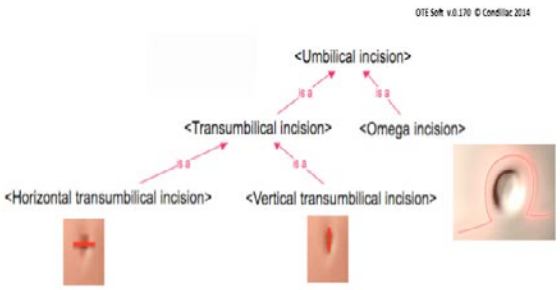


Figure 5: Types of <Umbilical incision>

Figure 6 contains a more detailed insight on the types of laparoscopes that exist, one of which - <Flexible video laparoscope> - is currently being used to perform LESS surgeries. In this case, the *EndoEYE* is presented as an instance of this concept.

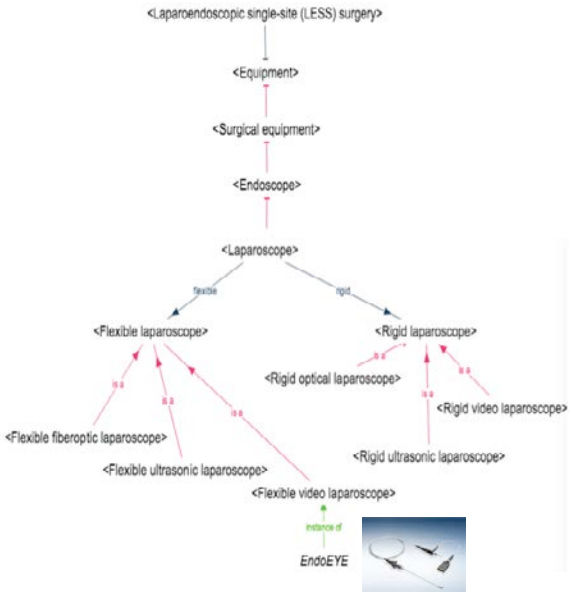


Figure 6: Types of <Laparoscope>

[28] These, along with other Aristotelian categories, are explored in Porphyry's *Isagoge* (2003).

[29] This also applies to the images in the remaining figures.

There is also a basic distinction within the expert community between flexible and rigid laparoscopes, as depicted by the map.

5 Concluding remarks

As a scientific discipline in the confluence of several others (e.g. logic, information science, cognitive sciences, linguistics, etc.), Terminology brings an unquestionable added value not only to the study of specialized language in various subject fields but, ultimately, also to the study of how knowledge is represented, organized and shared among a community of practice.

Nowadays, though, that contribution can only be further enhanced if the results of terminological work can be operationalized, i.e. represented in a computational format. Ontologies, in the sense of KE, represent a promising pathway that, however, must be based on collaborative work and on solid theoretical and methodological approaches. By acknowledging Terminology's linguistic and conceptual dimension and by applying that principle to the creation of tools, multidisciplinary teams integrating both "linguist-terminologists" and "engineer-terminologists" will be able to respond more effectively to the growing needs of expert communities – and, increasingly, of society as a whole.

Medicine is one of the fields where changes are more constant and substantial, and where terminological resources can play an even more vital role. Due to today's technological progress, it is likely that a sort of "multimedia knowledge base" may become a more and more common instrument in patient and expert education, which is why it is believed that the study and inclusion of **non-verbal elements** in these resources would represent an important qualitative leap in the joint research involving Terminology and Knowledge Engineering.

References

Adamson, G. *et al*. (2010). Creating solutions in endometriosis: global collaboration through the World Endometriosis Research Foundation. *Journal of Endometriosis*, *2*(1), 3–6.

ASRM. (1997). Revised American Society for Reproductive Medicine classification of endometriosis. *Fertility and Sterility*, *67*(5), 817–821.

Autorino, R. *et al*. (2011). LESS: an acronym searching for a home. *European Urology*, *60*(6), 1202–1204.

Berners-Lee, T. (2006). Linked Data. Retrieved April 20, 2014, from http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. *et al*. (2001). The Semantic Web. Retrieved March 14, 2014, from http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html

Bizer, C. *et al*. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, (Special Issue on Linked Data).

Box, G. *et al*. (2008). Nomenclature of natural orifice translumenal endoscopic surgery (NOTES) and laparoendoscopic single-site surgery (LESS) procedures in urology. *Journal of Endourology*, *22*(11), 2575–2581.

Carvalho, S. *et al*. Why read when you can watch? Video articles and knowledge representation within the medical domain. In *Proceedings of the 2015 TOTh Conference* (forthcoming).

CEN (2005). EN 12264: Health informatics. Categorial structures for systems of concepts. Brussels: European Committee for Standardization.

Chute, C. *et al*. (1996). The Content Coverage of Clinical Classifications. *Journal of the American Medical Informatics Association*, *3*(3), 224–233.

Cimino, J. J. (1998). Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. *Methods of Information in Medicine*, *37*(4-5), 394–403.

Cimino, J. J. (2001). Terminology tools: state of the art and practical lessons. *Methods of Information in Medicine*, *40*(4), 298–306.

Cimino, J. J. (2006). In defense of the Desiderata. *Journal of Biomedical Informatics*, *39*(3), 299–306.

Coiera, E. (2003). *Guide to Health Informatics* (2nd edition). New York: CRC Press.

Costa, R. (2013). Terminology and Specialised Lexicography: two complementary domains. *Lexicographica*, *29*(1), 29–42.

Costa, R. *et al*. (2011). L'organisation et la diffusion des connaissances terminologiques et textuelles au sein du Parlement portugais – le projet BDTT-AR. *Arena Romanistica. Journal of Romance Studies. Professional Communication and Terminology*, (7), 32 – 52.

Costa, R. *et al*. (2012). Mediation strategies between terminologists and experts. In *Proceedings of GLAT 2012 - Terminologies: textes, discours et accès aux savoirs spécialisés* (pp. 297–308). Genova.

Costa, R. *et al*. (2013). Methodology Design for Terminology in the Portuguese Parliament. In V. Jesenšek (Ed.), *Specialised Lexicography: Print and Digital, Specialised Dictionaries, Databases* (pp. 113–126). Berlin: Walter de Gruyter.

D'Ockham, G. (1988). *Somme de Logique*. (Traduit du latin par J. Biard). Trans-Europ-Repress.

Duclos, C. *et al*. (2014). Medical Vocabulary, Terminological Resources and Information Coding in the Health Domain. In A. Venot *et al*. (Eds.), *Medical Informatics, e-Health: Fundamentals and Applications* (pp. 11–41). Paris: Springer.

Dunselman, G. *et al*. (2014). ESHRE guideline: management of women with endometriosis. *Human Reproduction*, *29*(3), 400–412.

Galinski, C., & Picht, H. (1997). Graphic and Other Semiotic Forms of Knowledge Representation in Terminology Management. In S. E. Wright & G. Budin (Eds.), *Handbook of Terminology Management Volume 1: Basic Aspects of Terminology Management* (pp. 42–61). Amsterdam: John Benjamins Publishing Company.

Gill, I. *et al*. (2010). Consensus statement of the consortium for laparoendoscopic single-site surgery. *Surgical Endoscopy*, *24*(4), 762–768.

Grabar, N. *et al*. (2012). Ontologies and Terminologies: Continuum or Dichotomy? *Journal Applied Ontology*, IOS Press Amsterdam.

Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, *5*(2), 199–220.

Guarino, N. *et al*. (2009). What Is an Ontology? In S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (pp. 1–17). Heidelberg: Springer.

ISO (2007). ISO 17115: Health informatics -- Vocabulary for terminological systems. Geneva: International Organization for Standardization.

ISO (2009). ISO 704: Terminology work - Principles and methods. Geneva: International Organization for Standardization.

ISO (2012). ISO 1828: Health informatics -- Categorial structure for terminological systems of surgical procedures. Geneva: International Organization for Standardization.

Keizer, N. *et al*. (2000). Understanding Terminological Systems I: Terminology and Typology. *Methods of Information in Medicine*, *39*, 16–21.

Kennedy, S. *et al*. (2005). ESHRE guideline for the diagnosis and treatment of endometriosis. *Human Reproduction*, *20*(10), 2698–2704.

Kerremans, K., & Temmerman, R. (2004). Towards multilingual, termontological support in ontology engineering. In *Proceedings of Termino 2004. Université de Lyon* (p. n.a.). Lyon.

Kerremans, K. *et al*. (2004). Discussion on the Requirements for a Workbench supporting Termontography. In *Proceedings of the XI Euralex International Congress* (pp. 6–10).

Madsen, B. N. The Use of Linguistic and Non-linguistic Data in a Terminology and Knowledge Bank. In *Proceedings of the TOTh Workshop 2013: Verbal and nonverbal representation in terminology* (forthcoming).

Mori, T. (2014). Concept of Reduced Port Laparoscopic Surgery. In T. Mori & G. Dapri (Eds.), *Reduced Port Laparoscopic Surgery* (pp. 11–22). Tokyo: Springer.

Naitoh, T. (2014). Terminology. In T. Mori & G. Dapri (Eds.), *Reduced Port Laparoscopic Surgery* (pp. 23–26). Tokyo: Springer.

Nnoaham, K. *et al*. (2011). Impact of endometriosis on quality of life and work productivity: a multicenter study across ten countries. *Fertility and Sterility*, *96*(2), 366–373.

Nuopponen, A. (2011). Methods of concept analysis – tools for systematic concept analysis. *LSP Journal*, *2*(1), 4–15.

Nuopponen, A. (2014). Tangled Web of Concept Relations. Concept relations for ISO 1087-1 and ISO 704. In *Proceedings of Terminology and Knowledge Engineering 2014*. Berlin.

Picht, H. (1999). Einige Überlegungen zur nicht-sprachlichen Repräsentation von Gegenständen und Begriffen. *Synaps*, *3*, 1–50.

Picht, H. (2011). Non-verbal graphic representation of concepts. In *Ivanovo School of Lexicography: Traditions and Innovations. A Festschrift in Honour of Professor Olga Karpova* (pp. 220–236). Ivanovo: Ivanovo State University.

Porphyry. (2003). *Isagoge*. (Translated by J. Barnes). Oxford: Oxford University Press.

Prieto-Velasco, J. A. Depicting specialized concepts: strategies for the visualization of terminological knowledge. In *Proceedings of the TOTh Workshop 2013: Verbal and nonverbal representation in terminology* (forthcoming).

Rao, P. *et al*. (2011). Single-incision laparoscopic surgery - current status and controversies. *Journal of Minimal Access Surgery*, *7*(1), 6–16.

Roche, C. (2007). Saying Is Not Modelling. In *Proceedings of NLPCS 2007 (Natural Language Processing and Cognitive Science), Funchal, June 2007* (pp. 47–56). Funchal.

Roche, C. *et al*. (2009). Ontoterminology: A new paradigm for terminology. In *International Conference on Knowledge Engineering and Ontology Development, Oct 2009* (pp. 321–326). Funchal.

Roche, C. (2012). Ontoterminology: How to unify terminology and ontology into a single paradigm. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, May 23-25, 2012* (pp. 2626–2630). Istanbul: European Language Resources Association (ELRA).

Roche, C. Représentations formelles en terminologie. In *Proceedings of the TOTh Workshop 2013: Verbal and nonverbal representation in terminology* (forthcoming).

Roche, C. (2015). Ontological definition. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology - vol. 1* (pp. 128–152). Amsterdam: John Benjamins Publishing Company.

Santos, C., & Costa, R. (2015). Domain specificity: Semasiological and onomasiological knowledge representation. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology - vol. 1* (pp. 153–179). Amsterdam: John Benjamins Publishing Company.

Sarkissian, H., & Irwin, B. (2013). Overview: Rationale and Terminology. In A. Rane *et al*. (Eds.), *Scar-Less Surgery: NOTES, Transumbilical and Others* (pp. 13–24). London: Springer.

Shadbolt, N. *et al*. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*, (May/June), 96–101.

Simoens, S. *et al.* (2012). The burden of endometriosis: costs and quality of life of women with endometriosis and treated in referral centres. *Human Reproduction*, *27*(5), 1292–1299.

Studer, R. *et al.* (1998). Knowledge engineering: principles and methods. *Data & Knowledge Engineering - Special Jubilee Issue*, *25*(1-2), 161–197.

Temmerman, R., & Kerremans, K. (2003). Termontography: Ontology Building and the Sociocognitive Approach to Terminology Description. In *Proceedings of CIL17*. Matfyzpress.

Measuring the Relatedness between Documents in Comparable Corpora

Hernani Costa[a], **Gloria Corpas Pastor**[a] **and Ruslan Mitkov**[b]
[a]LEXYTRAD, University of Malaga, Spain
[b]RIILP, University of Wolverhampton, UK
{hercos,gcorpas}@uma.es, r.mitkov@wlv.ac.uk

Abstract

This paper aims at investigating the use of textual distributional similarity measures in the context of comparable corpora. We address the issue of measuring the relatedness between documents by extracting, measuring and ranking their common content. For this purpose, we designed and applied a methodology that exploits available natural language processing technology with statistical methods. Our findings showed that using a list of common entities and a simple, yet robust set of distributional similarity measures was enough to describe and assess the degree of relatedness between the documents. Moreover, our method has demonstrated high performance in the task of filtering out documents with a low level of relatedness. By a way of example, one of the measures got 100%, 100%, 95% and 90% precision when injected 5%, 10%, 15% and 20% of noise, respectively.

1 Introduction

Comparable corpora[1] can be considered an important resource for several research areas such as Natural Language Processing (NLP), terminology, language teaching, and automatic and assisted translation, amongst other related areas. Nevertheless, an inherent problem to those who deal with comparable corpora in a daily basis is the uncertainty about the data they are dealing with. Indeed, little work has been done on semi- or automatically characterising such linguistic resources and attempting a meaningful description of their content is often a perilous task (Corpas Pastor and Seghiri, 2009). Usually, a corpus is given a short description such as "casual speech transcripts" or "tourism specialised comparable corpus". Yet, such tags will be of little use to those users seeking for a representative and/or high quality domain-specific corpora. Apart from the usual description that comes along with the corpus, like number of documents, tokens, types, source(s), creation date, policies of usage, etc., nothing is said about how similar the documents are or how to retrieve the most related ones. As a result, most of the resources at our disposal are built and shared without deep analysis of their content, and those who use them blindly trust on the people's or research group's name behind their compilation process, without knowing nothing about the relatedness quality of the documents. Although some tasks require documents with a high degree of relatedness between each other, the literature is scarce on this matter.

Accordingly, this work explores this niche by taking advantage of several textual Distributional Similarity Measures (DSMs) presented in the literature. Firstly, we selected a specialised corpus about tourism and beauty domain that was manually compiled by researchers in the area of translation and interpreting studies. Then, we designed and applied a methodology that exploits available NLP technology with statistical methods to assess how the documents correlate with each other in the corpus. Our assumption is that the amount of information contained in a document can be evaluated via summing the amount of information contained in the member words. For

[1]I.e. corpora that include similar types of original texts in one or more language using the same design criteria (cf. (EAGLES, 1996; Corpas Pastor, 2001)).

this purpose, a list of common entities was used as a unit of measurement capable of identifying the amount of information shared between the documents. Our hypothesis is that this approach will allow us to: compute the relatedness between documents; describe and characterise the corpus itself; and to rank the documents by their degree of relatedness. In order to evaluate how the DSMs perform the task of ranking documents based on their similarity and filter out the unrelated ones, we introduced noisy documents, i.e. out-of-domain documents to the corpus in hand.

The remainder of the paper is structured as follows. Section 2 introduces some fundamental concepts related with DSMs, i.e. explains the theoretical foundations, related work and the DSMs exploited in this experiment. Then, Section 3 presents the corpora used in this work. After applying the methodology described in Section 4, Section 5 presents and discusses the obtained results in detail. Finally, Section 6 presents the final remarks and highlights our future work.

2 Distributional Similarity Measures

Information Retrieval (IR) (Singhal, 2001) is the task of locating specific information within a collection of documents or other natural language resources according to some request. This field is rich in statistical methods that use words and their (co-)occurrence to retrieve documents or sentences from large data sets. In simple words, these IR methods aim to find the most frequently used words and treat the rate of usage of each word in a given text as a quantitative attribute. Then, these words serve as features for a given statistical method. Following Harris' distributional hypothesis (Harris, 1970), which assumes that similar words tend to occur in similar contexts, these statistical methods are suitable, for instance to find similar sentences based on the words they contain (Costa et al., 2015) and automatically extract or validate semantic entities from corpora (Costa et al., 2010; Costa, 2010; Costa et al., 2011). To this end, it is assumed that the amount of information contained in a document could be evaluated by summing the amount of information contained in the document words. And, the amount of information conveyed by a word can be represented by means of the weight assigned to it (Salton and Buckley, 1988). Having this in mind, we took advantage of two IR measures commonly used in the literature, the Spearman's Rank Correlation Coefficient (SCC) and the Chi-Square (χ^2) to compute the similarity between documents written in the same language (see section 2.1 and 2.2). Both measures are particularly useful for this task because they are independent of text size (mostly because both use a list of the common entities), and they are language-independent.

The SCC distributional measure has been shown effective on determining similarity between sentences, documents and even on corpora of varying sizes (Kilgarriff, 2001; Costa et al., 2015; Costa, 2015). It is particularly useful, for instance to measure the textual similarity between documents because it is easy to compute and is independent of text size as it can directly compare ranked lists for large and small texts.

The χ^2 similarity measure has also shown its robustness and high performance. By way of example, χ^2 have been used to analyse the conversation component of the British National Corpus (Rayson et al., 1997), to compare both documents and corpora (Kilgarriff, 2001; Costa, 2015), and to identify topic related clusters in imperfect transcribed documents (Ibrahimov et al., 2002). It is a simple statistic measure that permits to assess if relationships between two variables in a sample are due to chance or the relationship is systematic.

Bearing this in mind, distributional similarity measures in general and SCC and χ^2 in particular have a wide range of applicabilities (Kilgarriff, 2001; Costa et al., 2015; Costa, 2015). Indeed, this work aims at proving that these simple, yet robust and high-performance measures allow to describe the relatedness between documents in specialised corpora and to rank them according to their similarity.

2.1 Spearman's Rank Correlation Coefficient (SCC)

In this work, the SCC is adopted and calculated as in Kilgarriff (2001). Firstly, a list of the common entities[2] L between two documents d_l and d_m is compiled, where $L_{d_l,d_m} \subseteq (d_l \cap d_m)$. It is possible to use the top n most common entities or all

[2]In this work, the term 'entity' refers to "single words", which can be a token, a lemma or a stem.

common entities between two documents, where n corresponds to the total number of common entities considered $|L|$, i.e. $\{n|n \in N^0, n \leq |L|\}$ – in this work we use all the common entities for each document pair, i.e. $n = |L|$. Then, for each document the list of common entities (e.g. L_{d_l} and L_{d_m}) is ranked by frequency in an ascending order ($R_{L_{d_l}}$ and $R_{L_{d_m}}$), where the entity with lowest frequency receives the numerical raking position 1 and the entity with highest frequency receives the numerical raking position n. Finally, for each common entity $\{e_1, ..., e_n\} \in L$, the difference in the rank orders for the entity in each document is computed, and then normalised as a sum of the square of these differences $\left(\sum_{i=1}^{n} s_i^2\right)$. The final SCC equation is presented in expression 1, where $\{SCC|SCC \in R, -1 \geq SCC \leq 1\}$.

$$SCC(d_l, d_m) = 1 - \frac{6 * \sum_{i=1}^{n} s_i^2}{n^3 - n} \quad (1)$$

2.2 Chi-Square (χ^2)

The Chi-square (χ^2) measure also uses a list of common entities (L). Similarly to SCC, it is also possible to use the top n most common entities or all common entities between two documents, and again, we use all the common entities for each document pair, i.e. $n = |L|$. The number of occurrences of a common entity in L that would be expected in each document is calculated from the frequency lists. If the size of the document d_l and d_m are N_l and N_m and the entity e_i has the following observed frequencies $O(e_i, d_l)$ and $O(e_i, d_m)$, then the expected values are $e_{i_{d_l}} = \frac{N_l * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$ and $e_{i_{d_m}} = \frac{N_m * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$. Equation 2 presents the χ^2 formula, where O is the observed frequency and E the expected frequency. The resulted χ^2 score should be interpreted as the interdocument distance between two documents. It is also important to mention that $\{\chi^2|\chi^2 \in R, 1 \geq \chi^2 < \infty\}$, which means that as more unrelated the common entities in L are, the lower the χ^2 score will be.

$$\chi^2(d_l, d_m) = \sum \frac{(O - E)^2}{E} \quad (2)$$

3 Corpora

INTELITERM[3] is a specialised comparable corpus composed of documents collected from the Internet. It was manually compiled by researchers with the purpose of building a representative corpus (Biber, 1988, p.246) for the Tourism and Beauty domain. It contains documents in four different languages (English, Spanish, Italian and German). Some of the texts are translations of each other (parallel), yet the majority is composed of original texts. The corpus is composed of several subcorpora, divided by the language and further for each language there are translated and original texts. For the purpose of this work, only original documents in English, Spanish and Italian were used, which for now on will be referred as int_en, int_es, int_it, respectively.

In order to analyse how the DSMs perform the task of ranking documents based on their similarity and filter out the unrelated ones, it is necessary to introduce noisy documents, i.e. out-of-domain documents to the various subcorpora. To do that, we chose the well-known Europarl[4] corpus (Koehn, 2005), a parallel corpus composed by proceedings of the European Parliament. As mentioned further in section 5.2, we added different amounts of noise to the various subcorpora, more precisely 5%, 10%, 15% and 20%. These noisy documents were randomly selected from the "one per day" Europarl v.7 for the three working languages: English, Spanish and Italian (eur_en, eur_es, eur_it, respectively).

	nDocs	types	tokens	$\frac{types}{tokens}$
int_en	151	11,6k	496,2k	0.023
eur_en	30	3.4k	29,8k	0.116
int_es	224	13,2k	207,3k	0.063
eur_es	44	5,6k	43,5k	0.129
int_it	150	19,9k	386,2k	0.052
eur_it	30	4,7k	29,6k	0.159

Table 1: Statistical information per subcorpora.

All the statistical information about both the INTELITERM subcorpora and the set of 20% of noisy documents, randomly selected for each working language, are presented in Table 1. In detail, this Table shows: the number of documents

[3] http://www.lexytrad.es/proyectos.html
[4] http://www.statmt.org/europarl/

(nDocs); the number of types (types); the number of tokens (tokens); and the ratio of types per tokens ($\frac{types}{tokens}$) per subcorpus. These values were obtained using the Antconc 3.4.3 (Anthony, 2014) software, a corpus analysis toolkit for concordancing and text analysis.

4 Methodology

This section describes the methodology employed to calculate and rank documents based on their similarity using Distributional Similarity Measures (DSMs). All the tools, libraries and frameworks used for the purpose in hand are also pointed out.

1) **Data Preprocessing**: firstly all the INTELITERM documents were processed with the OpenNLP[5] Sentence Detector and Tokeniser. Then, the annotation process was done with the TT4J[6] library, which is a Java wrapper around the popular TreeTagger (Schmid, 1995) – a tool specifically designed to annotate text with part-of-speech and lemma information. Regarding the stemming, we used the Porter stemmer algorithm provided by the Snowball[7] library. A method to remove punctuation and special characters within the words was also implemented. Finally, in order to get rid of the noise, a stopword list[8] was compiled to filter out the most frequent words in the corpus. Once a document is computed and the sentences are tokenised, lemmatised and stemmed, our system creates a new output file with all this new information, i.e. a new document containing: the original, the tokenised, the lemmatised and the stemmed text. Using the stopword list mentioned above a Boolean vector describing if the entity is a stopword or not is also added to the document. This way, the system will be able to use only the tokens, lemmas and stems that are not stopwords.

2) **Identifying the list of common entities between documents**: in order to identify a list of common entities (from now on we will use the acronym NCE), a co-occurrence matrix was built for each pair of documents. Only those that have at least one occurrence in both documents are considered. As required by the DSMs (see section 2), their frequency in both documents is also stored within this matrix ($L_{d_l,d_m} = \{e_i, (f(e_i, d_l), f(e_i, d_m)); e_j, (f(e_j, d_l), f(e_j, d_m)); ...; e_n, (f(e_n, d_l), f(e_n, d_m))\}$, where f represents the frequency of an entity in a document). With the purpose of analysing and comparing the performance of different DSMs, three different lists were created to be used as input features: the first one using the Number of Common Tokens (NCT), another using the Number of Common Lemmas (NCL) and the third one using the Number of Common Stems (NCS).

3) **Computing the similarity between documents**: the similarity between documents was calculated by applying three different DSMs ($DSMs = \{DSM_{NCE}, DSM_{SCC}, DSM_{\chi^2}\}$, where $_{NCE}$, $_{SCC}$ and $_{\chi^2}$ refer to Number of Common Entities, Spearman's Rank Correlation Coefficient and Chi-Square, respectively), each one calculated using three different input features (NCT, NCL and NCS).

4) **Computing the document final score**: the document final score $DSM(d_l)$ is the mean of the similarity scores of the document with all the documents in the collection of documents, i.e. $DSM(d_l) = \frac{\sum_{i=1}^{n-1} DSM_i(d_l,d_i)}{n-1}$, where n corresponds to the total number of documents in the collection and $DSM_i(d_l, d_i)$ the resulted similarity score between the document d_l with all the documents in the collection.

5) **Ranking documents**: finally, the documents were ranked in a descending order according to their DSMs scores (i.e. NCE, SCC or χ^2).

5 Results and Analysis

This experiment is divided into two parts. In the first part (section 5.1), we describe the corpus in hand by applying three different Distributional Similarity Measures (DSMs): the Number of Common Entities (NCE), the Spearman's Rank

[5] https://opennlp.apache.org

[6] http://reckart.github.io/tt4j/

[7] http://snowball.tartarus.org

[8] Freely available to download through the following URL https://github.com/hpcosta/stopwords.

Correlation Coefficient (SCC) and the Chi-Square (χ^2). As a input feature to the DSMs, three different lists of entities were used, i.e. the Number of Common Tokens (NCT), the Number of Common Lemmas (NCL) and the Number of Common Stems (NCS). By a way of example, Table 2 shows the NCT between documents, the SCC and the χ^2 scores and averages (av) along with the associated standard deviations (σ) per measure and subcorpus. Figure 1 presents the resulted average scores per document in a box plot format for all the combinations DSM vs. feature. Each box plot displays the full range of variation (from min to max), the likely range of variation (the interquartile range or IQR), the median, and the high maximums and low minimums (also know as outliers). It is important to mention that for the first part of this experiment (section 5.1) we did not use a sample, but instead the entire INTELITERM subcorpora in their original size and form, which means that all obtained results and made observations came from the entire population, in this case the English (int_en), Spanish (int_es) and Italian (int_it) subcorpora (for more details about the subcorpora see section 3). Regarding the second part of this experiment, we used the same subcorpora, but an additional percentage of documents was added to them in order to test how the DSMs perform the task of filtering out these noisy documents, i.e. out-of-domain documents (see 5.2). In detail, Figure 2 shows how the average scores decrease when injecting noisy documents and Table 3 presents how the DSMs performed when that noise was injected.

5.1 Describing the Corpus

The first observation we can make from Figure 1 is that the distributions between the features are quite similar (see for instance Figures 1a, 1d and 1g). This means that it is possible to achieve acceptable results only using raw words (i.e. tokens). Stems and lemmas require more processing power and time to be used as features – especially lemmas due to the part-of-speech tagger dependency and time consuming process implied. In general, we can say that the scores for each subcorpus are symmetric (roughly the same on each side when cut down the middle), which means that the data is normally distributed. There are some exception that we will discuss along this section. Another interesting observation is related with the high Number of Common Tokens (NCT) in English (int_en) when compared with Italian and Spanish (int_it and int_es, respectively), see Table 2 and Figure 1a. Later in this section, we will try to explain this phenomenon.

SubC.	Stats	NCT	SCC	χ^2
int_en	av	163.70	0.42	279.39
	σ	83.87	0.05	177.45
int_es	av	31.97	0.41	40.92
	σ	23.48	0.07	38.21
int_it	av	101.08	0.39	201.97
	σ	55.71	0.05	144.68

Table 2: Average and standard deviation of common tokens scores between documents per subcorpus.

Although the NCT per document on average is higher for the int_en subcorpus, the interquartile range (IQR) is larger than for the other subcorpora (see Table 2 and Figure 1a), which means that the middle 50% of the data is more distributed and thus the average of NCT per document is more variable. Moreover, longest whiskers (the lines extending vertically from the box) in Figure 1a also indicates variability outside the upper and lower quartiles. Therefore, we can say that int_en has a wide type of documents and consequently some of them are only roughly correlated to the rest of the subcorpus. Nevertheless, the data is skewed left and the longest whisker outside the upper quartile indicates that the majority of the data is strongly similar, i.e. the documents have a high degree of relatedness between each other. This idea can be sustained not only by the positive average SCC scores, but also by the set of outliers above the upper whisker in Figure 1b. The average of 0.42 SCC score and σ=0.05 also implies a strong correlation between the documents in the int_en subcorpus (Table 2). Likewise, the longest whisker and the set of outliers outside the upper quartile in the χ^2 scores also indicate a high relatedness between the documents.

Regarding the int_it subcorpus, the SCC and the χ^2 scores (Figures 1b and 1c) and the average of 101.08 common tokens per document and σ=55.71 (Figure 1a and Table 2) suggest that the data is normally distributed (Figure 1b) and highly

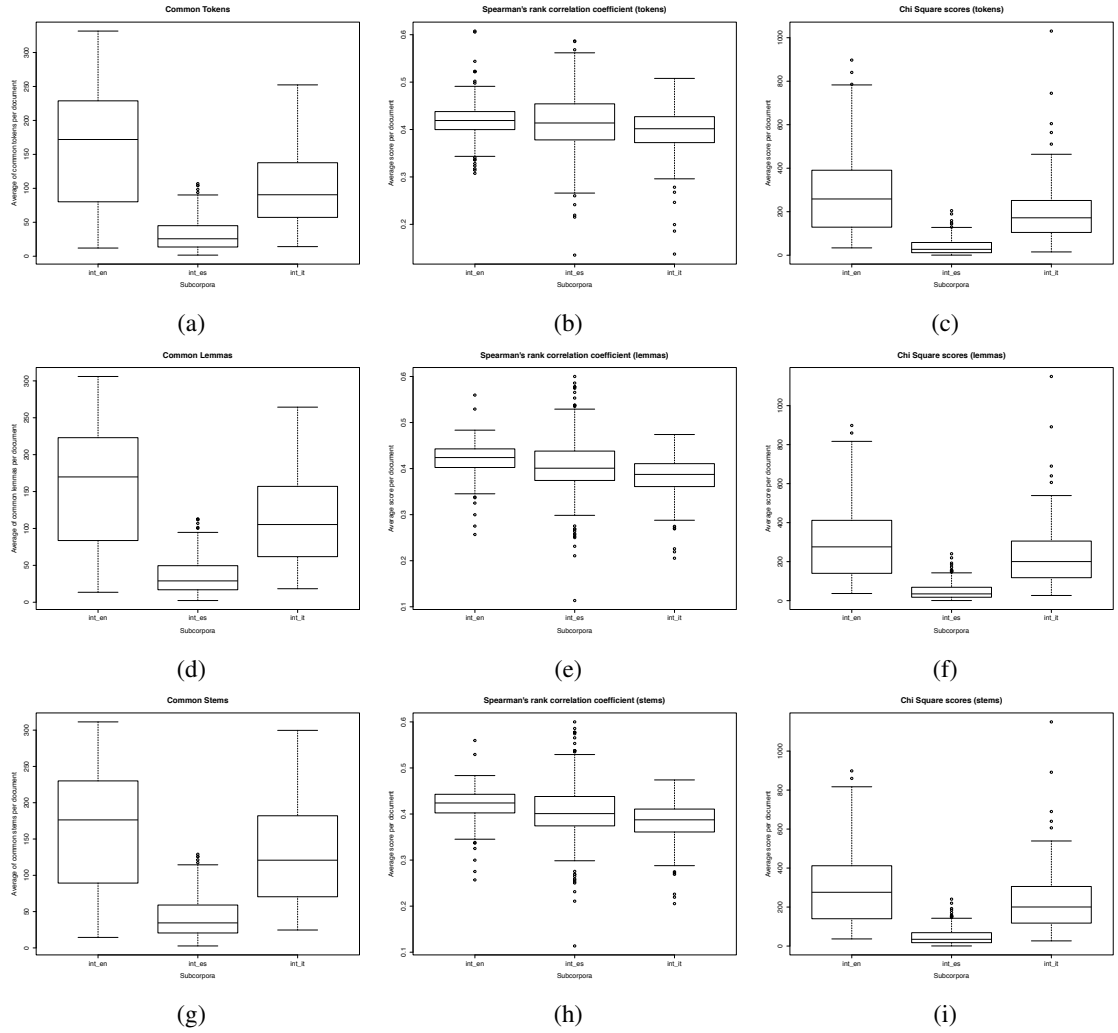


Figure 1: INTELITERM: average scores between documents per subcorpus.

correlated. Although this subcorpus got lower average scores for all the DSMs when compared to the English subcorpus, Table 2, Figure 1a, 1b and 1c show that the average scores and the range of variation are quite similar to the English subcorpus. Therefore, we can conclude that the documents inside the Italian subcorpus are highly related between each other.

From the three subcorpora, the int_es subcorpus is the biggest one with 224 documents (Table 1). Nevertheless, the average scores per document are slightly different from the other box plots (see Figures 1a, 1b and 1c). The χ^2 standard deviation practically equal to its average (38.21 and 40.92, respectively) and the SCC variability inside and outside the IQR indicates some inconsistency in the data. Moreover, Table 2 and Figure 1a reveal a lower NCT compared with int_en and the int_it subcorpora.

The subcorpus int_en has 163 common tokens per document on average with a σ=83, and the subcorpora int_it and int_es only have 101 and 31 common tokens per document on average with a σ=55 and σ=23, respectively (Table 2, NCT column). This means that the int_it and int_es subcorpora are composed of documents with a lower level of relatedness when compared with the English one. This fact could happen because Italian and Spanish have a richer morphology compared to English. Therefore, due to bigger number of inflection forms per lemma, there is a larger number of tokens and consequently less common tokens per document in Spanish. Another explanation could come from the fact

that the tourism and beauty services are more developed in Italy and Spain than in the UK and therefore there are more variety on the vocabulary used as well as in the services offered. Indeed, Table 1 offers some evidences about the employed vocabulary. The English subcorpus has a lower number of types and a higher number of tokens (11,6k and 496,2k, respectively) when compared with the Italian (19,9k types and 386,2k tokens) and Spanish subcorpora (13,2k types and 207,3k tokens). The high difference on the average of common tokens per document between Spanish and the other two languages can also be related with the marketing strategies used to advertise tourism and beauty services, which is somehow hard to confirm. Despite that our method is able to catch the lexical level of similarity between the documents, the semantic level is not taken into account, i.e. does not consider synonyms as similar words for example, and consequently would result on slightly different similarity scores (again, another explanation difficult to confirm).

To conclude, we can state from the statistical and theoretical evidences that the int_en and the int_it subcorpora look like they assemble highly correlated documents. We can not say the same for the int_es subcorpus. Due to the scarceness of evidences, we can only not reject the idea that this subcorpus is composed of similar documents. Nevertheless, as we will see in the next section, the fact that int_es is composed by low related documents (according to our findings) will affect the ranking task.

5.2 Measuring DSMs Performance

The second part of this experiment aims at assessing how the DSMs perform the task of filtering out documents with a low level of relatedness. To do that, we injected different sets of out-of-domain documents, randomly selected from the Europarl corpus to the original INTELITERM subcorpora. More precisely, we injected 5%, 10%, 15% and 20%[9] to the various subcorpora. As we can see in Figure 2, the more noisy documents are injected, the lower is the NCT. Then, the methodology described in Section 4 was applied to these "new twelve subcorpora" (int_en05, int_en10, ..., int_it15 and int_it20, see Figure 2). As a result, at this point we have the documents ranked in a descending order according to their DSMs scores.

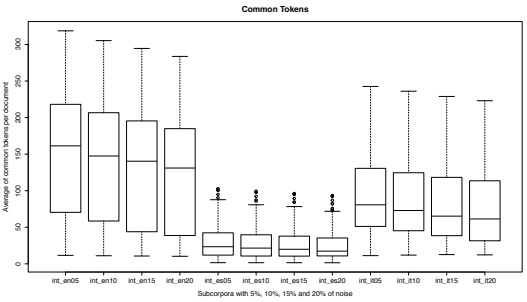


Figure 2: Average scores between documents when injecting 5%, 10%, 15% and 20% of noise to the various subcorpora.

In order to evaluate the DSMs precision, we analysed the first n positions in the ranking lists produced by the three DSMs (individually), and in this case n is the number of original documents in a given INTELITERM subcorpus. Table 3 presents the precision values obtained by the DSMs when injecting different amounts of noise to the various original subcorpora.

SubC	Noise	NCT	SCC	χ^2
int_en	*5%*	0.89	0.22	**1.00**
	10%	0.73	0.33	**1.00**
	15%	0.73	0.36	0.95
	20%	**0.80**	0.37	**0.90**
int_es	*5%*	0.00	0.00	0.38
	10%	0.07	0.07	0.20
	15%	0.09	0.09	0.17
	20%	0.14	0.18	0.23
int_it	*5%*	0.88	0.13	0.88
	10%	0.82	0.06	0.82
	15%	0.74	0.09	0.83
	20%	**0.73**	0.13	0.87

Table 3: DSMs precision when injecting different amounts of noise to the various subcorpora.

As expected, none of the DSMs got acceptable results for Spanish, being incapable of correctly identify noisy documents. However, we need to be aware that this happened due to the pre-existing low level of relatedness between the original documents in the int_es subcorpus (see Section 5.1 for more details). On the other hand, the DSMs show promising results for English and Italian. By

[9]The number of documents that correspond to these percentages can be inferred from Table 1.

a way of example, the χ^2 was capable of reaching 100% when injected 5% and 10% of noise to the int_en subcorpus, and even 90% when injected 20%. Although the NCT got lower precision, in general, when compared with the χ^2, it still reached 80% and 73% when injected 20% of noise to the English and to the Italian subcopora, respectively. From the evidences shown in Table 3, we can say that the NCT and the χ^2 are suitable for the task of filtering out low related documents with a high precision degree. The same cannot be say to the SCC measure, at least for this specific task.

6 Conclusions and Future Work

In this paper we presented a simple methodology and studied various Distributional Similarity Measures (DSMs) for the purpose of measuring the relatedness between documents in specialised comparable corpora. As input for these DSMs, we used three different input features (lists of common tokens, lemmas and stems). In the end, we conclude that for the data in hand these features had similar performance. In fact, our findings show that instead of using common lemmas or stems, which require external libraries, processing power and time, a simple list of common tokens was enough to describe our data. Moreover, we proved that it is possible to assess and describe comparable corpora through statistical methods. The number of entities shared by their documents, the average scores obtained with the SCC and the χ^2 measure resulted to be an important surgical toolbox to dissect and microscopically analyse comparable corpora.

Furthermore, these DSMs can be seen as a suitable tool to rank documents by their similarities. A handy feature to those who manually or semi-automatically compile corpora mined from the Internet and want to retrieve the most similar ones and filter out documents with a low level of relatedness. Our findings show promising results when filtering out noisy documents. Indeed, two of the measures got very high precision results, even when dealing with 20% of noise.

In the future, we intend not only to perform more experiments with these DSMs in other corpora and languages, but also test other DSMs, like Jaccard or Cosine and compare their performance.

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. The research reported in this work has also been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015); the R&D Project for Excellence TERMITUR (ref. nº HUM2754, 2014-2017); and the LATEST project (ref. 327197-FP7-PEOPLE-2012-IEF).

References

Laurence Anthony. 2014. AntConc (Version 3.4.3) Machintosh OS X. Waseda University. Tokyo, Japan. Available from `http://www.laurenceanthony.net`.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge, UK.

Gloria Corpas Pastor and Míriam Seghiri. 2009. Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In A. Beeby, P.R. Inés, and P. Sánchez-Gijón, editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Gloria Corpas Pastor. 2001. Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2010. The Impact of Distributional Metrics in the Quality of Relational Triples. In *19^{th} European Conf. on Artificial Intelligence, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ECAI'10, pages 23–29, Lisbon, Portugal, August.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2011. Using the Web to Validate Lexico-Semantic Relations. In *15^{th} Portuguese Conf. on Artificial Intelligence*, volume 7026 of *EPIA'11*, pages 597–609, Lisbon, Portugal, October. Springer.

Hernani Costa, Hanna Béchara, Shiva Taslimipoor, Rohit Gupta, Constantin Orasan, Gloria Corpas Pastor, and Ruslan Mitkov. 2015.

MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In *9th Int. Workshop on Semantic Evaluation*, SemEval'15, pages 96–101, Denver, Colorado, June. ACL.

Hernani Costa. 2010. Automatic Extraction and Validation of Lexical Ontologies from text. Master's thesis, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering, Coimbra, Portugal, September.

Hernani Costa. 2015. Assessing Comparable Corpora through Distributional Similarity Measures. In *EXPERT Scientific and Technological Workshop*, pages 23–32, Malaga, Spain, June.

EAGLES. 1996. Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P., May. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

Zelig Harris. 1970. Distributional Structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. D. Reidel Publishing Company, Dordrecht, Holland.

Oktay Ibrahimov, Ishwar Sethi, and Nevenka Dimitrova. 2002. The Performance Analysis of a Chi-square Similarity Measure for Topic Related Clustering of Noisy Transcripts. In *16th Int. Conf. on Pattern Recognition*, volume 4, pages 285–288. IEEE Computer Society.

Adam Kilgarriff. 2001. Comparing Corpora. *Int. Journal of Corpus Linguistics*, 6(1):97–133.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*.

Paul Rayson, Geoffrey Leech, and Mary Hodges. 1997. Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. *Int. Journal of Corpus Linguistics*, 2(1):133–152.

Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Amit Singhal. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.

A Combined Resource of Biomedical Terminology and its Statistics

Tilia Renate Ellendorff
Institute of Computational Linguistics
University of Zurich
tilia.ellendorff@uzh.ch

Adrian van der Lek
Institute of Computational Linguistics
University of Zurich
adrian.vanderlek@uzh.ch

Lenz Furrer
Institute of Computational Linguistics
University of Zurich
lenz.furrer@uzh.ch

Fabio Rinaldi
Institute of Computational Linguistics
University of Zurich
fabio.rinaldi@uzh.ch

Abstract

In this paper, we present a large biomedical term resource automatically compiled from the terminology of a selection of biomedical databases. The resource has a very simple and intuitive format and therefore can be easily embedded into a system for biomedical text mining and used as a linguistic resource. It is continuously updated and a user interface makes it possible to compile a new term resource according to individual requirements by selecting specific databases to be included. We present statistics for each included biomedical entity type separately as well as in the context of the combined terminology.

1 Introduction

Discovering entities such as genes, chemicals, diseases, species, etc. in the written text of research articles is an important part of biomedical text mining. This task is commonly known as "named entity recognition" (NER). Given a word from a text, it consists in deciding if this word is the name of an entity of interest. However, in the biomedical domain, the main focus of interest are not only the words in the text that refer to specific entities, but also under which database identifiers these entities are registered. The main purpose of using unique database identifiers is to provide unique conceptual referents for entities. Since biomedical entities tend to be highly ambiguous this disambiguation process is crucial. Furthermore, identifiers establish a reference to a database. This can be used either to retrieve additional information about the specific entity or to add related information to the database.

There are three major approaches for tackling the task of NER. A rule-based approach uses hand-crafted rules capturing structures of the word itself as well as its context. A machine learning approach extracts different lexical and contextual features from annotated corpora and applies a sophisticated statistical analysis. The third approach simply uses a dictionary look-up for discovering entity names which are already known and present in a given database.

Whereas the first and the second approach are possibly able to discover novel entity names which have never been seen in the literature before, the third approach is restricted to the dictionary, which means that only previously seen entity names can be discovered. The third approach, on the other hand, has the big advantage that by applying a dictionary look-up, a reference to a database can be established immediately.

In practice usually a combination of approaches is used. Combined approaches typically include a dictionary look-up as second step after using machine learning or rules for identifying entity mentions in the text. This second step provides database identifiers for the entity mentions discovered in the first step. Therefore, terminological resources are an important component of most systems for biomedical text mining.

There are a range of available biomedical databases containing terminological information for one or more entity types. Most of these databases are not designed to meet the needs of biomedical text mining and are typically available in very different formats. Therefore, when

building a text mining system, it can be time-consuming to extract their terminology and bring it into a format which can be easily used for term look-up.

Furthermore, different biomedical entity types have differences in their lexical properties. For instance genes tend to have a high degree of ambiguity. Chemicals tend to have many synonyms as their official name can be very long and complex, depicting different characteristics of the respective molecule. As a result of this, they are typically replaced by a shorter version in practice.

The aim of this paper is, on the one hand, to present a terminological resource in a very simple format which allows for easy integration as a dictionary in text mining systems, and, on the other hand, to analyze its term content regarding the lexical properties of the different terminologies that have been integrated, as well as the entity types in focus.

In the following sections of this paper we first give a very short introduction to the setting and purpose of biomedical text mining in general. Then we describe known properties of the different entity types. In the main part of the paper, we give a listing of a selection of databases containing terminological information about entities. We describe how we process these databases in order to transfer them into a simple and intuitive format. We consider the characteristics of the different entity types involved, namely genes and proteins, chemicals, diseases, species and cell lines. Finally, we present statistics of the terminology taken from each database on its own, as well as in the context of the whole combined term resource.

2 Biomedical Text Mining

It is essential to keep databases in the domain of life-sciences and biomedicine up to date in order to make knowledge easily accessible to researchers and support them in their daily work flow. New findings in the domain are typically published in the format of scientific articles. In the last decades the task of entering these findings into the database has still mainly relied on human manual work as an expensive and time-intensive process. Nowadays it becomes increasingly impossible for these specially trained curators alone to keep up with the increasing rate of publications in the domain. Automatizing this process, does not only help save money and time, but with the increasing rate of research and published papers it becomes the only way of coping with the huge inflow of information.

Biomedical text mining can be used to partially automate the process of biomedical literature curation by using computational power for discovering biomedical entities together with interaction and events in which they participate (Rinaldi et al., 2013). A successful biomedical text mining system is typically based on a pipeline which first discovers entities of interest in the text of a scientific article and subsequently looks for interactions between them. As described above, finding the unique database identifiers of the entities in focus is an important step in this process. A dictionary look-up considering all terms in an article is the only way for grounding them to their respective database identifiers. Which database identifiers are used in this process depends largely on the application for which a text mining system is built, or in other words, the database for which the system is designed to extract information.

3 Known Properties of Biomedical Entities

Compared to NER for other text genres, biomedical NER is known to be especially challenging. This is mainly due to the lexical properties of biomedical entities. The most prominent of these properties is the variability of entity names. Despite the existence of terminology recommendations by nomenclature organizations, authors are still free to use whatever variant of an entity name they prefer. Therefore, every author of the domain tends to use his or her own variant, often due to orthographic variations (Krauthammer and Nenadic, 2004). Furthermore, since many standard entity names are long and complicated, a very large range of abbreviations are continuously invented. Even though this is a characteristic shared by most entity types of the biomedical domain, each of them still has their own typical properties.

The biomedical entity type that has been most investigated in the past are genes and gene products, i.e. proteins. Usually, genes and proteins are treated as one group in biomedical NER, as their names are frequently used in place of each other. It is characteristic of genes/proteins to have a high level of ambiguity. One reason for this is that

genes are typically treated as different concepts, depending on the species for which the gene is described. Even though the concept is understood as different, often the same entity name is used. For example the gene *p53* (tumor suppressor gene) is found in a range of species, among them humans, mice and rats, and for all of them the same gene name is used. Another reason why gene names are so ambiguous is the fact that gene names can be very arbitrary, either because the discovering researcher chose an unusual name or because they are named according to their function, position on a gene or relation to other genes[1]. Furthermore, they can be built up of letters, numbers, punctuation, stopwords or non-alphabetical characters and frequently they are multi-word units (e.g. *daughters against decapentaplegic*, *short stop*, *cheap date*).

Chemicals have gained the reputation to belong to the most challenging entity names of the biomedical domain. They are highly heterogeneous as they can include generic names (e.g. *water*, *alcohol* or *cigarette smoke*), brand names (e.g. *Aspirin*), IUPAC (International Union of Pure and Applied Chemistry) names (*2-(Acetyloxy)benzoic Acid*) to name just a few (Rocktäschel et al., 2012). Chemical formulas are also used as entity names (e.g. *Al2(SO4)3*), some of which consist only of one letter. Apart from all these, partially highly ambiguous variants, authors introduce even more variants by using their own abbreviations, e.g. *NF-κB* instead of *NF-kappa-B inhibitor*. Furthermore, many standardized names, such as those from IUPAC, contain a lot of separating non-alphanumeric characters, such as slashes or commas, which can vary according to what the author prefers. On the other hand, if and where a name contains brackets, determines its meaning in many cases (Rocktäschel et al., 2012). For all these reasons, chemicals typically have a very large number of synonyms which can consist in completely different strings of letters, numbers and non-alphabetical characters, some of these as multi-word units. Because the synonyms have such different surface forms, usual normalization steps as for example fuzzy matching (checking for words similar to those in the dictionary), do not reach very good results for chemical entity names.

From all biomedical named entity types, disease names, as for example *HIV*, *Back Pain* or *Breast Cancer*, are possibly presenting the least difficulties to NER. In the past they have been shown to have less variability than named entities of genes and chemicals. This is mainly due to the reason that disease names are highly standardized throughout the literature (Jimeno-Yepes et al., 2008). One effect of this is that using a dictionary look-up on its own without further normalization can reach reasonable results as long as the dictionary is complete, containing all disease entity names of interest. Maybe even more than for chemicals and genes/proteins, it is very typical for disease names to consist in multi-word units.

In biomedical NER, names of organisms and species are usually treated together under the type "species". Species names can also have a high level of ambiguity, often the same species name is used to refer to several different entities (*C. elegans* can be used to refer to up to 41 different species in the NCBI taxonomy (Gerner et al., 2010)). Similar to gene entity names, species entity names can contain common English words which are part of the name and they share a range of acronyms with other entity types, like genes (Gerner et al., 2010). All of this is prone to introducing a high rate of false positives. Another characteristic that species names share with gene and chemical names is the high variability, introduced through the usage by the authors and sometimes even by misspellings.

An overview of cell line nomenclature has been given by Sarntivijai et al. (2008). Cell lines names share the characteristics of other biomedical entity names: there is no obligation to use standardized names and authors are free to use whatever variants they like. One more reason for ambiguity of cell lines is the scientific experimental setting: cell lines mutate or become subject to contamination, which also bring along a change of concept (Sarntivijai et al., 2008).

4 Combining Terminology from Different Databases

With the aim of building a terminological resource for biomedical text mining, we decided to focus on a selection of commonly used databases, each of which contains terminological information about one or more of the most typical biomedical entity types. In this section, we first describe the struc-

[1] http://www.curioustaxonomy.net/gene/fly.html

ture of the resulting combined terminological resource before we give details about the databases from which we take the terminological information, which files we use, how we process them and the general architecture that we apply.

4.1 Structure of the Combined Terminological Resource

The terminological resource which we compile using terminological data extracted from the selected databases is contained in one single file. This file has the very simple format of comma separated values (csv). The fieldnames, defining the contents of each of the six columns are the following: 'oid', 'resource', 'original_id', 'term', 'preferred_term', 'entity_type' (Table 1).

For each ID in each original database, an internal Base36 identifier 'oid' is generated. Base36 is a binary-to-text encoding scheme[2], using digits and all letters (in this case capital) of the English alphabet. A five digit sequence may e.g. encode over 60 million decimal values, while taking up only five bytes as opposed to eight. Synonyms are assigned the same oid as the main term. As such, the oid is not a unique identifier. Hence, the primary row key of the output file is a combination of the oid and the (synonymous) term.

The contents of the term field are matched in the text by the dictionary look up. The preferred term is the most standard term for a concept, which is preferred over other term variants.

Finally, the entity type field contains the type of entity, in this case normalized to the following entity types: gene/protein, chemical, disease, species and cell line.

By restricting the database to these fields, we focused on the most important information from the selected databases. The intention is to exclude any redundant information from the term resource and keep it as lightweight as possible by only focusing on information that is absolutely necessary for its application in a biomedical text mining system.

4.2 Included Resources by Entity Types

We decided to include terminology from the databases described in the current section, sorted by entity types. This selection is only made for illustrative purpose and to cover a sample of the most commonly used databases. However, as mentioned before, the format of the terminological resource is flexible and allows for easy integration of terminology from further databases.

Genes and Proteins

NCBI Gene NCBI Gene (Brown et al., 2015) ("Entrez Gene") is the gene database of the National Center Biotechnology Information (NCBI)[3]. It contains gene data from a wide range of species. Entrez Gene uses its own unique gene identifiers to track gene records. NCBI provides a downloadable file[4] which is updated on a daily basis. This file contains one gene identifier per line together with the gene symbol and synonyms, among other information.

UniProtKB/SwissProt The UniProt Knowledgebase (UniProtKB) of the Universal Protein Resource provides various functional information on proteins. Only the section "UniProtKB/SwissProt" is considered, because it is manually annotated and reviewed, whereas the (much larger) UniProtKB/TreMBL resource is automatically curated and not manually reviewed. UniProt uses a mnemonic identifier and one or more accession numbers. Both identifiers and accession numbers are considered in our work. The identifier has the quality of being a human-readable mnemonic code, unlike the accession number[5]. In order to reduce redundancy, only the first accession number is considered, although a separate mapping is maintained for accession numbers that refer to the same gene.

Diseases

MeSH diseases MeSH (Medical Subject Headings)[6] is a controlled vocabulary maintained by the United States National Library of Medicine (NLM) and updated annually. It contains keywords used to manually annotate PubMed abstracts and NLM's book database with the aim of facilitating search by providing the subjects of a text. These so called subject headings are available in the format of descriptors which are hierarchically sorted. A connected tree number defines the position in the hierarchy and, at the same time,

[2] https://en.wikipedia.org/wiki/Base36

[3] http://www.ncbi.nlm.nih.gov/gene

[4] ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/All_Data.gene_info.gz

[5] UniProt KB User Manual

[6] https://www.nlm.nih.gov/mesh/

Table 1: Overview of the fields in the combined terminology.

oid	*The internal identifier created by our system*
resource	*The original database from which the terminology of the term was extracted*
original_id	*The original unique identifier from the database of origin*
term	*The term itself which is searched in the text during the process of a dictionary look-up*
preferred_term	*This is the standard term for a concept (if available in the database of origin)*
entity_type	*The entity type of the term (e.g. gene/protein, chemical, disease, species or cell line)*

provides information about the entity type described by a descriptor. The tree branches that we considered recursively with all their subbranches are those for chemicals and drugs (branch D), diseases (branch C) and organisms (branch B). Apart from MeSH descriptors we also included MeSH supplementary records. Supplementary records (SCRs) are updated weekly and contain additional terms which might occur in the literature as concept data (MeSH, 2015). SCRs exist mainly for chemical substances and rare diseases and are mapped to descriptors. This mapping defines their position in the hierarchical structure of MeSH according to which we determined their entity type. MeSH descriptors as well as supplementary records use their own unique identifiers by which information about specific database entries can be retrieved from the database.

Chemicals

MeSH Chemicals and Drugs The chemical and drug branch (branch D) of MeSH. Its structure and method applied are the same as described above for diseases.

ChEBI ontology ChEBI (Chemical Entities of Biological Interest)[7] is a "freely available dictionary of molecular entities focused on 'small' chemical compounds". ChEBI's terminology follows IUPAC (International Union of Pure and Applied Chemistry) and NC-IUBMB (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology) but establishes its own unique and stable identifiers. Data is not only manually curated by ChEBI curators but also integrated from different sources, such as IntEnz, KEGG and PDBeChem.

[7] https://www.ebi.ac.uk/chebi/

Organisms and Species

NCBI Taxonomy The NCBI Taxonomy is a "curated classification and nomenclature"[8] of species referenced in other Entrez databases. At the time of writing this paper, the coverage of the taxonomy is 10% of all described species of life (NCBI, 2015). From the various files that constitute the NCBI taxonomy, only the mapping from taxonomy IDs to names, synonyms and other properties is processed when creating the resource. During preprocessing, the node file, describing parent-child relationships between nodes in the taxonomy, is used to filter all names that are not assigned to leaf nodes.

MeSH Organisms The organisms branch (branch C) of MeSH (description above).

Cell Lines

Cellosaurus Cellosaurus[9] is a thesaurus of cell lines. It is described as a "controlled vocabulary of cell lines" that is used in biomedical research. The resource is freely available under the Creative Commons Attribution-No Derivs License 3.0.

4.3 Implementation

All components of the program were implemented in CPython 2.7.

Automatic downloading and preprocessing

All resources are downloaded automatically using a standalone downloader. The script queries a text file specifying source URLs and, if present, loads timestamps from previous downloads from a log file. For each FTP or HTTP URL specified, the script attempts to query the modification timestamp from the server. If it deviates from the timestamp recorded in the log file, the file is downloaded and, if compressed, extracted from its archive. Finally, the updated log file is written.

[8] http://www.ncbi.nlm.nih.gov/taxonomy

[9] http://web.expasy.org/cellosaurus/

For resources that are provided in a form not ready for automatic processing, a preprocessing step is triggered. In the Entrez Gene gene info file, the header row is reformatted to correspond to the tab-separated format and only columns relevant to term resource are kept. In the NCBI taxonomy names listing, only names for records with the taxonomic rank *species* are kept.

Parsing

For each resource, a parser was implemented, tailored to extract the identifiers and terms from each original database. For each term in the database, the parser produces a row hash with name-value pairs for each field. If a record specifies synonyms or multiple IDs, additional hashes are created for each synonymous term or additional ID. The main term is specified as the preferred term for each synonym.

In order to keep the memory footprint to a minimum, most parsers process resources in an iterative way, row by row. An exception is the MeSH resource, which depends on two separate, interlinked files, hence, the entire resource is loaded into memory.

Finally, a resource builder iterates over all specified resource parser objects and writes each row hash to a row in the output TSV.

Uniprot and Cellosaurus parser Based on a parser described on *Mannheimia goes programming*[10] Uniprot sequence entries and Cellosaurus cell lines are specified as lists of space-separated key value pairs, separated by delimiter lines. Keys may be mandatory or optional and can occur once or multiple times. Multiple occurrences of the species key are concatenated, as these constitute continuations of previous values, not additional values (see Uniprot knowledgebase). Relevant information (terms and accession numbers) are mapped to the row hashes mentioned above. Some keys may have multiple values, each separated by a semicolon. For the Uniprot resource, a copy of the row hash is created for each accession number indicated in a sequence entry.

NCBI Taxonomy parser The names file is provided in a pipe-separated TSV-like format. For each id, a list of terms is provided, each paired with a name class, specifying the type of term. The entry with the name class "scientific term" is used as preferred name. All name classes are considered with the exception of "authority", as these terms specify not only terms but authorship and publication dates. For obscure reasons, entries of the name class "synonym" also occasionally cite authorship. Regular expressions are used to detect and remove citations and additional information in parenthesis, as well as to strip double quotes from these entries. A small subset of entries specify a unique name. For these cases, additional row hashes are generated.

Entrez Gene gene info parser The gene info format is a standard TSV file. A preprocessor converts the non-standard headers and generates a reduced file containing only relevant columns (containing the gene ID, symbol and all synonyms). Synonyms are specified as pipe-separated sequences. For each synonym, an additional row hash is generated.

MeSH XML parser MeSH is provided in two separate files, a descriptor record set and an associated supplement record set. First, both files are parsed into memory. Only the tree structures *Organisms* [B], *Diseases* [C] and *Chemicals and Drugs* [D] are considered. Additionally, a look-up table is generated from the descriptor record set, mapping the ID of a descriptor record to IDs of all trees, in which it occurs. Each supplement is mapped to its descriptor record using the look-up table. Finally, for each descriptor and supplement record, a row hash is generated.

CHEBI OBO parser The CHEBI OBO parser wraps the OBO parser from the *Orange Bioinformatics* add-on for the open-source data mining utility *Orange*[11]. A row hash is generated for each term and, if present, for each synonym. Placeholder synonyms (containing only periods as terms) are discarded.

Web interface

The combined resource is generated and accessed through a web interface[12]. The interface is controlled by a Python Common Gateway Interface (CGI) script, which allows direct control of the creation pipeline. Visitors to the website

[10] http://mannheimiagoesprogramming.blogspot.ch/2012/04/uniprot-keylist-file-parser-in-python.html

[11] http://orange.biolab.si/

[12] pub.cl.uzh.ch/purl/biodb

can select the desired resources through checkboxes. Additionally, the user may provide pattern/replacement pairs for changing how the resources and their entity types are labeled. The reason for this lies in the extensive labeling of certain resources. This is particularly true for MeSH, where labels like *mesh desc(Anatomy)*, *mesh desc(Chemicals and Drugs)* etc. give a detailed description of origin. By replacing the regular expression `mesh desc.*` with `MeSH`, for example, the number of resource labels can be reduced.

After submitting the request, the resource creation process is started in the background. As the creation process may take several minutes, the user is provided an individual link that allows downloading the resource file as soon as it is ready.

5 Term Resource Statistics

Graphs depicting frequency distributions for the terms in the term resource, for each entity type as well as for the whole term resource, can be found in the appendix[13]. Each graph shows the ambiguity of a term (how many IDs per each term), and the reverse property, i.e. how many terms are available for each ID. For example, while cell lines and organisms are mostly unambiguous (the vast majority of them shows a 1:1 correspondence between IDs and terms), chemicals show a much higher degree of ambiguity. Since common matching strategies in dictionary look-up are to lowercase the terms, or to strip them of non-alphanumeric characters with the aim of increasing recall, we also show the additional ambiguity generated by this process.

6 Related Work

Recently, the focus of research on biomedical named entity recognition has rather been on machine learning approaches than on dictionary based approaches, however, assigning unique database identifiers is often a necessity and is frequently used as a second step after the application of a machine learning system.

The task of retrieving database IDs for named entities is promoted by shared tasks, such as the BioCreative workshop which includes normalization sub-tasks by asking participants to provide database identifiers for the named entities in the text. BioCreative II included a gene normalization task encouraging the development of systems which are able to assign an Entrez Gene identifier to genes found in PubMed articles (Morgan et al., 2008).

Thompson et al. (2011) have compiled a very large biomedical dictionary called BioLexicon which, similar to our work, brings together terms from different resources. But additionally, BioLexicon contains terms extracted from text as well as further linguistic information, such as grammatical information and semantic verb roles, all of which is located in a relational database.

Kaljurand et al. (2009) have also compiled a term resource from different database terminologies but focuses mainly on the identification of protein mentions. Furthermore, the current work considers a richer set of terminology, and presents more detailed statistics.

One system, that has successfully used a dictionary look up for gene and protein NER has been described by Hanisch et al. (2005). The lexical properties of the gene ontology (which we have not used in our work so far) have been explored by McCray et al. (2002).

Compared to these related resources, the combined term resource presented in this paper has several advantages. It is continuously updated with the newest version of all included databases, which keeps it up to date. A user can produce a customized file with selected databases, according to specific needs of a project. Apart from this, it can be easily extended with further resources, which we are planning to do in the future. Last but not least the format is very simple and therefore allows for easy integration into any kind of text mining system. Furthermore, also due to its simplicity, the format is more lightweight, taking up less memory than other formats commonly used for terminological databases.

7 Conclusion

In this paper we presented a new terminological resource for biomedical text mining. The resource is compiled of terminology extracted from a selection of biomedical databases. Its very simple format facilitates integration into any system for biomedical text mining in order to perform a dictionary look-up. The novelty of our approach

[13] The graphs were generated using *plotly* (https://plot.ly/)

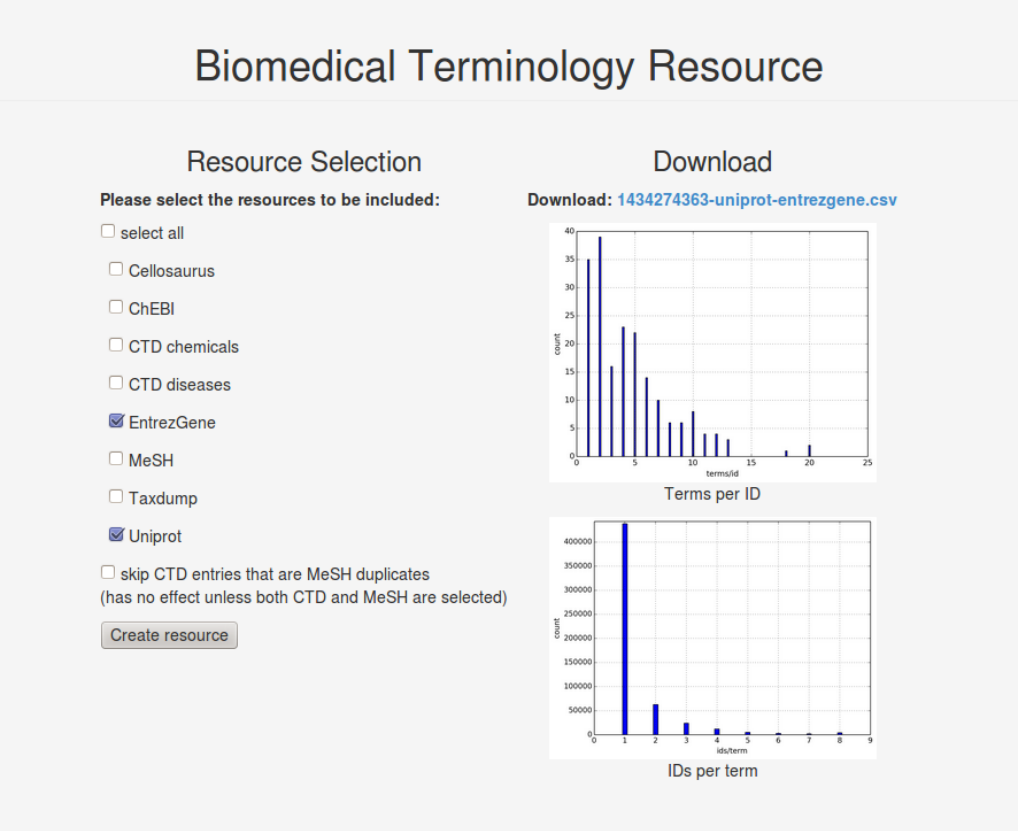


Figure 1: Web interface.

is that the resource is continuously updated with new terms found in the respective databases. Furthermore, a user interface provides a user with a choice of different databases and entity types so that an individual resource can be compiled. Various statistics for the compiled termfile give insight to the lexical properties of the terms contained in the resource. This sheds light on differences between the lexical properties of different entity types. Additional terminological resources will be included in future versions of the system.

References

Garth R. Brown, Vichet Hem, Kenneth S. Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D. Pruitt, Donna R. Maglott, and Terence D. Murphy. 2015. Gene: a gene-centered information resource at ncbi. *Nucleic Acids Research*, 43(D1):D36–D42.

Martin Gerner, Goran Nenadic, and Casey M. Bergman. 2010. Linnaeus: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11:85.

Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. 2005. Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, pages –1–1.

Antonio Jimeno-Yepes, Ernesto Jiménez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga Llavori, and Dietrich Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(S-3).

Kaarel Kaljurand, Fabio Rinaldi, Thomas Kappeler, and Gerold Schneider. 2009. Using existing biomedical resources to detect and ground terms in biomedical literature. In Carlo Combi, Yuval Shahar, and Ameen Abu-Hanna, editors, *AIME*, volume 5651 of *Lecture Notes in Computer Science*, pages 225–234.

Michael Krauthammer and Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512 – 526. Named Entity Recognition in Biomedicine.

Alexa T McCray, Allen C Browne, and Olivier Bodenreider. 2002. The lexical properties of the gene ontology. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 504–508.

MeSH. 2015. Introduction to MeSH – 2015.

https://www.nlm.nih.gov/mesh/introduction.html. Accessed: 2015-06-16.

Alexander Morgan, Zhiyong Lu, Xinglong Wang, Aaron Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jorg Hakenberg, Chengjie Sun, Heng-hui Liu, Rafael Torres, Michael Krauthammer, William Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K Bretonnel Cohen, and Lynette Hirschman. 2008. Overview of biocreative ii gene normalization. *Genome Biology*, 9(Suppl 2):S3.

NCBI. 2015. NCBI Taxonomy - Frontpage. http://www.ncbi.nlm.nih.gov/taxonomy. Accessed: 2015-07-02.

Fabio Rinaldi, Allan Peter Davis, Christopher Southan, Simon Clematide, Tilia Renate Ellendorff, and Gerold Schneider. 2013. Odin: a customizable literature curation tool. In *Fourth BioCreative Challenge Evaluation Workshop*, volume 1, pages 219–223. Biocreative, October.

Tim Rocktäschel, Michael Weidlich, and Ulf Leser. 2012. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

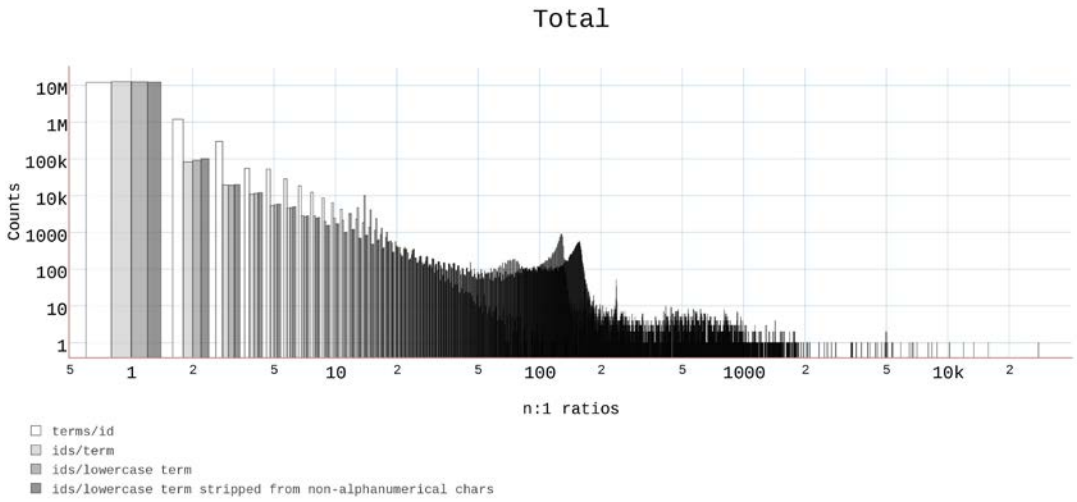
Sirarat Sarntivijai, Alexander S. Ade, Brian D. Athey, and David J. States. 2008. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics*, 24(23):2760–2766.

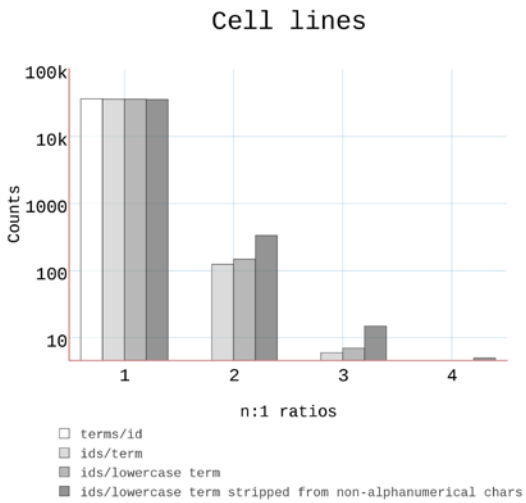
P. Thompson, J. McNaught, S. Montemagni, N. Calzolari, R. del Gratta, V. Lee, S. Marchi, M. Monachini, P. Pezik, V. Quochi, C. J. Rupp, Y. Sasaki, G. Venturi, D. Rebholz-Schuhmann, and S. Ananiadou. 2011. The biolexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics*, 12:397.

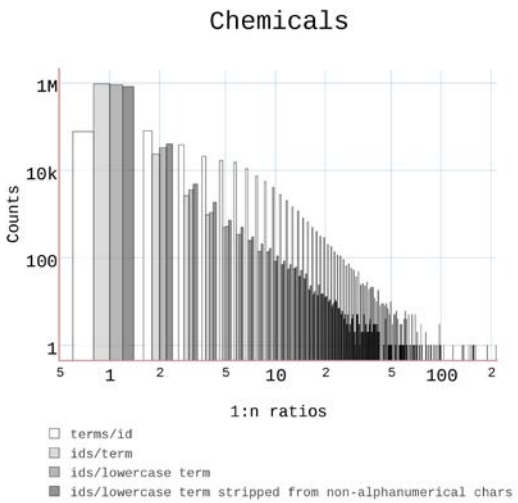
A Termfile Statistics

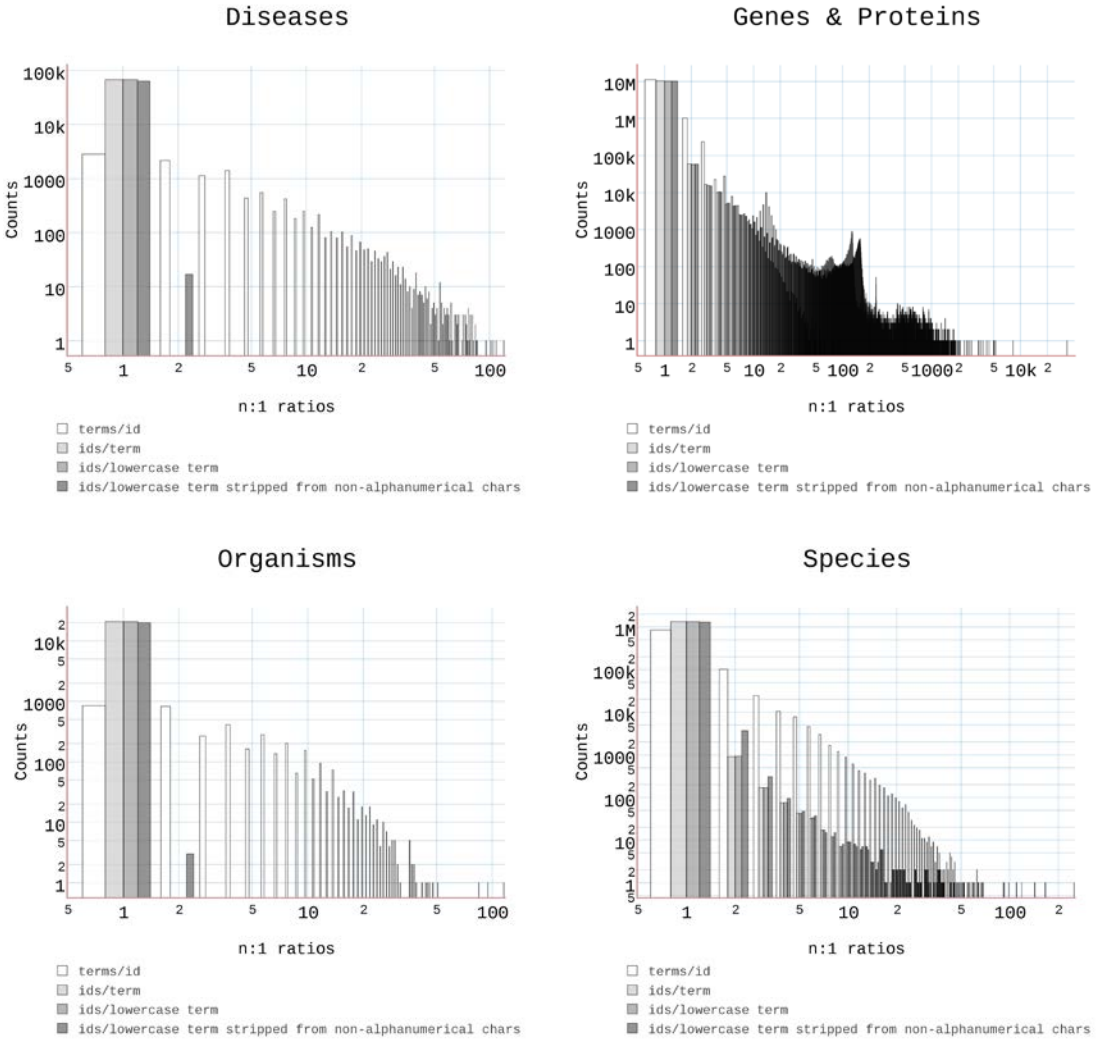
Table 2: Overview of termfile statistics.

	genes/proteins	chemicals	diseases	species	cell lines	all entity types (whole resource)
Number of Terms in Resource	10,429,162	979,418	67,614	1,333,903	36,249	12,846,346
Average of Term Length (number letters)	11.73	37.49	26.98	22.87	7.611	14.92
Average of terms per original IDs	1.1455	3.545	6.018	1.326	1.000	1.328
Average of original IDs per term	1.371	1.049	1.000	1.003	1.004	1.306
Average of original IDs per term (case insensitive)	1.383	1.062	1.000	1.003	1.004	1.317
Average of original IDs per term (case insensitive, non-alphanumeric characters removed)	1.387	1.086	1.000	1.006	1.010	1.324








Diseases
Genes & Proteins
Organisms
Species
Counts
n:1 ratios
terms/id
ids/term
ids/lowercase term
ids/lowercase term stripped from non-alphanumerical chars

A logical information system proposal for browsing terminological resources.

Annie Foret
IRISA, University of Rennes 1
Campus de Beaulieu, 35042 Rennes cedex, France
foret@irisa.fr

Abstract

This article presents an automated construction of a logical information context from a terminological resource, available in xml ; we apply this to the resource FranceTerme and to Camelis tool and we discuss how the resulting context can be used with such a tool dedicated to logical contexts.

The purpose of this development and the choices related to this experiment is two-fold : to facilitate the use of a rich linguistic resource available as open-data in xml ; to test and envision a systematic transformation of such xml resources to logical contexts. A logical view of a context allows to explore information in a flexible way, without writing explicit queries, it may also provide insights on the quality of the data. Such a context can be enriched by other information (of diverse natures), it can also be linked with other applications (according to arguments supplied by the context).

Keywords : Scientific terminology, Technological terminology, Multilingual applications, Information extraction, Textual data mining, Information retrieval, Linguistic resources, Open Data, Information Quality, Legal Information.

1 Introduction

This study aims to make linguistic data easier to exploit through the *logical information systems* approach : whereas such data are not always easy to use without assistance or expertise, logical information systems are especially designed to offer a flexible browsing of data when organized as a *logical context*. Some other works use a similar frameworkbut their data are of different nature, and their goals as well : (Cellier et al., 2011) apply Logical Concept Analysis to explore sets of patterns obtained by data-mining, (Quiniou et al., 2012) consider stylistic patterns, (Foret and Ferré, 2010) consider type-logical grammars, (Falk et al., 2014) uses several features including a thematic one to help identify new words.

In this proposal, we want both :
– to facilitate the use of a valuable linguistic resource (with a rich structure) and available in XML, and to allow its flexible querying and exploration without prior knowledge ;
– we want to test and consider a systematic transformation (*a transducer*) from such resources (in XML) to logical contexts ; such contexts can be loaded in a software allowing rich and flexible browsing on data, combining various heterogeneous criteria ; the way we represent the information in the context may also have an impact on its ease of use.

The aim is here to perform a transducer so as to present the data in a logical information system without losing information content, but gaining in ease of exploration. Other advantages are provided by a safe navigation (no dead-end property) and serenpidity.

The resulting context is freely available [1].

Terminological resource. The selected resource concerns the scientific and technical fields, it also interests us for the richness of its structure : its multilingual aspects, with definitions, synonymous relations, etc. its confirmed status (with source and date of publication), variations according to domain/subdomains or according to linguistic criteria (several variants of English, for

1. at http://www.irisa.fr/LIS/softwares

example), possible absence or possible repetition of certain types of information.

This rich structure also allows further extensions : either with existing data or with new data that we organize in a similar pattern.

Logical context. A logical context is defined by a finite set of objects $\mathcal{O}$,[2] and a finite set of logic descriptions $d(o_i)$ expressed using a well-formed logical language L.

A *Logical context management system* can load and manage such a context, allowing querying a context by logical requests (explicit or interactive) ; then the answer is a sub-context of objects satisfying the query. We used CAMELIS (version 1)[3] for the experiment reported in this article. This software is based on *Logical concept analysis (LCA)* as defined in (Ferré and Ridoux, 2004). LCA is an an extension of the formal concept analysis (FCA, see (Ganter and Wille, 1999)) : a *logical concept*, denoted c is a pair formed of an extent $ext(c)$ (a set of objects) and an intent $int(c)$ (a formula) such that the elements of $ext(c)$ are exactly those which satisfy $int(c)$. These concepts form a lattice underlying the incremental *logical navigation tree* in the left window of the software. The software CAMELIS is also designed for managing sets of objects of different types. Object descriptions in a given logical context can have several origins : they can be retrieved by a transducer or come from extrinsic judgments (personal notes, for example) ; combining these modes allows to enrich the context and adapt it according to a user preferences.

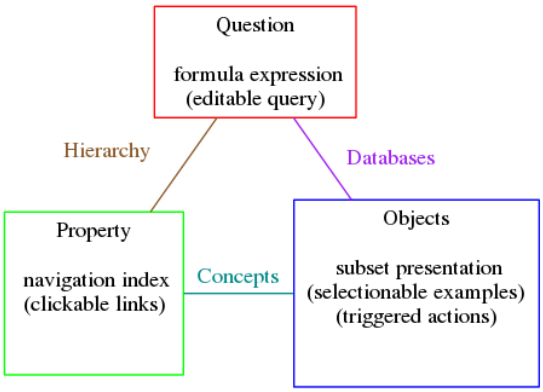


FIGURE 1: LIS and LCA

2. (several objects can have the same label)
3. http://www.irisa.fr/LIS/ferre/camelis/

Figure 1 illustrates how LCA generalizes Database and Hierarchical systems, the figure also follows the interface that enables three modes and shows synchronized related windows (as in figure 5) : a query on the top, links in the navigation index on the left, or objects on the right.

Thereafter, we present in section 2 our transducer implemented in XSLT[4] and we specify the construction methodology. We present in section 3 how to exploit the transducer output on the FranceTerme resource containing terms of different scientific and technical fields ; we discuss several scenarios and benefits of this approach through this experiment. Additions and adjustments are proposed and discussed in 4 before concluding in section 5.

2 The transducer methodology

2.1 Some key aspects

The transducer is designed to present data in a logical information system without losing information content, but gaining in ease of exploration. The diagram in figure 2 illustrates the approach, where the automated steps (solid arrows) are distinguished from manual or semi-manual ones (dotted line).

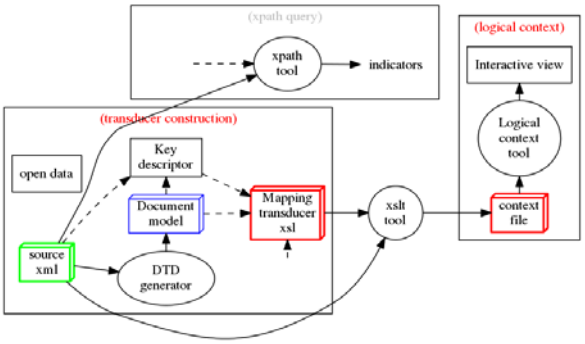


FIGURE 2: global architecture

Source Document Schemas. The transducer is designed to apply automatically on documents conforming to some document specifications. In general, such a specification can be automatically produced from an XML instance. We used DTD

4. a web availability is planned.

generator [5]

Logical information system capabilities. We recall that a logical context system must allow the loading and management of a context described by its objects and its objects properties. More precisely, we assume that :
- the logical context allows for some *inferences* : at least from classical logic (such as *if A then A or B*) possibly with axioms useful to model the context and organize its presentation (for example to reflect a taxonomy and only see some salient properties) ;
- some general form of information retrieval and multi-faceted means are provided, as *"logical facets"* and "logical criteria" combinations ;
- a *modular and dynamic* construction is allowed, for both sets of objects and sets of properties.

CAMELIS **system.** In our experiment, we used the logical context tool CAMELIS, that, to our knowledge, is the only logical context management system. The tool interface displays three connected windows (see figure 3) :
- a query window (top) ;
- an object window (right) ;
- an index window (left) as a *navigation tree*.

Properties in the navigation tree are organised as a clickable summary grouping hierarchy properties : it is important to note that a navigation link there corresponds to a *sub-context* (as in figure 5, the link/sub-context cardinality is given and a color is also associated with a concept :two navigation links with the same color lead to the same sub-context).

In the browsing mode, this tool allows three forms of query that select a *sub-context* :

expert/query mode by *editing in the query window* ; the displayed objects become those statisfying the logic query ; the navigation tree is then presented in a form adapted to the new context ;

example/object mode by *selecting a set of objects in the object window* ; the query automatically becomes an expression for the common properties of objects ;

index/property mode by selecting a property (or more) in the navigation tree ; the objects displayed become all those verifying the selected property (links) ; the query window is then automatically updated.

CAMELIS **general properties.** The *consistency* between the three windows is ensured. In addition, a session will not lead to an empty set of objects when following the links in the navigation tree : this important property is the *navigational safety*.

CAMELIS *update and logic modularity.* Using the same interface, we can add and dynamically update objects and properties, then export as a new *logical context file* (useful for example to generate a documentation for the objects of a selected sub-context). The tool can also be adapted to choose a dedicated logic, obtained by combination of logic functors (Ferré and Ridoux, 2004).
We do not detail these last aspects in this article.

Control. Part of the context information visible in the tool, can be retrieved by other means. We built some XPath queries to control the process and to produce complementary indicators.

We also built a control-context (figure 7).

2.2 Transduction overview

We give here the characteristics and main stages of realization of the transducer, in its basic version. This construction is guided by the information in the DTD generated by XPath queries and control.

Key selection. Defining a key in the source (by means of its document schema or of an XPath expression) is a preliminary stage, and the key definition plays a central role in the context construction. For this experiment with `FranceTerme.xml`, we considered `//Article/@id` (XPath).

Components. The program is designed to facilitate its updating, structured by source components and similar typical treatments. The treatment of a given source component depends on its kind (XML element, attribute) the relevant part of context, its status (optional, repeatable or not), the domain of the source content, and the desired rendering (data type, property name, property hierarchies).

Main Loop. For each source item `Article` :
— each object get a unique label (extracted by `Terme[@statut='privilegie']`), used for the object presentation and for a string property in the navigation index ;

5. it is available at http ://saxon.sourceforge.net/dtdgen.html ;the terminological XML source file we used is accompanied with an XML Schema xsd, but without guarantee

— the key becomes a number property `articleID = ...` (xslt [6]) ;
— the publication date is processed to appear in the index at different levels of date detail ;
— most other components are processed to produce strings of the form :
`property_name is "string_value"`
— property names may depend on several XML components (such as `Terme` element with a `@statut` attribute), they are organized to allow their grouping and multiple levels of detail (`Terme ?` is more general than `Terme SYNONYME`) ;
— we also give a common prefix `_Plus` to properties for data in the source, but not visible in the ressource site (such as data about committees) ; [7]
— for XML elements that can be repeated for a given key/object, (such as `Terme`) we use an inner loop ;
at this stage the output file context contains the description of one object per line, with its main properties ;
— other components (such as foreign equivalents or antonymes, optional or repeated) are rendered by rules of the form :
`rule_extr (key=id) --> (prop1 is val1)`
that automatically associate the property to the object designated by the key.

2.3 Modularity of context

For treating a logical property related to an XML component, such as `<Attention>` child (optional and repeatable) of an `<Article>` identified by its attribute `id`, several alternatives are possible :
-to indicate the name of property and its value by assembling and repeating the property name for the object, using this pattern :

```
mk "object" key=id, ..., is prop1
val1, val2 prop1 is ... is ... prop2
```

-to indicate each property value by transformation rules, using this pattern, when the object is assumed to be already created and associated with the key :

```
rule _extr (key=id) --> (prop1 is val1)
rule _extr (key=id) --> (prop1 is val2)
```

this will automatically add each property value to the object with key `id`.

This second solution brings some modularity since we can put rules in separate files, the properties being effectively added to objects after a file import. We chose this approach by means of rules and keys for some repeatable components. [8]

3 Logical Context and facets

In this part, we discuss several possible search modes in the resulting context, where navigation links (incremental) correspond to logical concepts that can be selected.

3.1 Simple searches on several data types.

Multilingual data. The `FranceTerme` resource contains translations in several languages, with variations for the same language. Those data are attached to various domains and subdomains (possibly several ones for a given object).

Scenario. An exploration of the logical context can be conducted that way, for example :

- open the `Domaine ?` property in the index ;
- select-click `Domaine is "Informatique"` (computer science), this yields the corresponding sub-context (with 3 coherent views) ;
- we may further select-click `Domaine is "Droit"` (Law), also automatically expressed as `Domaine is "Informatique" and Domaine is "Droit"` in the top window ;
- open the `Equivalent ?` property then open `Equivalent _en is "..."` in the index etc.
- open the `PubliArticle ?` property then `PubliArticle date = 2014` in the index etc.

Another simple search on ”streaming” is shown in figure 3.

Sub-context Cardinalities. In the property index-tree, we may choose an order for displaying a given facet. This is useful for example to read directly which `Equivalent _en` correspond to the greatest number of French terms (figure 3).

6. `<xsl:value-of select="concat('articleID='$varArtId)"/>`
7. in the navigation tree, compound names of the properties are grouped by prefixes, details appear by opening a link `_Plus ?`
8. Here is a typical xslt fragment (with some special symbols treatments for compatibility) :

```
<xsl:for-each select="Attention"> <xsl:value-of
select="concat($varRulePart1,'idArticle=',
$varArtId,$varRulePart2, 'Attention is ', $varQuote,
translate(normalize-space(./text()), $varQuote, $varBQuote),
$varQuote)"/> <xsl:text> </xsl:text> </xsl:for-each>
```

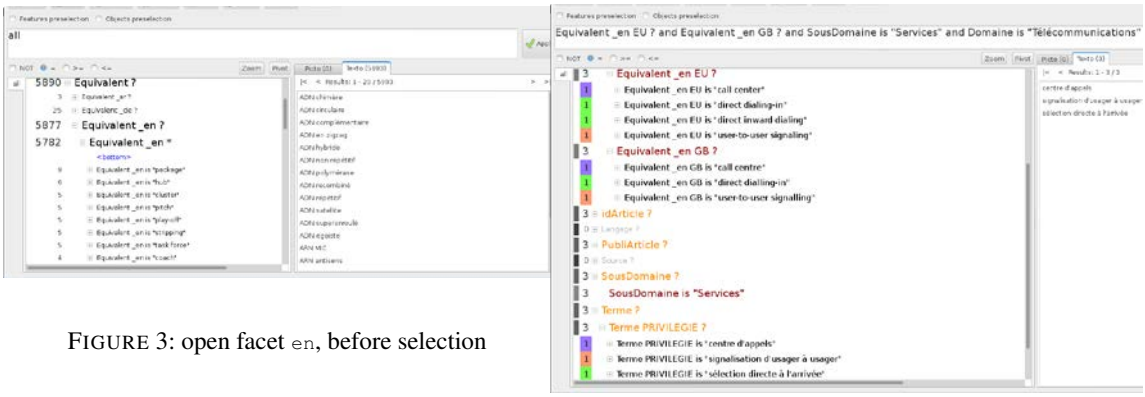


FIGURE 3: open facet `en`, before selection

FIGURE 5: Variants

Data types. Data types other than attributes and strings can be handled, Figure 4) shows a possible use of dates, allowing for more or less fine-grained selections.

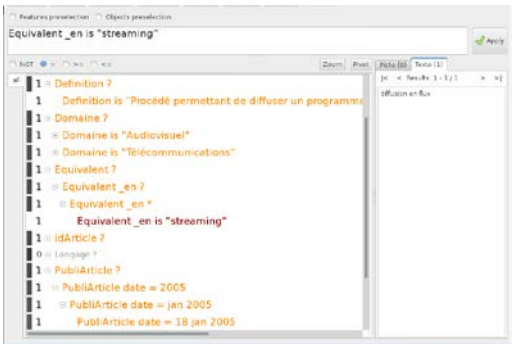


FIGURE 4: Facet `en` selected, facets date and domain opened

Exploring variants and false-friends. Figure 5 comes from selections in the index tree ; links of a same color characterize the same set of objects (concept).

This example illustrates the identification of potential linguistic errors (in a domain / subdomain)

Other examples such as ”package” (`Equivalent_en`) or some abreviations (”ABS”) can be highlighted as ambiguous : the dynamic navigation links (domains, etc.) then provide hints to disambiguate.

3.2 Focused search on elements and exceptions : summaries.

Using the `Not` button at a given stage, we can arrive at a subcontext characterized by $A_2 = not(A_1)\ and\ A_0$ as in Figure 6. Property A_2 expresses a search for exceptions to A_1, we get a kind of subcontext summary and extra informations (here `Attention` is the focus element).

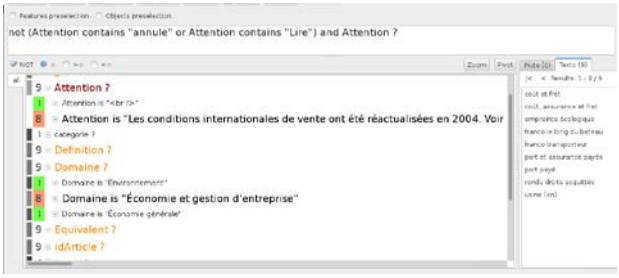


FIGURE 6: Focus on `Attention`

3.3 Other scenarios : data quality

This navigation mode allows to detect abnormalities, in particular pseudo-empty properties appear on other facets (through a link like `Definition is ""`) these cases often correspond to existing but empty XML elements (but are not XML errors). Low cardinalities in the navigation tree may suggest to explore the link, by selecting it and opening other facets simultaneously ; we can analyse this way ”the words without translation, following the `not Equivalent?` link.

In case of redundancies, these may become easily noticeable througs browsing : exploring the `Antonymes` facet, we can see XML structuring redundancies (this information being carried by two source elements).

3.4 Control and actions from a context

The logical context software can assign actions to objects by properties ; clicking on an object label then shows a contextual action menu.

This is useful in particular to inspect objects in their source xml file.

Other kinds of control (for coverage, counts, etc.) are made
- by XPath queries on the XML document used to verify if certain characteristics of particular subcontexts (planned or explored) are consistent with the source document ;
- a meta context built, following the DTD schema, whose objects are : element names, and the pairs (attribute name, element name). These objects are associated with actions parameterized by their label, in our case (Figure 7), the action is an XPath query using BaseX ; This can be adapted easily to another set of controls.

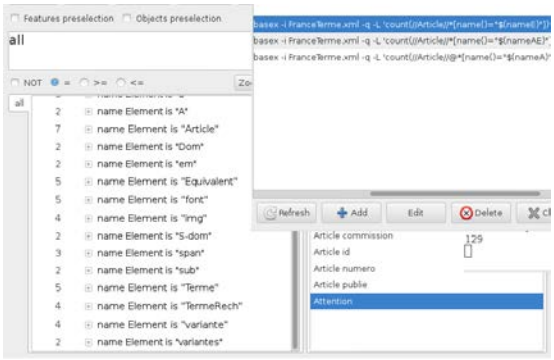


FIGURE 7: control (meta) context

4 Refinements and user preferences

A logical context tool such as CAMELIS by its genericity and its features, allows many alternatives to represent and use a terminological resource like `FranceTerme`.

Some initial choices can be easily revised or completed ; for example the name of a property can be changed directly through the interface (or by simple transformation of the context file).

The choices and refinements should provide a better context. Several quality criteria can be considered : effectively obtaining a desired result (usefulness / completeness) ; the number of steps to get there (effectiveness) ; rich browsing indexes (multiple views) and efficient indexes to pursue the navigation (through the proposed increments) ; a flexible mode of interrogation and ease of interpretation.

We indicate some possibilities through modifications in the experiment.

4.1 Adapting facets using rules

Domains and SubDomains data have been translated. The resulting two context files contain update rules of the form :

```
rule_extr Domaine is "Acoustique" -->
Domain is "Acoustic"
```

when loaded in the context, properties on the right hand side are added to all objects verifying the left hand side.

4.2 Improving grammatical categories using rules and axioms

In the original context, we can see (with an appropriate ordering) that among the terms with a category attribute, the names `(categorie is "nm" or categorie is "n")` are the majority, followed by `categorie is "adj."`. However these grammatical categories are listed with various values, we can observe : which may include in particular :
- a disjunction, as in : `categorie is "adj. ou n.m." and Equivalent _en is "crossmedia (n. ou adj.)"` which selects the term `transmédia` ; but we also observe its permutation `categorie is "n.m. ou adj."` ;
- a more or less fine granularity, as in : `categorie is "n.m.inv."`

The addition of rules and axioms in the logical context permits to harmonize these properties, resulting a more structured navigation tree according to this facet. A few lines in the resulting context define a hierarchy of categories, such as :

```
rule_extr categorie is "n.m.inv." -->
Categorie_n_m_inv
...
Categorie_n_m_inv axiom, Categorie_n_m
Categorie_n_m axiom, Categorie_n
...
```

Note that such improvements could apply to other terminological resources and result from linguistic analysis or other lexicons.

4.3 Axioms for property variants

We have seen that a property `Terme?` covers three statutes (`PRIVILEGIE`, `SYNONYME`, `ANTONYME`).

A context of axioms can facilitate a search on all or a subset of these variants :

```
axiom PRIVILEGIE, SET
axiom SYNONYME, SET
axiom SET, ANY
axiom ANTONYME, ANY
```

A query may then group several properties (having a status below another expression like `SET`) as examplified in figure 8.

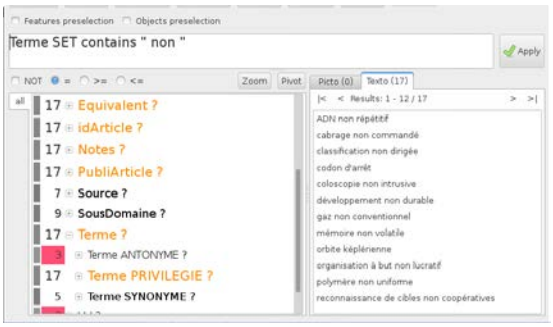


FIGURE 8: with axioms on property names

This example shows a query for Terms (including status variants) containing negation (”non”).

Note that such axioms can be added or modified in a modular way.

4.4 Linking data and resources

At the property level in the navigation tree. The website `FranceTerme` allows to select new terms, from the description of a current term (denoted as t_1) by a link `See also`. This is rendered in the context navigation tree by the facet `Voir aussi (See also) is "... information on` t_j`"` (denoted as f_j) where the term to see t_j is shown with its key `Article` . Two modes of translation of this piece of information have been tested :
- in a basic mode, as a simple property f_j of the current term t_1
- in a full (reflexive) mode, where t_1 has f_j and also f_1 : `Voir Aussi is "... information on` t_1`"`. [9]

This second mode allows this type of scenario : while t_1 has property f_j, select this f_j link in the navigation tree ; by reflexive closure, t_j is also in the sub-context, and can be further selected.

We treated in the same way the reciprocal link of `Voir Aussi`, by adding (by the transducer) a `Voir Depuis (See From) "..."` property. This enables to group into a sub-context the terms pointing to (or pointed to by) a particular word (or more).

Notes.
- In the context for `FranceTerme`, we observe some terms (15) satisfying this query :
`Voir Aussi? and not Voir Depuis?`
these are the ”terms pointing to an article, but not pointed to” ;
- according to the resource schema, other elements (synonyms, antonyms, ...) could be treated similarly, with navigation links in context. [10]

Linking to other resources at the level of actions. As explained about control, the logical context management software allows to associate actions to objects.

We can use this mode to associate an object with a process (or more) on this item that may be introduced in the interface from a context menu related to the object (or group of objects). The setting can be provided at the transducer level. A file describing these actions can be later loaded from the interface.

We generated connections to :
- a parser, installed locally : the processing chain (open) `Bonsai` (Candito et al., 2010) which takes as parameter the label of the object ; a selection of this action on the object provides a syntactic analysis of the label expression ;
- a web link to another terminological resource for French (CNRTL, `http://www.cnrtl.fr/`) the parameter being the term as above ; a selection of this action on the object opens the browser on the website page for this term, if there exists one (none for some `FranceTerme` expressions) ;
- an XML link to a subpart of the source file, through an XPath tool (BaseX) the parameter being the object key (attribute `id` of `Article`) ; a selection of this action on the object executes the software with a prepared BaseX request using the object key.

This action list is not exhaustive and can be adapted. In particular, we could consider links (local or not) with other analyzers, or other linguistic resources and retrieve results to enrich the logical context with new properties. The capabilities of *Full text search* (of BaseX) could also be ex-

9. no addition to the terms that have no link

10. this treatment is not currently done for the other elements (in the source xml these terms do not necessarily correspond to an article/object).

ploited.

5 Conclusion

The general aim of this proposal was to show how a logical concept analysis (LCA) framework and tools could be beneficial for browsing terminological resources ; through this experiment the purpose was twofold :
- to facilitate the use of a useful language resource (rich structure) and available in XML,
- to envision a systematic transformation of such resources as XML to logical contexts.

Improvements may also be suggested and brought to the data ; other treatments may also be eased, for example a selected sub-context can be exported as text and generate other results (such as a documentation).

We illustrated how a logical context allows to explore linguistic information, in a flexible way, without a priori knowledge, and also get guidance on data quality (in the navigation tree, counts and colors for concepts, ...) New linguistic information (personal, enterprise, ...) could be incorporated easily in the initial context (if they comply with the document model and the key assumption).

Additional data to compare and enrich the content can also be added in several ways and for many languages, (for French : Wordnet Wolf (Sagot and Fiser, 2012), Lefff lexicon (Sagot, 2010), etc.) :
- by adding objects without confusion between sources (since a property indicating the source is associated with the object) ;
- by adding properties to expand the browsing possibilities ;
- by adding triggered actions on objects.

Other actions corresponding to linguistic processing can be added to the context :parsing the expression (several languages), syntactic head, etc. We could also consider inverse properties (such as translation) and enrich the context with these objects.

Moreover, it seems that the development method could be transposed to other open data and linguistic xml ressources. To some extent, the construction of the transducer presented here could be automated if it relies on a determination of a key and a grid indicating for a *source component*, its label, its type, its repeatability, and the way it should be rendered. A similar experiment could be carried out by adapting the standards and software tools of the semantic web. Finally, we mainly discussed browsing, future work could also concern updates.

References

Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. 2010. Benchmarking of statistical dependency parsers for french. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 108–116. Chinese Information Processing Society of China.

Peggy Cellier, Sébastien Ferré, Mireille Ducassé, and Thierry Charnois. 2011. Partial orders and logical concept analysis to explore patterns extracted by data mining. In Simon Andrews, Simon Polovina, Richard Hill, and Babak Akhgar, editors, *Conceptual Structures for Discovering Knowledge - 19th International Conference on Conceptual Structures, ICCS 2011, Derby, UK, July 25-29, 2011. Proceedings*, volume 6828 of *Lecture Notes in Computer Science*, pages 77–90. Springer.

Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From non word to new word : Automatically identifying neologisms in french newspapers. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 4337–4344. European Language Resources Association (ELRA).

Sébastien Ferré and Olivier Ridoux. 2004. Introduction to logical information systems. *Inf. Process. Manage.*, 40(3) :383–419.

Annie Foret and Sébastien Ferré. 2010. On categorial grammars as logical information systems. In Léonard Kwuida and Baris Sertkaya, editors, *Formal Concept Analysis, 8th International Conference, ICFCA 2010, Agadir, Morocco, March 15-18, 2010. Proceedings*, volume 5986 of *Lecture Notes in Computer Science*, pages 225–240. Springer.

Bernhard Ganter and Rudolf Wille. 1999. *Formal concept analysis - mathematical foundations*. Springer.

Solen Quiniou, Peggy Cellier, Thierry Charnois, and Dominique Legallois. 2012. What about sequential data mining techniques to identify linguistic patterns for stylistics ? In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing - 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Pro-*

ceedings, Part I, volume 7181 of *Lecture Notes in Computer Science*, pages 166–177. Springer.

Benoît Sagot and Darja Fiser. 2012. Cleaning noisy wordnets. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 3468–3472. European Language Resources Association (ELRA).

Benoît Sagot. 2010. The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Constructing a Syndromic Terminology Resource for Veterinary Text Mining

Lenz Furrer
Institute of Computational Linguistics
University of Zurich
lenz.furrer@uzh.ch

Susanne Küker
Veterinary Public Health Institute
University of Bern
susanne.kueker@vetsuisse.unibe.ch

John Berezowski
Department of Clinical Research and Veterinary Public Health
University of Bern
john.berezowski@vetsuisse.unibe.ch

Horst Posthaus
Institute of Animal Pathology
University of Bern
horst.posthaus@vetsuisse.unibe.ch

Flavie Vial
Veterinary Public Health Institute
University of Bern
flavie.vial@vetsuisse.unibe.ch

Fabio Rinaldi
Institute of Computational Linguistics
University of Zurich
fabio.rinaldi@uzh.ch

Abstract

Public health surveillance systems rely on the automated monitoring of large amounts of text. While building a text mining system for veterinary syndromic surveillance, we exploit automatic and semi-automatic methods for terminology construction at different stages. Our approaches include term extraction from free-text, grouping of term variants based on string similarity, and linking to an existing medical ontology.

1 Introduction

In the project Veterinary Pathology Text Mining, we are developing tools to exploit veterinary post-mortem data for epidemiological surveillance and early detection of animal diseases. This paper describes the work in progress on the construction of a veterinary terminology resource as a basis for a text mining tool to classify, with minimal human intervention, free-text veterinary reports with respect to multiple clinical syndromes that can be monitored.

In human medicine, text mining has been successfully applied to clinical records in many public health surveillance systems (Botsis et al., 2011; Steinberger et al., 2008; Brownstein et al., 2008; Wagner et al., 2004). The approaches range from hand-written rule-based systems to fully automated methods using machine learning. For example, Chapman et al. (2004) use heuristical keyword-driven as well as supervised machine learning techniques (Naïve-Bayes classifier) for detecting mentions of fever in free-text clinical records. Similarly, the BioCaster system (Collier et al., 2006; Collier et al., 2008) relies on a carefully constructed medical ontology combined with a Naïve-Bayes classifier as an input filter. Friedlin et al. (2008) use a regular-expression based term-extraction system to find positive and negative mentions of methicillin-resistant *Staphylococcus aureus* in culture reports. Hartley et al. (2010) give an overview of surveillance systems that mainly focus on world-wide monitoring of web sources, including news feeds and informal medical networks.

The text mining of veterinary reports faces additional challenges such as multiple species and a less controlled vocabulary (Smith-Akin et al., 2007; Santamaria and Zimmerman, 2011). Up to this point, approaches for classifying veterinary diagnostic data into syndromes for surveillance have been restricted to the use of rule-based classifiers (Dórea et al., 2013; Anholt et al., 2014). To build these classifiers, a group of experts manually creates a large set of rules. The rules are then used to classify veterinary diagnostic submissions into syndromes based on the presence or absence of specific words within various fields in the diagnostic submission data.

We propose to develop a process for using text mining methodologies (natural language processing) to efficiently extract relevant health information from veterinary diagnostic submission data with minimal human intervention. Given a sufficient amount of data (i. e. at least a few hundreds of manually classified reports), a machine

learning approach will allow us to directly classify these data into syndromes that can be monitored for surveillance.

As recognized in the Swiss Animal Health Strategy 2010+, methods for early disease detection, based on the increasing abundance of data on animal health stored in national databases, can contribute to valuable and highly efficient surveillance activities. Post-mortem data, available from pathology services, are often under-exploited. The main purpose of post-mortem investigations of food production animals is to provide information about the cause of disease or death with regard to treatment, and prevention options for the affected herd. Besides these major diagnoses, all additional pathological findings are also recorded as text and electronically archived as necropsy reports. In addition to the value of this information for veterinarians and farmers, systematic evaluation of necropsy data may be of value the early detection of spatio-temporal clusters of syndromes which may result from a new disease emerging into a population or from changing patterns of endemic diseases. As such, it has the potential to be of value for both nation-wide and international (veterinary) public health early-warning systems.

The rest of this paper is organized as follows: We present our efforts in constructing and exploiting a veterinary terminology resource in Section 2. Section 3 describes our work towards report classification in the context of building a surveillance tool. The next steps and further application scenarios are given in Section 4.

2 Terminology Construction

In the process of report classification, we have put a lot of effort in the construction of a terminology resource that suited our needs. The resulting term inventory is tailored to a very specific task. Still, the methods, insights and even the resource itself can be of use for other applications. Similar to the work by Rinaldi et al. (2002), we extracted a set of terms from a collection of raw text and used automatic methods to organize them into a hierarchical structure. Section 2.1 introduces the categories we used for classification. In Sections 2.2 and 2.3, we describe the steps that led to the construction of the term inventory. Sections 2.4 and 2.5 show how this resource can be automatically enhanced for a more general usage.

2.1 Syndrome and Diagnosis Classification

The work described here is based on post-mortem reports that were compiled by the Institute of Animal Pathology (ITPA) of the Vetsuisse faculty at the University of Bern. The data were entered into a database by veterinary pathologists between 2000 and 2011. We used a subset of approximately 9 000 report entries regarding pigs and cattle. The reports are written in German, with a small fraction (less than 3 %) in English and French.

For subsequent quantitative analysis, we classified all reports using two categorization levels. As a coarse-grained categorization, we annotated each report with the syndromic groups that were affected by a medical issue. Each report was assigned zero, one or more of 9 syndrome categories (gastro-intestinal, respiratory, urinary, cardio-vascular, lymphatic, musculo-skeletal, reproductive, neural, other). This categorization approximately meets the level of granularity found in other work (Dórea et al., 2013; Warns-Petit et al., 2010). For a finer-grained categorization of the reports, we additionally annotated post-mortem diagnoses mentioned (directly or implicitly) in the reports, such as *enteritis*, *lipidosis*, or *injuries from foreign bodies*. The set of diagnoses was not defined a priori, but continuously updated in the classification process. The final set comprised some 100 classes and is shown in Table 1. The diagnoses are modeled as subcategories of the syndromes. While some category names occur in more than one syndromic category, it does not mean that they are ambiguous, as they are triggered by different terms. For example, *atresia* is classified as a congenital abnormality of the gastro-intestinal system, whereas the *ventricular septal defect* is a congenital abnormality of the cardio-vascular system.

2.2 Term Normalization

The medical reports have a high number of surface variants per term. The variation is caused by inflection, inconsistent spelling and typographical errors. On a higher level, variation is increased by synonymy, i. e. the use of different terms for the same concept (e. g. *Lipidose/Verfettung* 'lipidosis'). From the perspective of the given text mining task, certain derivative forms can be considered synonymous variants as well (e. g. *Ulzeration* besides *Ulkus*).

We split the report texts into tokens, which we

gastro-intestinal	
abomasal ulcer	178
abomasitis	120
acidosis	41
cheilitis	6
cholangitis	84
colitis	368
congenital abnormality	21
dilatation	159
displaced abomasum	55
duodenitis	12
enteritis	2458
esophagitis	46
gastric ulcer	206
gastritis	121
glossitis	16
hepatitis	200
HIS	361
Hoflund syndrome	10
icterus	39
ileitis	60
invagination	59
jejunitis	30
lipidosis	93
obstipation	13
omasitis	18
pancreatitis	2
parasites	13
perforation	35
pharyngitis	9
proctitis	13
reticulitis	98
rumenitic ulcer	4
rumenitis	92
sialoadenitis	2
steatorrhea	105
stenosis	10
stomatitis	57
trauma	25
typhlitis	37
volvulus	479
respiratory	
bronchiolitis	256
bronchitis	466
broncho-pneumonia	1040
laryngitis	23
pharyngitis	1
pleuritis	40
pneumonia	769
rhinitis	11
rhinitis atrophicans	196
sinusitis	8
tracheitis	28
urinary	
congenital abnormality	1
cystitis	56
hydronephrosis	30
nephritis	416
renal degeneration	116
trauma	8
urolithiasis	75
cardio-vascular	
cardiomyopathia	46
congenital abnormality	84
endocarditis	179
epicarditis	62
heart degeneration	50
hydropericard	319
myocarditis	77
pericarditis	427
pleuritis	41
lymphatic	
lymph-adenopathy	245
splenitis	77
tonsillitis	88
musculo-skeletal	
arthritis	231
arthrosis	31
bone degeneration	17
callus	14
congenital abnormality	30
fracture	94
luxation	9
myodegen-eration	70
myopathy	56
osteochondrosis	30
osteomyelitis	136
polyarthritis	343
synovitis	45
tendinitis	25
tendovaginitis	18
reproductive	
abortion	642
congenital abnormality	8
dystocia	9
metritis	70
perforation	4
placentitis	126
retained placenta	3
uterine perforation	4
uterine torsion	3
vaginitis	8
neural	
congenital abnormality	5
encephalitis	116
hydrocephalus	18
intoxication	5
meningitis	226
myelitis	13
neural degeneration	54
neuropathy	78
other	
crushed	81
dermatitis	184
enterotoxemia	285
eye related	22
foreign body	118
hernia	73
hydrothorax	104
inanition	74
intoxication	95
iron deficiency	65
mastitis	66
neoplasia	68
otitis	20
perforation	257
peritonitis	866
pleuritis	643
pododermatitis	19
polyserositis	297
rumen drinker	33
sepsis	647
splenic torsion	18
umbilicus related	117

Table 1: The diagnoses used for classification, grouped by syndrome, with number of occurrences.

defined as consecutive runs of alphanumeric characters or hyphens. We then performed a series of normalization steps in order to reduce the number of term variants when compiling an index.

The bulk of the spelling variation stems from Latin/Greek-originated terms, such as *Zäkum* 'cecum'. Besides the German spelling (using the letters *ä, ö, z/k*), the Latin spelling is often used (*ae, oe, c*, respectively), and even combinations of the two are encountered. For the previous example, the following variants are present, among others: *Caecum, caecum, Cäcum, Cäkum, Zaecum*. We normalized the usage of these letters by replacing *ä* with *ae* and *ö* with *oe* unconditionally, while treating *c* differently based on its right context: before a front vowel it was replaced by *z*, before *h* and *k* it was kept as *c*, and in all other cases (including word-final position) we replaced it with *k*. The complexity of this rule is owed to the fact that this normalization is applied to all words, i. e. including originally German words like *Kinn/Zinn* 'chin'/'tin', which would be confused by an unconditional conflation of *c, k, z*. As a side effect, the normalization of German terms occasionally captured closely spelled English terms (which were not systematically gathered), such as *Enzephalitis/encephalitis*.

Subsequently, we removed inflectional suffixes using the NLTK[1] implementation of the "Snowball" stemmer for German (Porter, 1980). Stemming is the process of removing inflectional and (partially) derivational affixes, thus truncating words to their stems. For example, *minimally* and *minimize* are both reduced to *minim* in Porter's English stemmer, which is not a proper word, but nev-

[1]Natural Language Toolkit: `www.nltk.org`

variants	normalized form	explanation
Zäkumtorsion, Caecumtorsion	zaekumtorsion	*ä/ö/c/k/z* normalization
Kokzidiose, Coccidiose	kokzidios	
Aborts, Abort, Abortes, Aborte, Aborten	abort	stemming
perforierter Ulcus, perforierten Ulkus	perforiert ulkus	both
Kardiomyopathie, Cardiomyopathie, Kardiomyopathien	kardiomyopathi	

Table 2: Normalization examples.

ertheless a useful key for lumping together etymologically related words.

Stemming is based on orthographical regularities and uses only a minimal amount of lexical information. Although the method is not flawless – it may be prone to errors with very short and irregularly inflected words – it generally works well for languages with alphabetic script and has been successfully applied to many European languages. Using a stemmer, we were able to considerably reduce the number of inflectional/derivational variants. However, a number of inflectional forms were still missed by the stemmer – especially plural forms with Latin inflection, such as *Ulkus/Ulzera*, or *Enteritis/Enteritiden*, which are not covered by the stemming rules for general German grammar. The stemmer also failed to capture most of the spelling errors. Table 2 illustrates the conflation with examples.

2.3 Focus Terms

For the syndromic classification of the veterinary reports, we manually created a list of focus terms which served as indicators for the clinical syndromes and diagnoses. Starting from a frequency-ranked list of the words found in all of the reports (already grouped by their normalized form), we manually selected terms that were likely to indicate (positive) diagnoses in the reports. The list was refined by inspecting the reports that produced hits for the focus terms.

The focus terms typically consist of a single token, but we also allowed multi-word expressions. The terms are grouped by diagnosis. Thus, each diagnosis refers to a set of terms which either constitute a common name of the diagnosis or describe some of its aspects. For example, concerning injuries caused by foreign bodies, we consider *Draht* 'wire' and *Nagel* 'nail' as focus terms, even though these words only refer to the cause, but not to the injuries themselves.

As each focus term is represented by its normalized form, a number of variant forms is already matched, as described above. We aimed to additionally cover variants produced by misspellings as well as inflected forms not recognized by the stemmer. Using approximate string matching, we searched the reports for similar terms for each of the focus terms. We used the `simstring` tool (Okazaki and Tsujii, 2010) for retrieving similarly spelled terms among the entire text collection. Approximate matching is a difficult task, as it is hard in general to formally define similarity among (the orthographical representations of) words in a way consistent with human judgement. `simstring` measures similarity as a function of the number of shared n-grams (runs of n characters) in two words, which is only a rough approximation of the task. However, compared to other similarity measures – e. g. Levenshtein's edit distance[2] – it is considerably more efficient for retrieval in large amounts of text. In the inevitable trade-off of good precision and high recall, we strove for recall by choosing a low similarity threshold for retrieval. As expected, this resulted in a high number of hits, including many false positives, i. e. words with a high n-gram similarity score, that are not actually similar to the input term (e. g. *arthritis* and *arteritis*). Due to the limited number of focus terms it was feasible to manually clean the list of similar words.

Figure 1 illustrates how term variants were gathered around the concept of a diagnosis. A number of synonymous and hyponymous terms were added to a specific diagnosis by a human expert. These terms were used as seeds to automatically find more variants, such as inflectional and spelling variants as well as misspellings. Please note that the labeled edges are only added for illustration purposes – the relations between term

[2] For a study of agreement between human judgement and different similarity measures, see e. g. Efremova et al. (2014); for a general overview of similarity measures cf. Navarro (2001) and Christen (2006).

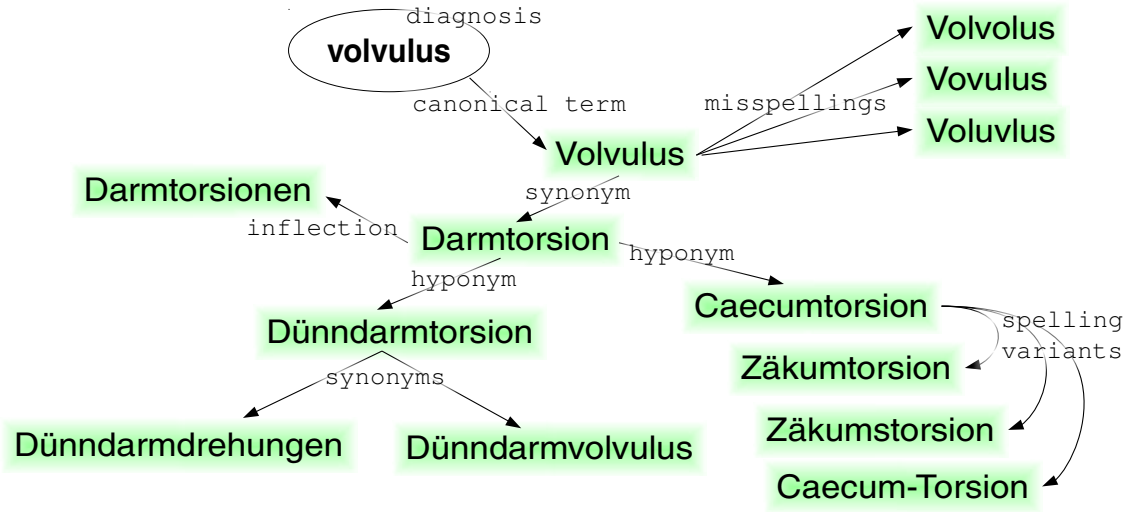


Figure 1: Term variants for the diagnosis *volvulus*.

forms (such as *synonym, misspelling*) were not captured during this phase, as they were not needed for syndrome/diagnosis classification. However, we examined ways to partly recover this underlying structure in an automated way, as is described in the following sections.

2.4 Further Term Conflation

The UMLS Metathesaurus[3] is a large collection of various medical terminology resources. One of its key features is the assignment of unique concept identifiers to entries from different vocabularies in many languages, thus establishing equivalence relations across them. By creating links to Metathesaurus concepts, we can enrich our own terminology resource with information contained in the Metathesaurus, as well as making it more valuable when sharing it with others.

We used the 2014AA release of the Metathesaurus for this work. For each concept that was represented in a German vocabulary, we normalized its lemma and tried to match it against an entry among our focus terms. With this approach, we were able to establish a link to one or more UMLS concepts for 80.6 % of the diagnoses.

Since our data were organized by diagnosis, each covering a number of terms with sometimes quite disparate meanings, the connection to the Metathesaurus produced a high number of one-to-many mappings (cf. Figure 2). This difference in granularity hinders the exploitation of the linked information, as the meaning of many diagnoses appears highly ambiguous in terms of the Metathesaurus. In order to better match the semantic range of the UMLS concepts, we passed on to perform the mapping at the level of terms rather than diagnoses. This required us to add a hierarchical layer to our data structure: We needed to distinguish *term variants* (spelling and inflectional alternations, such as *Caecumtorsion* vs. *Zäkumstorsion*) from *separate terms* (e. g. *Zäkumstorsion* vs. *Darmtorsion*). Please note that synonyms such as *Darmtorsion* and *Darmdrehung* are considered separate terms, even though they have the same meaning.

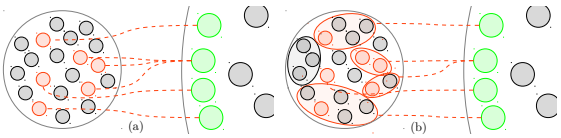


Figure 2: Ontology matching before (a) and after (b) term conflation. In both graphics, the left-hand side represents a diagnosis as a set of terms, some of which are linked to a UMLS concept (connected bullets) on the right-hand side.

For each diagnosis, we organized all term forms into groups of term variants. The arrangement was performed automatically, based on string similarity. While string similarity is only an unreliable approximation of human similarity judgement, and while there are a number of concurring ways of computing it, it is also difficult to determine a

[3] www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

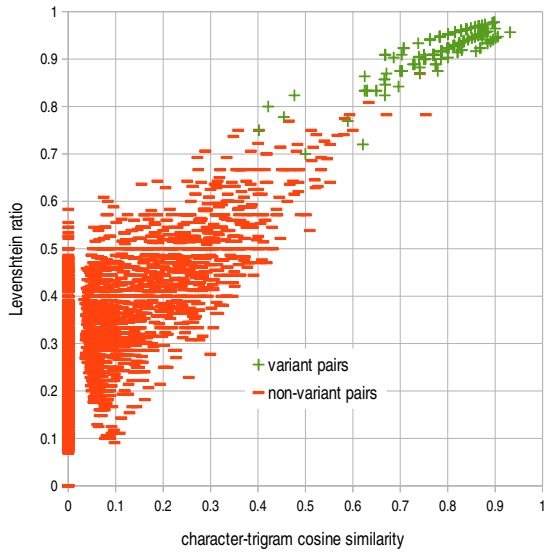


Figure 3: Two different similarity measures for pairs of similar and dissimilar words.

threshold that clearly separates similar from dissimilar pairs of words. We therefore chose to perform supervised machine learning, i. e. automatic learning by example. We compiled a training set of positive instances of inflectional/spelling alternation as well as negative instances, i. e. pairs of unrelated words. For each pair, we computed two different string similarity measures (cf. Figure 3): cosine similarity of character trigram vectors, and Levenshtein ratio. These two measures cover different aspects of similarity, and thus their combination might capture more information than just one of them. We trained a Support Vector Machine on the two-dimensional space of the similarity measures, using a polynomial kernel function.

The automatic term grouping yielded very satisfactory results. We manually evaluated the resulting groups, requiring that all members be orthographical or inflectional variations of each other. We also allowed derivational variants (e. g. *Weissmuskelkrankheit/…erkrankung* ‘white muscle disease’) to be in the same group, although the separation of derivatives (e. g. *Ulkus/Ulzeration*) was not counted as false negative. We found that less than 6.7 % of the groups contained unequal terms (false positives), and only 1.9 % of the groups were erroneously isolated instead of being merged with the correct equivalents (false negatives). Many false positive judgements were caused by terms with only small differences in meaning, such as *Muskeldegeneration* ‘muscle degeneration’ and *Muskelfaserdegeneration* ‘muscle fiber degeneration’, which might even be regarded equal in a less strict evaluation. As for the false negatives, the number of misses could be reduced by extending the stemmer with Latin-inflection endings like *Ulkus – Ulzera*.

2.5 Connecting to UMLS

Each group of term variants was then linked to a UMLS concept if there was a match between at least one member of the group (i. e. a term variant) and of the German concept descriptions, respectively. Only exact agreement of the normalized forms was counted as a match, as preliminary experiments had shown that fuzzy matching introduced a great amount of false positives (connections between similarly spelled, but otherwise unrelated words) while adding only very few desired links. However, we were able to improve the linkage with simple heuristics, such as the removal of boilerplate expressions like *nicht näher bezeichnet* ‘not otherwise specified’.

In 42.1 % of the terms, we could find a match with a UMLS concept. Only 6.7 % of the matching terms point to more than one concept, which means that 93.3 % of the terms with a match can be mapped to the Metathesaurus unambiguously. However, for more than half of the terms no corresponding UMLS concept could be found at all, which is mainly due to the different domains of our veterinary texts and the predominantly human-medicine-based UMLS. Table 3 shows some examples of the mapping.

The connections to the Metathesaurus allowed us to further enrich our data. For example, every UMLS concept has a semantic type assigned to it, such as “Disease or Syndrome” or “Pathologic Function”. Additionally, we used the concept descriptions in Metathesaurus to find more focus terms. By matching the descriptions of connected concepts against our text collection, we were able to enlarge the set of focus terms by almost 10 %.

As next steps, we plan to create links to other widely-used terminology resources, such as the *Central key for health data recording* by the International Committee for Animal Recording (ICAR).[4]

[4] See `www.icar.org`

diagnosis/terms	UMLS
stenosis	
Darmstenose	C0267465 Darmstenose/Darmstriktur/Stenose des…
Dünndarmstenose	C0151924 Dünndarmstenose/Stenose des Dünndarms
Rectumstenose, Rektumstenose	–
myodegeneration	
Belastungsmyopathie	–
Muskelfaserdegeneration, Muskelfaserndegeneration	C0234958 Muskeldegeneration/Degeneration des …
Muskelfasernekrose	–
Muskelläsionen	–
Muskelnekrose, Muskelnekrosen	C0235957 Muskelnekrose/Myonekrose
Myodegeneration, myodegeneration	–
Myonekrose, myonecrosis	C0235957 Muskelnekrose/Myonekrose
Rhabdomyolyse	C0035410 Rhabdomyolyse
Weiss-Muskel-Krankheit, Weiss-Muskelkrankheit, Weissmuskelerkrankung, Weissmuskelkrankeit, Weissmuskelkrankheit	C0043153 Muskeldystrophie, nutritive/ Weißmuskelkrankheit

Table 3: Mapping to the UMLS Metathesaurus.

3 Annotation Tool

The terminology resource described above is a key component in our efforts to create a veterinary surveillance system. We wrote a pipeline of Python scripts that assists our semi-automatic annotation of the pathology reports. The tool performs automatic annotation of syndromes and diagnoses based on the term resource, while also keeping track of manual verfications and rejections. Through a web interface, it accepts a Microsoft Excel workbook as input and produces a modified version in the same format, which allows a veterinary domain expert to inspect and modify the automatic annotations. All relevant information – such as the term resource and the assigned categories, negations (see below), and the previous manual annotations – are contained within this file.

3.1 Negation Detection

In a keyword-based system for detecting evidence, negative expressions can play a crucial role. Occasionally, negative outcomes of an analysis are reported in the texts, and suspected diagnoses are rejected quite frequently, such as *keine Hinweise auf eine Pneumonie* 'no evidence of a pneumonia'. Therefore, we aimed at identifying occurrences of focus terms that are mentioned in a negated context.

Besides the identification of negated expressions, negation detection heavily depends on the correct determination of their scope. Tanushi et al. (2013) compare different approaches to negation scope detection in Swedish clinical reports. According to them, "[e]mploying a simple, rule-based approach with a small amount of negation triggers and a fixed context window for determining scope is very efficient and useful, if results around 80 % F-score are sufficient for a given purpose" (Tanushi et al., 2013, p. 393). We included a simple negation-detection module in our pipeline, which looks for a set of negative expressions in a context window of 5 tokens to either side of the focus term. The context can be restricted for each expression (e. g. only to the right of or only immediately preceding a focus term). The context window is shortened at sentence boundaries and other indicators of a break. However, as the results of the negation detection are not yet satisfactory, we plan to integrate an existing library for this task, e. g. the Python package pyConTextNLP (Chapman et al., 2011).

3.2 Inter-Annotator Agreement

In order to validate the quality of our annotations, we organized a multi-annotator evaluation. We performed an experiment with six experts of veterinary pathology, which were asked to classify a number of reports with respect to the syndromic categories described in Section 2.1. For this purpose, we created a web interface which displayed the report text together with some metadata, one report at a time, and allowed to mark each of the syndromes as present or absent. The reports were randomly sampled, keeping the distribution

syndrome	reports		D_o	D_e	α
gastro-int.	52	(13)	0.059	0.251	**0.764**
respiratory	28	(10)	0.045	0.207	**0.781**
urinary	9	(3)	0.014	0.083	**0.836**
cardio-vasc.	15	(9)	0.041	0.115	**0.644**
lymphatic	3	(3)	0.014	0.018	**0.240**
musc.-skel.	13	(3)	0.014	0.125	**0.891**
reproductive	9	(1)	0.005	0.094	**0.952**
neural	5	(2)	0.009	0.052	**0.825**
other	38	(21)	0.095	0.226	**0.577**
avg.					**0.723**

Table 4: Inter-annotator agreement of the syndromic categories, measured with Krippendorff's Alpha. The second column gives the number of reports where at least one annotator marked the corresponding syndrome as present; following in parentheses is the number of reports with disagreement. D_o and D_e are the observed and expected disagreement, respectively.

of species and year of creation as close to the entire collection as possible (approaching stratified sampling). Each annotator was provided with a sample of 20 reports, which was extended to twice or three times the size when an annotator asked for more. In order to increase sample size, the same report was given to only two or three annotators, rather than all of them. In total, 81 distinct reports were annotated.

We evaluated the inter-annotator agreement with Krippendorff's Alpha (Krippendorff, 2013, pp. 267–309), as is shown in Table 4. For computing the agreement, we regarded each syndrome as an independent, binary variable (each syndrome is either present or absent in a report). The agreement value α ranges from 1 (perfect agreement) to 0 (agreement as by chance) or even below (systematic disagreement). A high agreement means that identifying syndromes is a clear task, while a low agreement indicates that the decisions cannot be easily made. Most of the syndromes have a good (>0.8) or acceptable (>0.6) α score,[5] whereas some are clearly identified as problematic. For the lymphatic system, the sparse representation (only 3 reports) does not allow for valid conclusions; further investigation is required in this case. The "catch-all" class *other*, however, most likely suffers from having an unclear scope. As a consequence of this evaluation, we decided to reduce the ambiguity of *other* by including additional classes in the next revision of the syndromic categorization.

[5]For a discussion of the interpretation of absolute agreement scores see Artstein and Poesio (2008, p. 591)

4 Outlook

We will assess the performance of the text-mining tool based on a small number of diseases which have been relevant in Switzerland in the last 10 years:

1. Bovine Viral Diarrhoea in cattle (an eradication campaign for the disease was introduced in 2008)
2. Porcine Circovirus type 2 infection in pigs
3. Gastro-intestinal syndromes in pigs (for which we observe an increasing amount of pathology submissions)

Time-series analyses will be performed to quantify trends, seasonality and other effects (day of week, day of month etc.) on the number of submissions for syndromes potentially related to these diseases. For each disease, "in-control" data (data collected in the absence of an outbreak) will be used to establish a baseline model describing the amount of normal "noise" in the data (expected number of submissions in the absence of disease outbreaks). Retrospective analyses of the time-series will be done to see whether alerts (signals) were produced when the number of submissions for syndromes potentially linked to the disease was higher than expected from our baseline model (event detection). This will allow us to evaluate whether the system would have worked as an early-warning system.

The tools developed in this project will be adapted to reports from different pathology institutes throughout Switzerland, thus contributing to a nation-wide syndromic surveillance system. Similarly, the methodology developed may be applicable to the analysis of text-based disease information which is recorded in other contexts. For example, there is a great potential of using such a system to systematically analyse health data which are recorded by veterinary practitioners in their practice management software, slaughter data or by animal health services in their central database.

Acknowledgements

This work was funded by the Swiss Federal Food Safety and Veterinary Office (Bundesamt für Lebensmittelsicherheit und Veterinärwesen).

References

R. Michele Anholt, John Berezowski, Iqbal Jamal, Carl Ribble, and Craig Stephen. 2014. Mining free-text medical records for companion animal enteric syndrome surveillance. *Preventive Veterinary Medicine*, 113(4):417–422.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Taxiarchis Botsis, Michael D. Nguyen, Emily Jane Woo, Marianthi Markatou, and Robert Ball. 2011. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *Journal of the American Medical Informatics Association*, 18(5):631–638.

John S. Brownstein, Clark C. Freifeld, Ben Y. Reis, and Kenneth D. Mandl. 2008. Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med*, 5(7):e151.

Wendy W. Chapman, John N. Dowling, and Michael M. Wagner. 2004. Fever detection from free-text clinical records for biosurveillance. *Journal of Biomedical Informatics*, 37(2):120–127.

Brian E. Chapman, Sean Lee, Hyunseok Peter Kang, and Wendy W. Chapman. 2011. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *Journal of Biomedical Informatics*, 44(5):728–737.

Peter Christen. 2006. A comparison of personal name matching: Techniques and practical issues. Technical Report TR-CS-06-02, The Australian National University, Dec.

Nigel Collier, Ai Kawazoe, Lihua Jin, Mika Shigematsu, Dinh Dien, Roberto A. Barrero, Koichi Takeuchi, and Asanee Kawtrakul. 2006. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language Resources and Evaluation*, 40(3-4):405–413.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, et al. 2008. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

Fernanda C. Dórea, C. Anne Muckle, David Kelton, J. T. McClure, Beverly J. McEwen, W. Bruce McNab, Javier Sanchez, and Crawford W. Revie. 2013. Exploratory analysis of methods for automated classification of laboratory test orders into syndromic groups in veterinary medicine. *PLOS one*, 8(3):e57334.

Julia Efremova, Bijan Ranjbar-Sahraei, and Toon Calders. 2014. A hybrid disambiguation measure for inaccurate cultural heritage data. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, page 47–55, Gothenburg, Sweden, April. Association for Computational Linguistics.

Jeff Friedlin, Shaun Grannis, and J. Marc Overhage. 2008. Using natural language processing to improve accuracy of automated notifiable disease reporting. *AMIA Annual Symposium Proceedings*, 2008:207–211.

David Hartley, Noele Nelson, Ronald Walters, Ray Arthur, Roman Yangarber, Larry Madoff, Jens Linge, Abla Mawudeku, Nigel Collier, John Brownstein, Germain Thinus, and Nigel Lightfoot. 2010. Landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, 3(e3).

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Thousand Oaks, CA, 3rd edition.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, March.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, page 851–859, Beijing, China, August.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Fabio Rinaldi, James Dowdall, Michael Hess, Kaarel Kaljurand, Mare Koitand, Kadri Vider, and Neeme Kahusk. 2002. Terminology as knowledge in answer extraction. In *TKE-2002: 6th International Conference on Terminology and Knowledge Engineering*, Nancy, France, August.

Suzanne L. Santamaria and Kurt L. Zimmerman. 2011. Uses of informatics to solve real world problems in veterinary medicine. *Journal of veterinary medical education*, 38(2):103–109.

Kimberly A. Smith-Akin, Charles F. Bearden, Stephen T. Pittenger, and Elmer V. Bernstam. 2007. Toward a veterinary informatics research agenda: An analysis of the PubMed-indexed literature. *International Journal of Medical Informatics*, 76(4):306–312.

Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter, and Roman Yangarber. 2008. Text mining from the web for medical intelligence. In Françoise Fogelman-Soulié, Domenico Perrotta, Jakub Piskorski, and Ralf Steinberger, editors, *Mining Massive Data Sets for Security*, volume 19 of *NATO Science for Peace and Security Series – D: Information and Communication Security*, page 295–310. IOS Press.

Hideyuki Tanushi, Hercules Dalianis, Martin Duneld, Maria Kvist, Maria Skeppstedt, and Sumithra Velupillai. 2013. Negation scope delimitation in clinical text using three approaches: NegEx, PyConTextNLP and SynNeg. In *Proceedings of the*

19th Nordic Conference of Computational Linguistics (NODALIDA 2013), page 387–397, Oslo, Norway.

Michael M. Wagner, J. Espino, F-C. Tsui, P. Gesteland, W. Chapman, O. Ivanov, A. Moore, W. Wong, J. Dowling, and J. Hutman. 2004. Syndrome and outbreak detection using chief-complaint data – experience of the Real-Time Outbreak and Disease Surveillance project. *Morbidity and Mortality Weekly Report*, 53:28–31.

Eva Warns-Petit, Eric Morignat, Marc Artois, and Didier Calavas. 2010. Unsupervised clustering of wildlife necropsy data for syndromic surveillance. *BMC Veterinary Research*, 6(56):1–11.

Acquisition of medical terminology for Ukrainian from parallel corpora and Wikipedia

Thierry Hamon
LIMSI-CNRS, Orsay
U Paris 13
Sorbonne Paris Cité
France
hamon@limsi.fr

Natalia Grabar
UMR8163 STL
CNRS, U Lille 3
Villeneuve d'Ascq
France
natalia.grabar@univ-lille3.fr

Abstract

The increasing availability of parallel bilingual corpora and of automatic methods and tools for their processing makes it possible to build linguistic and terminological resources for low-resourced languages. We propose to exploit various corpora available in several languages in order to build bilingual and trilingual terminologies. Typically, terminology information extracted in French and English is associated with the corresponding units in the Ukrainian corpus thanks to the multilingual transfer. According to the used approaches, precision of the term extraction varies between 0.454 and 0.966, while the quality of the interlingual relations varies between 0.309 and 0.965. The resource built contains 4,588 medical terms in Ukrainian and their 34,267 relations with French and English terms.

1 Introduction

The acquisition of terminology has gone through a very active period and provides nowadays several automatic tools and methods (Kageura and Umino, 1996; Cabré et al., 2001; Pazienza et al., 2005) for several European languages and Japanese. Nevertheless, other languages remain low-resourced and require specific Natural Language Processing (NLP) developments.

Our main objective is to create terminological resources for Ukrainian, for which very little digitized or electronic resources are available. Yet, the terminology extraction tools usually require the morpho-syntactic tagging of texts, which can be problematic if the corresponding automatic tools are not available for a given language. For instance, the UGtag Part-of-Speech (POS) tagger (Kotsyba et al., 2009) developed for Ukrainian does not perform the syntactic and morphological disambiguation of the tags. Hence, it becomes impossible to use it for the pre-processing of corpora before the traditional terminology acquisition process.

In this situation, we propose first to compile terminological resources for Ukrainian in order to build the basis for the observation of the specificities of terminological units in this language. Such observations will allow to develop and parameter the terminology extraction tool for Ukrainian.

The motivation of our work is double. We want to

1. automatically build terminologies for Ukrainian,

2. design specific methods for the acquisition of such terminological resources.

The work is carried out with medical data, and in three languages (Ukrainian, French, and English).

The work we present starts from the exploitation of two kinds of corpora (Section 2.1): Wikipedia in Ukrainian which provides several useful kinds of information (such as term labels and their codes) with a high level of quality, and the parallel corpus MedlinePlus. The term detection and extraction can be either manual or automatic. Since, there is no appropriate POS-tagging and term extraction tools for Ukrainian, we propose to use such tools in French and English, and to take advantage of these to transfer English and French extracted terms on the Ukrainian corpus.

Indeed, the transfer methodology can be considered as suitable for such objectives. Suppose we

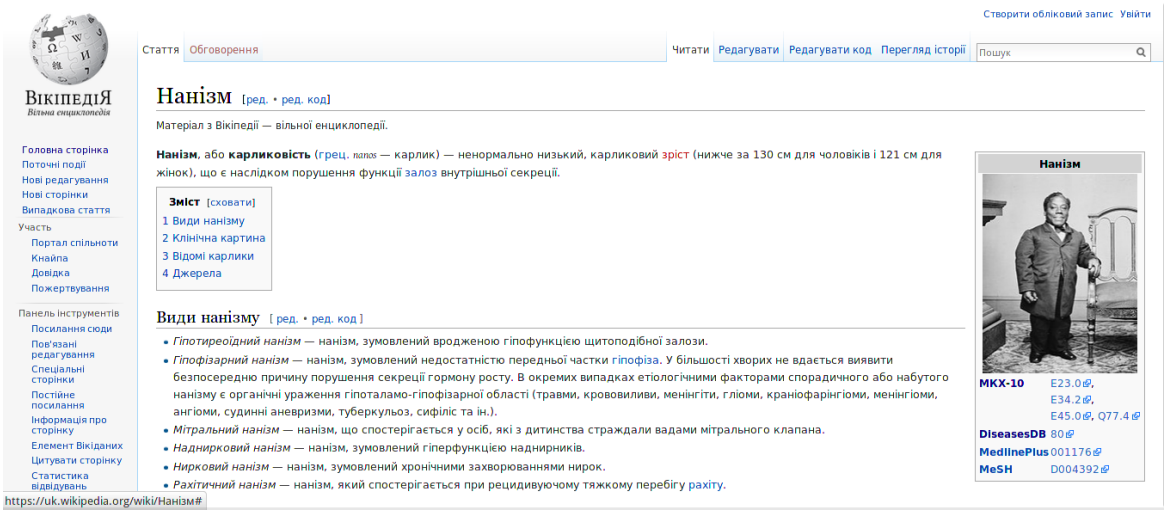


Figure 1: Example of the Ukrainian Wikipedia source pages (Dwarfism). The infobox with the coding is on the right.

have parallel and aligned corpora with two languages $L1$ and $L2$, and we have several types of syntactic or semantic annotations and information associated to $L1$. The transfer approach permits to transpose these annotations or information from $L1$ to $L2$, and to obtain in this way the corresponding annotations and information in the $L2$ text. From this point of view, $L1$ is considered as the source language while $L2$ is considered as the target language. This kind of approach is particularly interesting when working with low-resourced languages for which less tools and semantic resources are available. An increasing availability of parallel bilingual corpora, and of automatic methods and tools for their processing makes it possible to build linguistic and terminological resources using the transfer methodology (Yarowsky et al., 2001; Lopez et al., 2002). Very few works have been done in this direction, and we assume they open novel and efficient ways for the processing of multilingual texts in particular from low-resourced languages (Zeman and Resnik, 2008; McDonald et al., 2011). Notice that the modeling of cross-language features aims at using language-independent features to create various types of annotations. Among such features, we can mention part-of-speech, semantic categories or even acoustic and prosodic features.

We propose to apply this method for the acquisition of bilingual or trilingual terminologies involving Ukrainian. In our work, each corpus is exploited through dedicated methods. The MedlinePlus corpus provides the basis for the building of the terminology, while the Wikipedia corpus permits to enrich this information and helps the word-level alignment of the MedlinePlus corpus.

Terminology-related research on Ukrainian is an active area, although the main terminological work shows mainly theoretical and linguistic orientation (Коссак, 2000; Dmytruk, 2009; Рожанківський and Кузан, 2000; Ivashchenko, 2013; Oliinyk, 2013). Very few works are oriented on the use of terminologies and their automatic processing, such as the software localization (Shyshkina et al., 2010).

In the following of this paper, we first present the material used for the acquisition of bilingual terminology (section 2), and the methods designed for achieving this objective (section 3). We then discuss the results we obtain (section 4), and conclude with directions for the future work (section 5).

2 Material

2.1 Corpora

We use two kinds of corpora:

- *MedlinePlus:* parallel medical corpus from MedlinePlus. These data are built by MedlinePlus from the National Library of Medicine[1]. They contain patient-oriented

[1] *www.nlm.nih.gov/medlineplus/healthtopics.html*

Corpus	*Size (occ of words)*
Wikipedia/UKmed	246,368,411
MedlinePlus/UK	43,184
MedlinePlus/FR	53,067
MedlinePlus/EN	46,544

Table 1: Size of the exploited corpora.

brochures on several medical topics (body systems, disorders and conditions, diagnosis and therapy, demographic groups, health and wellness). These brochures have been created in English and then translated in several other languages, among which French and Ukrainian;

- *Wikipedia:* medicine-related articles from Wikipedia. This corpus is extracted from the Ukrainian part of the Wikipedia using medicine-related categories, such as *Медицина* (*medicine*) or *Захворювання* (*disorders*). The corpus potentially covers a wide range of medical notions. In Figure 1, we indicate an example of the source pages which propose the navigation frame on the left, the text with explanations and the infobox with illustration and coding on the right.

In Table 1, we indicate the size of the corpora. Not surprisingly, the Wikipedia corpus is much larger although only part of its information is exploited, as we will see in the next section.

2.2 UMLS: Unified Medical Language System

The UMLS (Unified Medical Language System) (Lindberg et al., 1993) merges several (over 100) biomedical terminologies, such as international terminologies MeSH (NLM, 2001) and ICD (Brämer, 1988). Such international terminologies may exist in several languages. For instance, French and English versions of MeSH are included in the UMLS. No terminologies in Ukrainian are part of the UMLS. Each UMLS term is provided with unique identifiers, which allows to find the corresponding terms in other terminologies or languages.

3 Methods

The methods we propose for the extraction of bilingual terminology are adapted to each kind of corpora and of data they contain: the MedlinePlus corpus (section 3.1) and the Wikipedia corpus (section 3.2). We then present their cross-fertilization (Section 3.3), and the evaluation of the results (Section 3.4).

3.1 Extraction of bilingual terminology from the MedlinePlus corpus

Prior to the exploitation of the MedlinePlus data, the documents are first transformed in a suitable format:

- the source PDF documents are converted in the text format;
- in each language, the documents are segmented in paragraphs;
- the alignments French/Ukrainian and English/Ukrainian are generated, in which n_{th} paragraph from one language is associated with the n_{th} paragraph from the other language;
- the alignment between the two pairs of languages is then verified manually.

In Figure 2, we present an excerpt from the English/Ukrainian aligned corpus.

Then in French and English, we can use the existing terminology extraction tools which results bootstrap the acquisition of bilingual terminology. Hence, we use the YATEAterm extractor (Aubin and Hamon, 2006), that is applied to documents POS-tagged. The extracted terms are then projected on the French and English corpora. In Figure 2, candidate terms are marked in bold.

The exploitation of the MedlinePlus parallel and aligned corpus is performed in several ways (Figure 3).

Transfer 1 First, the simplest situation is when the two aligned lines contain term candidates in either language: these terms are recorded as candidates for the alignments. For instance, in Figure 2, the pairs {*Tiredness*, *Втома*} and {*Pain*, *Біль*} are issued from this kind of alignment.

Transfer 2 Secondly, when the paragraphs contain complex expressions or sentences, the processing is done as follows (Figure 4):

1. the paragraph-aligned corpora are aligned at the word level using GIZA++ (Och and Ney, 2000),

English	*Ukrainian*
Cancer cells grow and divide more quickly than **healthy cells**. **Cancer treatments** are made to work on these **fast growing cells**.	**Ракові клітини** ростуть і діляться швидше, ніж **здорові клітини**. При **лікуванні раку** здійснюється вплив на ці **клітини, що швидко ростуть**.
- **Tiredness**	- **Втома**
- **Nausea** or **vomiting**	- **Нудота** або **блювота**
- **Pain**	- **Біль**
- **Hair loss** called **alopecia**	- **Втрата волосся**, що називається **алопецією**

Figure 2: Example of the paragraph-aligned MedlinePlus corpus (English/Ukrainian).

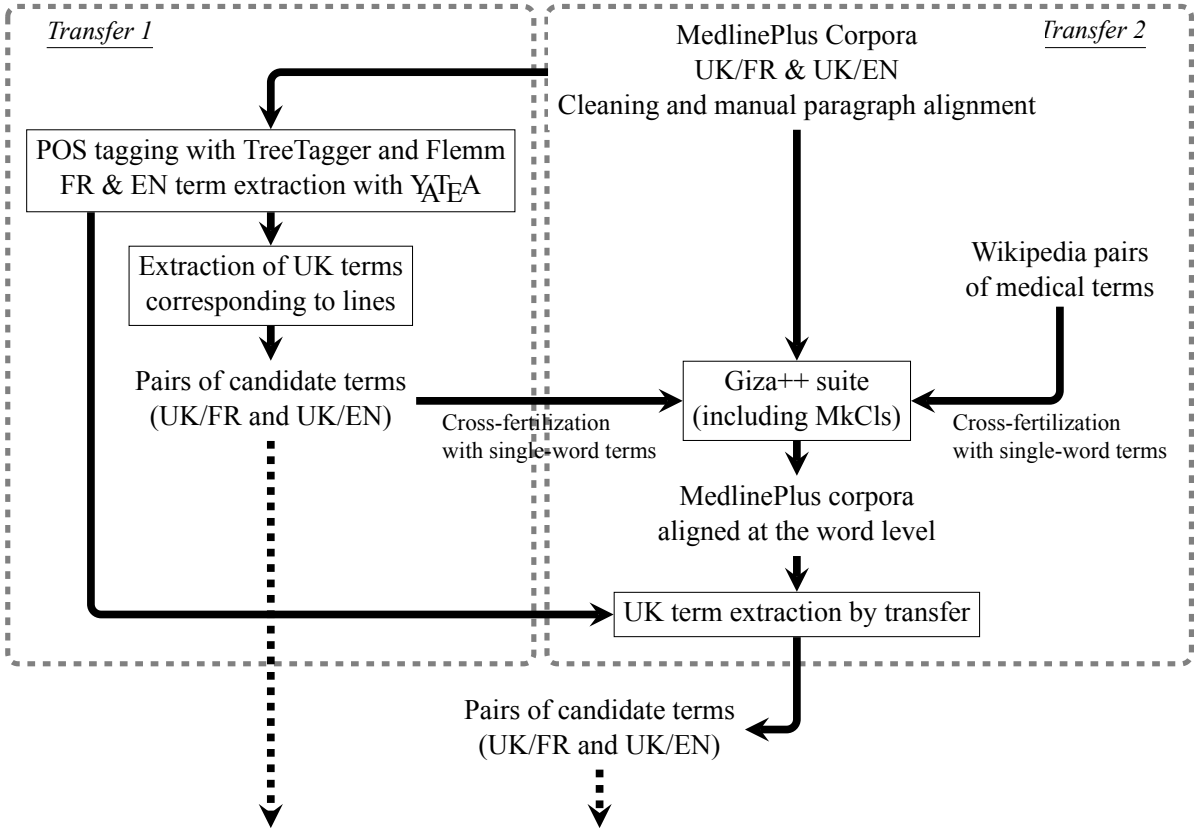


Figure 3: Extraction of medical terms from MedlinePlus corpora (Ukrainian = UK, French = FR, English = EN).

2. using the word-aligned corpora, in each paragraph pair (French/Ukrainian and English/Ukrainian), the terms recognized in French and English are transferred on the Ukrainian paragraph (conceived as the target language);

3. the alignments extracted are recorded as candidates for building the bilingual terminology.

For instance, in Figure 2, the term *Cancer cells* is automatically extracted from the English corpus. GIZA++ proposes that *Cancer cells* is aligned with *Ракові клітини*. Thus, through the word-aligned text, we can propose that *Cancer cells* is the translation of *Ракові клітини*. This processing is performed on the two pairs of languages (French/Ukrainian and English/Ukrainian).

As indicated in Table 1, the size of our corpora is rather small for the statistical alignment performed

by GIZA++. For this reason, we provide GIZA++ with a bilingual dictionary in order to help the alignment at the word level (see Section 3.3). Besides, in preliminary experiments, we also observe that word level alignment errors lead to the extraction of Ukrainian stopwords as term candidates (*на (on)*, *або (or)*, etc.). To remove such obvious errors, we filter out such candidates if they occur in a list of 385 stop-word forms issued from an existing resource dedicated to the localization of graphical interfaces[2].

3.2 Extraction of bilingual terminology from the Wikipedia corpus

The Wikipedia corpus is used to complete and to help the method applied to the MedlinePlus corpus. The content we propose to exploit is included in infoboxes (on the right in Figure 1) and is reachable through the MediaWiki source code of the Wikipedia. This provides the label of the medical terms in Ukrainian and their MeSH codes. The process is the following (Figure 4):

1. the infobox content is extracted and parsed[3] in order to obtain the term label and its MeSH code,
2. the MeSH code is used to query the UMLS, and to get the corresponding French and English terms,
3. the term pairs French/Ukrainian and English/Ukrainian are then built and provide good candidates for the bilingual terminology.

This part of the method exploits specific and intentionally created content for a given medical notion in Ukrainian: term for a given medical notion and its MeSH code. This information is reliable. For instance, in Figure 1, the term *нанізм* is extracted, as well as its MeSH code D004392. Through the UMLS, the corresponding English terms are *dwarfism* and *nanism*, while the corresponding French term is *nanisme*. Notice that similar method has been used for the building of medical terminology in the Arabic language (Vivaldi and Rodríguez, 2014).

[2] `https://github.com/fluxbb/langs/blob/master/Ukrainian/stopwords.txt`

[3] We use the Perl module `Text::MediawikiFormat` (`http://search.cpan.org/\~szabgab/Text-MediawikiFormat`)

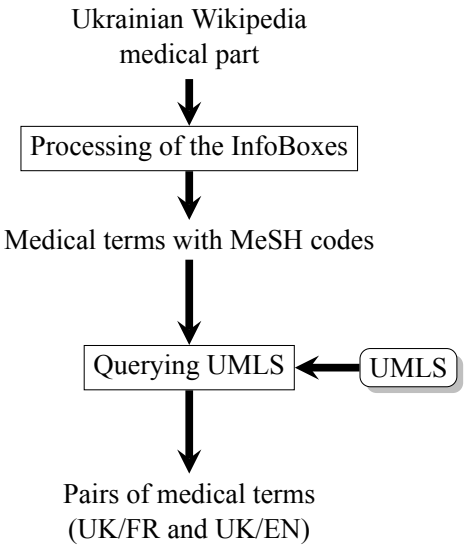


Figure 4: Extraction of medical terms from Wikipedia (Ukrainian = UK, French = FR, English = EN).

3.3 Cross-fertilization and Experiments

The cross-fertilization of the two methods (Sections 3.1 and 3.2) is done in two ways:

- the Wikipedia terms are used to enrich the extracted terminology,
- the single-word terms extracted by other approaches can be provided to GIZA++, as an additional bilingual dictionary, in order to help the alignment of MedlinePlus at the word level.

During preliminary experiments, we test several combinations of parameters for the pre-processing and the alignments. While pre-processing the French corpus, the Part-of-Speech is performed by TreeTagger (Schmid, 1994) and can be improved by the morphological analyzer Flemm (Namer, 2000). We also experiment with the use of GeniaTagger (Tsuruoka et al., 2005) on the English corpus. We also experiment with the use of the terms extracted from Wikipedia, or by the MedlinePLus method Transfer 1, or both, for guiding the Giza++ alignment.

Thus, based on the results of the preliminary experiments, we choose to pre-process the English corpus with TreeTagger and the French corpus with TreeTagger and Flemm. Single-word terms extracted from Wikipedia and by the method Transfer 1 are used as bilingual dictionary to help

the Giza++ word level alignment. We only present the results obtained with this configuration in the following.

3.4 Evaluation

The evaluation is performed manually in order to check whether the candidates extracted for building the bilingual terminologies are correct. It has been performed by an Ukrainian native speaker having knowledge in medical informatics. Terms are validated independently in each language, but we also evaluation the bilingual and trilingual relations between the Ukrainian, English and French terms. With this kind of evaluation, precision of the results can be computed, i.e. the ratio between the correct answers and all the answers.

4 Results and Discussion

Table 2 presents the results and the precision for the extracted terms by the three methods. Table 3 presents the results and the precision concerning the pairs and triples of terms.

4.1 Extraction of bilingual terminology from the Wikipedia corpus

The exploitation of the Wikipedia infobox allow to collect 357 Ukrainian medical terms among which 177 are single-word terms. By querying UMLS with the MeSH codes, those terms are associated with 1428 French terms (among them, 339 single-word terms) and 3625 English terms (among them, 448 single-word terms). The number of French and English terms compared to the number of Ukrainian terms are due to the synonyms proposed by MeSH. As for the bilingual pairs of terms, we obtain 1,515 Ukrainian/French term pairs and 3,789 Ukrainian/English term pairs, including, respectively, 270 and 405 pairs between single-word terms. Since each Ukrainian term is associated with at least one French and English terms, this allows to build 28,840 triples. We consider that the precision of this terminology is 1 because the collecting manner.

4.2 Extraction of bilingual terminology from the MedlinePlus corpus

The use of the first method of transfer (Transfer 1) allows to extract 436 Ukrainian terms with a high precision unsurprisingly (0.966). These terms are associated with 316 French terms and 354 English terms in 282 triples between Ukrainian, French and English terms, 63 pairs only between Ukrainian and French terms and 115 pairs only between Ukrainian and English terms, with 0.954, 0.937 and 0.965 precision, respectively. Thus, the Transfer 1 method allows to collect 334 Ukrainian/French term pairs (among them 108 pairs between single-word terms) and 380 Ukrainian/English term pairs (among them 135 pairs between single-word terms). We observe that these relations can involve synonyms in either language: {*фаллопієва труба*, *trompes de fallope/trompe utérine*} (*fallopian tube*), {*втрата слуху/втрачається слух*, *hearing loss*}, {*втома*, *fatigue/tiredness*}. Besides, in Ukrainian, several case forms can be associated to a same English of French form: {*вагітність*, *pregnancy*} and {*вагітності*, *pregnancy*}.

As the precision values suggest, this first transfer method leads to few errors. Their analysis shows that they mainly concern partial match between one language and another involved by the translation: {*появу виразок у роті*, *mouth sores*} -- lit. *(appearance of) mouth sores*, {*ви можете спати*, *dormir/sleep*} -- lit. *you can sleep*. The silence of the method can be explained by two reasons. First, again the variation due to the translation prevents the transfer 1 method to extract term in French or English. For instance, since the title *Soins* in the French corpus is the English translation of *Your care*, the French term matches with the line, contrary to the English term. The Transfer 2 method will solve this problem. However, the main reason of the silence is the incapacity of the term extractor to identify French or English terms because its extraction strategy or errors in the POS tagging.

As for the second transfer method (transfer 2), we present the results obtained when the pairs of single-words terms issued from the MedlinePlus corpus and from Wikipedia are used to help the GIZA++ alignment. In that context, the transfer 2 method allows to extract 9,040 Ukrainian terms with 0.454 precision (exact match). Precision of the French and English terms is higher: 0.674 and 0.761 respectively (exact match). Moreover, the number of French and English terms is dramatically lower (about -45% and -40%) than in Ukrainian: the rich morphology of the Ukrainian language provides several inflected

Source	*UK*		*FR*		*EN*	
	#terms	Prec.	#terms	Prec.	#terms	Prec.
Wikipedia	357	1	1,428	1	3,625	1
MedlinePlus$_{Transfer1}$	436	0.966	316	0.971	354	0.989
inexact match		0.998		0.987		0.997
MedlinePlus$_{Transfer2}$	9,040	0.454	3,671	0.674	3,597	0.761
inexact match		0.84		0.726		0.799
Total	9,529	0.481	5,200	0.769	7,335	0.883
Total of correct terms	4,588		3,998		6,476	

Table 2: Number of terms extracted (Ukrainian = UK, French = FR, English = EN).

Source	*UK/FR*		*UK/EN*		*UK/FR/EN*		Total	
	#rel.	Prec.	#rel.	Prec.	#trpl.	Prec.	#trpl.	Prec.
Wikipedia	1,515	1	3,789	1	28,840	1	28,840	1
MedlinePlus$_{Transfer1}$	63	0.937	115	0.965	282	0.954	460	0.954
inexact match		0.984		1		0.982		0.987
MedlinePlus$_{Transfer2}$	3,724	0.309	4,745	0.401	4,724	0.419	13,218	0.381
inexact match		0.751		0.84		0.586		0.724
Total	3,798	0.318	4,819	0.41	33,845	0.918	42,462	0.807
Total of correct relations	1,207		1,974		31,086		34,267	

Table 3: Number of term pairs and triples (Ukrainian = UK, French = FR, English = EN).

forms for a given term ({*напад*, *нападу*} -- *attack*, {*припадків*, *припадки*} -- *seizure*, {*костей*, *кістки*} -- *bones*). Besides, the method allows also to extract synonymous terms ({*приступам*, *припадків*} -- *attacks/seizures*, {*биття*, *удару*} -- *beats*). The precision values with the inexact match (the correct term is included or includes the term candidates) are much higher and gain 0.40 points for the Ukrainian terms and 0.05 for the French and English terms. We assume this difference on Ukrainian candidate terms is mainly due to the alignment quality. As for the interlingual relations, the Transfer 2 method collects 3,724 pairs of Ukrainian/French terms with 0.309 precision, 4,745 pairs of Ukrainian/English terms with 0.401 precision and 4,724 triples with 0.419 precision.

An analysis of the results shows that most of the errors are due to the alignment problems. Indeed, we observe that when the alignment is correct, the Ukrainian terms are correctly extracted by the transfer. Otherwise, the errors occur.

Moreover, even if the documents (patient-oriented brochures) are not highly specialized, most of the extracted terms are specific to the medical domain ({*трахеотомією*, *tracheostomy*}), {*фактори ризику*, *risk factors*}, {*шприца*, *syringe*}, {*холестерину*, *cholesterol*}). Other terms also refer to close and approximating notions which reflects this type of documents: {*діти*, *children*}, {*здорову їжу*, *healthy diet*}, {*серцевий напад*, *heart attack*}, {*склянок рідини*, *glasses of liquid*}. An interesting observation is that some French and English terms correspond to propositions in Ukrainian: {*не до кінця приготовлену їжу*, *undercooked foods*} (lit. food which is not fully cooked), {*При цьому обстеженні Ви не відчуєте жодного болю*, *indolore (painless)*} (lit. With this exam you will feel no pain).

Finally, all the methods combined allow to build a terminological resource containing 4,588 Ukrainian medical terms and their 34,267 relations with French and English terms.

5 Conclusion and Future Work

In this work, we propose to exploit two kinds of freely available multilingual corpora in French, English and Ukrainian. Each corpus is exploited with appropriate methods which allows to extract the term candidates and to create term pairs Ukrainian/French and Ukrainian/English. In particularly, French and English corpora are processed with NLP and term extraction tools. Then,

thanks to the transfer methods these terms are transposed on the Ukrainian language. We also propose to use existing terminologies and to exploit simple terms for improving the alignment performed at the word level with GIZA++.

Our future work will address the enrichment of the created resource with terms from other corpora. Besides, in the Wikipedia corpus, we can use other codes, such as those from MKX-10 (ICD10) or MedlinePlus. This will also augment the coverage of the term pairs extracted in the current work. Another perspective of this work is the improvement of the bilingual alignment of documents at the word level. In that respect, we plan to investigate the use of other alignment algorithms, such as Fast-Align (Dyer et al., 2013) or the Lingua::Align toolbox (Tiedemann and Kotzé, 2009). Other curators will be involved. Further improvements of the proposed transfer method can be obtained with statistical and morphological cues.

Acknowledgments

This work is funded by the LIMSI-CNRS AI project *Outiller l'Ukranien*. We are thankful to the reviewers for their useful comments which permitted to improve the quality of the paper.

References

S Aubin and T Hamon. 2006. Improving term extraction with terminological resources. In *FinTAL 2006*, number 4139 in LNAI, pages 380--387. Springer.

GR Brämer. 1988. International statistical classification of diseases and related health problems. tenth revision. *World Health Stat Q*, 41(1):32--6.

MT Cabré, R Estopà, and J Vivaldi, 2001. *Automatic term detection: a review of current systems*, pages 53--88. John Benjamins.

Veronica Dmytruk. 2009. Typological features of word-formation in computing, the internet and programming in the first decade of the XXI century. In *УДК*, pages 1--11.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL/HLT*, pages 644--648.

VL Ivashchenko. 2013. Historiography of terminology: metalanguage and structural units. In *UDC*, pages 1--22.

K Kageura and B Umino. 1996. Methods of automatic term recognition. In *National Center for Science Information Systems*, pages 1--22.

Natalia Kotsyba, Andriy Mykulyak, and Ihor V. Shevchenko. 2009. Ugtag: morphological analyzer and tagger for the ukrainian language. In *Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009)*.

DA Lindberg, BL Humphreys, and AT McCray. 1993. The unified medical language system. *Methods Inf Med*, 32(4):281--291.

Adam Lopez, Mike Nossal, Rebecca Hwa, and Philip Resnik. 2002. Word-level alignment for multilingual resource acquisition. In *LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Data*, Las Palmas, Spain.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62--72, Stroudsburg, PA, USA. Association for Computational Linguistics.

F Namer. 2000. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues (TAL)*, 41(2):523--547.

National Library of Medicine, Bethesda, Maryland, 2001. *Medical Subject Headings*. *www.nlm.nih.gov/mesh/meshhome.html*.

FJ Och and H Ney. 2000. Improved statistical alignment models. In *ACL*, pages 440--447.

OY Oliinyk. 2013. Terminology for description of linguistic landscape in native and foreign linguistics. *Terminolohichnyi visnyk*, 2(1):1--7.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology extraction: An analysis of linguistic and statistical approaches. In Spiros Sirmakessis, editor, *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*, pages 255--279. Springer Berlin Heidelberg.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44--49.

Nataliia Shyshkina, Galina Zorko, and Larisa Lesko. 2010. Terminology work and software localization in Ukraine. In *Problems of Cybernetics and Informatics*, pages 17--20.

Jörg Tiedemann and Gideon Kotzé. 2009. A discriminative approach to tree alignment. In Iustina Ilisei, Viktor Pekar, and Silvia Bernardini, editors, *Proceedings of the International Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography and Language Learning (in connection with RANLP'09)*, pages 33 -- 39.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust

part-of-speech tagger for biomedical text. *LNCS*, 3746:382--392.

J Vivaldi and H Rodríguez. 2014. Arabic medical term compilation from Wikipedia. In *Proc of CIST 2014*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT*.

D Zeman and P Resnik. 2008. Cross-language parser adaptation between related languages. In *NLP for Less Privileged Languages*.

Орест Коссак. 2000. Українська комп'ютерна термінологія. In *Сучасні проблеми в комп'ютерних науках*, pages 39--42.

Р Рожанківський and М Кузан. 2000. Комп'ютерні проблеми стандартизації термінології. In *Сучасні проблеми в комп'ютерних науках (CCU'2000)*, pages 42--44.

Terminology acquisition and description using lexical resources and local grammars

Cvetana Krstev
University of Belgrade
cvetana@matf.bg.ac.rs

Ranka Stanković
University of Belgrade
ranka@rgf.bg.ac.rs

Ivan Obradović
University of Belgrade
ivan.obradovic@rgf.bg.ac.rs

Biljana Lazić
University of Belgrade
biljana.lazic@rgf.bg.ac.rs

Abstract

Acquisition of new terminology from specific domains and its adequate description within terminological dictionaries is a complex task, especially for languages that are morphologically complex such as Serbian. In this paper we present an approach to solving this task semi-automatically on basis of lexical resources and local grammars developed for Serbian. Special attention is given to automatic inflectional class prediction for simple adjectives and nouns and the use of syntactic graphs for extraction of Multi-Word Unit (MWU) candidates for termbases, their lemmatization and assignment of inflectional classes.

1 Introduction

In this paper we present a semi-automatic procedure for terminology acquisition in Serbian. Rapid changes in many knowledge domains mean that new terms are continuously being created and introduced in Serbian making important the automation of their retrieval and incorporation in Serbian terminological dictionaries. Due to specific features of Serbian grammar, especially its rich morphology, this is a complex task, and corresponding language resources in the form of morphological e-dictionaries and grammars need to be applied (Vitas et al., 2012). For that reason, in the case of Serbian, it is not enough to extract terminology from the domain, but it also has to be adequately described, for instance, in the form of e-dictionaries.

The field of terminology is strongly related to research on multiword terms, which relates closely to MWEs (Baldwin & Kim, 2010; Frantzi et al., 2000). An analysis of terms from technical dictionaries for different domains (fiber optics, medicine, physics and mathematics, psychology) showed that 97% of multi-words in these sources consist of nouns and adjectives only, and more than 99% consist only of nouns, adjectives, and a preposition. (Justeson & Katz, 1995) Identifying the adjectives and the prepositional phrase is thus important for terminology acquisition (Daille, 2000).

There are two mainstream approaches (Enguehard & Pantera, 1995; Cerbah & Daille, 2007) to terminology acquisition. One relies on using statistical measures (Nakagawa & Mori, 2003; Ramisch et al., 2012; Quochi et al., 2012; Zhang et al., 2006) and the other is based on linguistic rules. A rule-based approach for the extraction of terms based on a cascade of transducers using CasSys tool incorporated in Unitex[1] corpus processing platform, as well as the use of TMF standard for the representation of terms is proposed in (Ammar et al., 2015) and applied on Arabic scientific and technical corpus. In (Savary et al., 2012) terminology extraction in the domain of ecsonomy is presented for Polish. It has two modules: a grammatical lexicon of terminological MWEs and a fully lexicalized shallow grammar, obtained by an automatic conversion of the lexicon. Przepiorkowski and associates (2007) present results of automatic extraction of term definitions from unstructured texts in Bulgarian, Czech and Polish by use of regular grammars.

There are also combinations of the two approaches (Rodrıguez et al., 2007). Sag et al. reported that modern statistical Natural Language Processing (NLP) is in great need of better language models and linguistic tools must come to

[1] Corpus processing System Unitex: http://www-igm.univ-mlv.fr/~unitex/

grip with problems of disambiguation and MWUs (Sag et al., 2002).

2 Process description

The processing steps (Fig.1) of integrating new terms from specific domains in terminological dictionaries using lexical resources and local grammars in our approach are:

1. Linguistic preprocessing of the input plain text file from the chosen domain using Unitex.
2. Analysis of unrecognized words as the most probable source of terminology and expanding the dictionary of simple words:
   - 2.1 Retrieval of unrecognized words;
   - 2.2 Manual filtering, preparation of a list of extracted terms in canonical forms (for instance, nominative singular for nouns) and annotating with semantic labels (e.g. human) and some grammatical categories (e.g. adding the gender for the nouns);
   - 2.3 Automatic prediction of the inflectional class and the production of dictionary entry in DELA format (detailed description of the algorithm is given in section 3);
   - 2.4 Compiling the dictionaries of newly acquired terms and integrating them with other resources for linguistic text processing;
   - 2.5 Repeated linguistic preprocessing with expanded dictionaries for verification of recognition of new lemmas.
3. MWUs extraction
   - 3.1. Application of syntactic graphs to extract MWUs with different syntactic structures from the same text (detailed description of the algorithm is given in section 4);
   - 3.2. Removing duplicate extractions: if a sequence of words is recognized with different graphs as having different syntactic structures the most probable candidate is chosen according to the pre-established order of precedence;
   - 3.3. Two-step generation of MWU canonical forms: in the first step lemmatization of simple words that form the MWU is performed, while in the second step the lemma of the MWU is produced based on the results from step 1.
4. Selection of terms from new MWUs
   - 4.1. Frequency calculation for all forms of MWUs and their basic forms with ranking of results;
   - 4.2. Removing MWUs already in e-dictionaries and those with rank under the specified threshold;
   - 4.3. Linguistic evaluation of grammatical correctness of remaining MWUs;
   - 4.4. Assessment of domain relevance of each MWU by comparing its frequency in the domain text with its frequency in the Corpus of Contemporary Serbian (Utvić, 2014).
5. Expanding MWU dictionaries
   - 5.1. Creation of complete MWU lemmas in compliance with DELAC format (Savary, 2009);
   - 5.2. Compiling the dictionaries of newly acquired multi-word terms and integrating them with other resources for linguistic text processing;
   - 5.3. Linguistic pre-processing with expanded dictionaries for verification of recognition of new MWU lemmas.

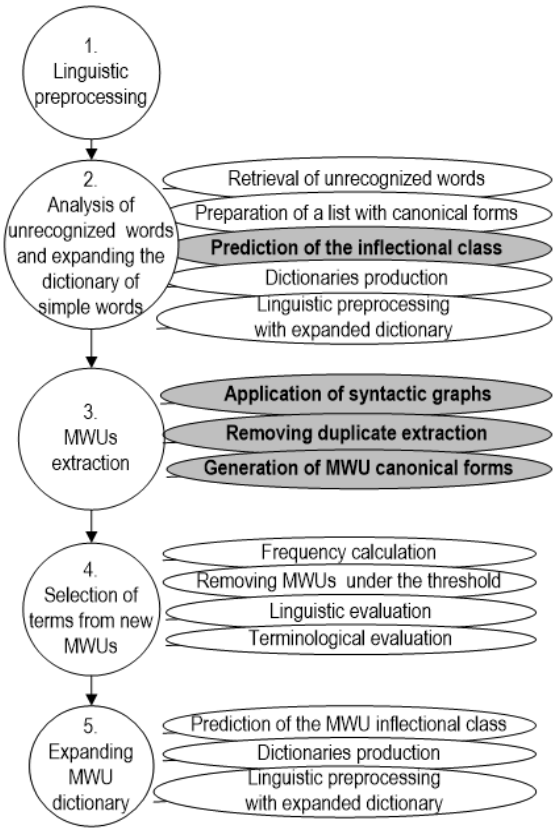


Figure 1: Diagram of terminology acquisition using lexical resources and local grammars

The newly acquired terms, both simple and MWU, can be exported to termbases, TBX and other standard formats for terminological resources. In this paper we will focus on (marked

gray in Figure 1): inflectional class prediction (step 2.3) and extraction of MWU candidates for termbases using syntactic graphs (step 3).

3 Prediction of inflectional class for simple words

Prediction of inflectional class for a new word in Serbian is not an easy task because of complex inflectional grammar with numerous rules and exceptions. Morphological electronic dictionaries of Serbian for NLP are being developed for many years now. Their development follows the methodology and format (known as DELAS/DELAF) presented for French in (Courtois, 1990). E-dictionaries in the same format have been produced for many other languages.

In dictionary of lemmas (DELAS) each lemma is described in full detail so that a dictionary of forms containing all necessary grammatical information (DELAF) can be generated from it, and subsequently used in various NLP tasks.

Serbian e-dictionaries of simple forms have reached a considerable size: they have about 135,000 lemmas generating more than 5 million forms and 13,000 compound lemmas, that is, multi-word units (Krstev, 2008). The number of simple lemmas by Part-Of-Speech (POS) is depicted in Figure 2 (left).

POS	lemmas		FSTs	
Nouns	81,866	61%	372	44%
Verbs	17,071	13%	372	44%
Adjectives	31,071	23%	69	8%
Other	4,632	3%	41	5%
Total	134,640		854	

Figure 2: Statistics of lemmas and inflectional FSTs

Inflectional classes are described with metadata including most important features for class distinction e.g. for nouns grammatical gender and number, case, and animateness are given.

Grammatical inflectional rules are encoded by 854 inflectional Finite-State Transducers (FST) Inflectional FSTs are a special kind of FSTs used for modeling inflectional paradigms, that is, inflectional classes. Each FST of this kind is used for production of all inflected forms for all lemmas belonging to the same class. The number of Inflectional FSTs by POS is depicted in Figure 2 (right).

Productiveness of all inflectional classes are not the same: some classes are used for a large number of regular cases, while other pertain to (rare) exceptions. Our approach is addressing the first group, having in mind that terminology usually inflects regularly. Figure 3 presents the number of inflectional classes and percent of lemmas that belong to them. For example, 10 classes for adjectives account for 98% of lemmas, 10 classes for nouns account for 61.8% of lemmas, and 10 classes for verbs account for 59.6% of lemmas.

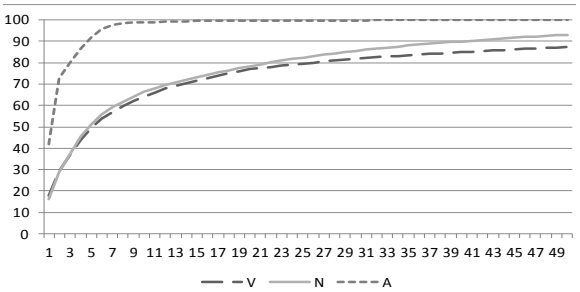


Figure 3: FST classes and the percentage covering the dictionary of lemmas

FST class prediction can be divided into two parts: one is extraction of implicit knowledge and the other is actual prediction of FST class for a new lemma. Extraction of implicit knowledge in the form of a dataset with word endings, grammatical categories and FST classes proceeds as follows:

1. Calculate frequencies for each POS and relative frequencies for each FST class within POS in the current dictionary of simple lemmas.
2. Create a dataset from DELAS lemma endings of length 3,4,5 and 6 characters with corresponding grammatical categories retrieved from DELAF (e.g. for nouns in that dataset: POS, lemma, FST, gender, animateness, pronunciation).
3. Create another dataset with frequencies for each combination of FST code and grammatical category and for each ending of length 3,4,5,6, as an estimate of the probability that the FST class is the appropriate one. The dataset includes: ending, POS, gender, animateness, pronunciation, FST and probability (chance rank 0-100) for FST (table 1, column Rel. freq.).

ending	POS	length	gender	anim	FST	Total	Frequency	Rel. freq.	Example
alica	N	5	f	-	N650	97	95	98	sij**alica**
alica	N	5	f	-	N1650	97	1	1	Sk**alica**
alica	N	5	f	+	N651	27	26	96	var**alica**
alica	N	5	f	+	N1651	27	1	4	L**alica**
ica	N	3	m	+	N1683	145	142	98	Milo**jica**
ica	N	3	m	+	N1741	145	3	2	Pr**ica**
ica	N	3	m/f	+	N683	40	33	83	tvrd**ica**
ica	N	3	m/f	+	N679	40	7	18	ub**ica**

Table 1: Excerption from dataset with ending.
(m: masculine gender; f: feminine gender; m/f: nouns change gender in their inflectional paradigm)

Analysis of the relation between word endings and inflectional FST classes shows that the prediction of inflectional class by the abovementioned statistical analysis of existing dictionaries is justified. Figure 4 illustrates this relation for word endings of length 3, 4, 5 and 6. For example, in the case of word endings of length 3, for 33% of words from the existing dictionary there is only one corresponding FST class, for approximately 20% of words there are two classes, and so on, whereas for word endings of length 6 there is a single class for as much as 90% of words.

In order to facilitate prediction of FST class, a set of rules based on inflectional class metadata is used. Distinction between inflectional classes based on grammatical categories can be done only to some extent, so implicit knowledge from the existing dictionary of simple words is used to improve prediction.

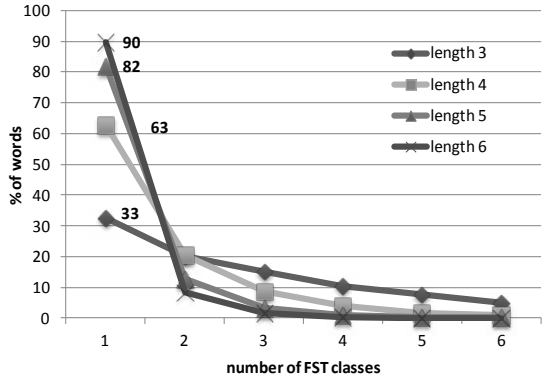


Figure 4: Relation between word endings and inflectional FST classes

The process of automatic prediction of inflectional FST class for a new entry follows a hybrid approach: one part is rule-driven with explicit codification of knowledge about FST classes and the other is statistical, based on existing dictionary of simple word lemmas with implicit knowledge about dependence between FST classes and dictionary entries.

After preparing the list of new entries in the form: lemma, POS, Grammatical_Categories (e.g. *grabuljar,N,Rud* 'rake') the following procedure is applied:

1. For each candidate lemma filter the dataset prepared from previous step as follows:
   1.1. if the lemma has specific marks for pronunciation, then retain only dataset members with the same mark and remove the rest;
   1.2. if the grammatical gender or animateness is assigned, retain only dataset members with the same grammatical category and remove the rest;
   1.3. if the first letter of the lemma is in upper case additional filtering can take place taking into account FST classes which have only inflected singular forms.
2. After filtering and ranking the dataset, prediction (FST assignment) for the lemma is repeated with threshold from 99 to 95 for relative frequency, for suffixes 6,5,4, and 3 respectively;
3. For thresholds under 95 and over 80 lemma prefix (if longer than 2 characters) is used: if the prefix is in the dictionary of prefixes and the remainder of the lemma is a word in DELAS, then the lemma is the inflectional class of the corresponding DELAS word is assigned to the lemma.
4. For thresholds 80 and less steps 1 and 2 only are repeated.

From a sample of domain texts and dictionaries we manually filtered 623 new terms from domains of mining, geology and e-learning and applied the described procedure for FST class prediction: to 582 (93%) of them the correct FST

class was assigned, 27 (4%) had a partly correct class assigned (for instance, inflection is correct but falsely allows plural forms), and to 14 (2%) of them an incorrect class was assigned.

4 Syntactic graphs for MWU recognition

4.1 Structure of terms in termbases

In order to analyze the structure of terms in different domains, primarily the number of components they consist of, we used samples from three terminological resources for Serbian. Two terminological resources, GeolISSTerm [2] and RudOnto[3] have been developed at University of Belgrade, Faculty of Mining and Geology. GeolISSTerm is a bilingual thesaurus of geological terms in Serbian and their English equivalents (Stankovic et al., 2011), divided in several subdomains: petrology, mineralogy, hydrogeology, geophysics, structural geology etc. RudOnto is covering the larger area of mining engineering and mine safety terminology (Stankovic et al., 2012). The third termbase used is the Dictionary of Library and Information Sciences (RNBS),[4] developed by the National Library of Serbia. It contains terminology in Serbian, English and German, related to theory and practice of librarianship and information sciences and a wide range of close or related fields.

Table 2. Frequencies of terms of different lengths in samples from 3 termbases

Dictionary	Term length (in number of words)				
	1	2	3	4	≥5
GeolISS Term	1436	2356	749	305	243
RNBS	3302	6180	2062	806	415
RudOnto	1004	1351	1350	1031	2341

Table 2 presents the distribution of terms consisting of 1, 2, 3, 4 and more components for the three termbases. These results are consistent with the results presented in (Justeson et al., 1995), at least for GeolISSTerm and RNBS, and show that terms with 5 or more components are much less frequent than the shorter ones. The results are somewhat different for RudOnto, as it contains very specific terms, such as causes of injuries, employee positions, types of injuries, or technical characteristics of machines, which are often longer MWUs than the less specific terminology of the two other termbases. Two examples from RudOnto can illustrate this: a term for employee position "Geologist for mineralogy, petrology, sedimentology and geochemical research" and a term for technical characteristics of machines "Length of the caterpillar transporting device measured from the vertical excavator rotation axis to the front edge of the caterpillar".

4.2 Extraction of MWUs from domain texts

The extraction of MWUs from a text is preceded by the retrieval of new simple word terms from it and their incorporation in the existing system of morphological e-dictionaries as MWU extraction relies heavily on existing lexical resources.

In the Serbian e-dictionary of MWUs, all entries are distributed in classes according to their syntactic structure, or more precisely, according to the information needed for their inflection. The names of classes correspond to the names of special FSTs that are used for MWU inflection. For instance, the class AXN pertains to MWUs with the syntactic structure: an adjective (A) followed by a noun (N), where the two components agree in gender, number, case and animateness. In class names X stands for a component that does not inflect when a MWU inflects or for a component separator. In the case of AXN, X stands for the separator, usually a space. Sometimes, MWUs with different syntactic structure belong to the same class. For instance, the class N4X implies that MWUs belonging to it consist of a noun followed by two other components (separated by two separators) that do not inflect. The syntactic structure of these components can be a noun followed by two adjectives/nouns in the genitive case (e.g. *eksploatacija mineralnih sirovina* 'exploitation of mineral resources') but also a noun followed by a prepositional phrase (e.g. *bager na šinama* 'excavator on rails').

There are 29 such classes for Serbian nominal MWUs.[5] However, 10 of them are used for the inflection of more than 98% of all nominal MWUs. Four of these classes are used for the inflection of two component MWUs, four for the inflection of 3-component MWUs and two for the inflection of 4-component MWUs. Given that

[2] http://geoliss.mprrpp.gov.rs/term
[3] http://rudonto.rgf.bg.ac.rs/
[4] http://rbi.nb.rs/en/home.html

[5] The number of FSTs (80) is greater than the number of classes because they deal with other details of inflection: does the MWU inflect in number, are some components optional, etc.

they cover the large majority of MWUs, we have developed syntactic FSTs for the extraction of MWUs belonging to these 10 classes. They are, listed in the descending order of their frequency:

1. **AXN** – an adjective followed by a noun; the adjective and the noun have to agree in all four grammatical categories; e.g. *zemni gas* 'natural gas'.
2. **2XN** – a noun preceded by a word that does not inflect in the MWU. Usually it is a word used only in one or a few MWUs, a prefix or an adverb derived from an adjective, while the separator is usually a hyphen; e.g. *anker-mreža* 'anchor network'.
3. **N2X** – a noun followed by a word that does not inflect in the MWU. Usually this word is a noun in the genitive or in the instrumental case; e.g. *patrona eksploziva* 'explosive cartridge' and *upravljanje krovinom* 'roof control'.
4. **N4X** – a noun followed by two words that do not inflect in the MWU. Two syntactic structures are possible:
   a. **NNgi** - A noun followed by two adjectives/nouns in the genitive case or in the instrumental case; e.g. *otkopavanje širokim čelom* 'broad forehead excavation'.
   b. **NprepNp** - A noun followed by a prepositional phrase; e.g. *lanac sa grabuljama* 'chain with a rake'.
5. **AXN2X** – a noun preceded by an adjective that agrees with it in gender, number, case and animateness and followed by a word that does not inflect in the MWU, usually a noun in the genitive or instrumental case; e.g. *geološko kartiranje terena* 'geological field mapping'.
6. **NXN** – a noun followed by a noun that agrees with it in number and case, where the separator can be a hyphen; e.g. *bager kašikar* 'shovel excavator'.
7. **AXAXN** – a noun preceded by two adjectives that agree with it in gender, number, case and animateness; e.g. *površinski istražni radovi* 'surface exploration works'.
8. **N6X** - a noun followed by three words that do not inflect in the MWU. Three syntactic structures are possible:
   a. **NNgiPrepNp** - a noun followed by a noun in the genitive case and a prepositional phrase (as in case 4b); e.g. *priprema ležišta za otkopavanje* 'deposit preparation for mining'.
   b. **NNgiNgiNgi** - a noun followed by three nouns/adjectives in the genitive case; e.g. *istraživanje ležišta mineralnih sirovina* 'exploration of mineral deposits'.
   c. **NprepNpNgi** - a noun followed by a prepositional phrase; e.g. *bakar sa primesama zlata* 'copper with a sprinkling of gold'.
9. **AXN4X –** a noun preceded by an adjective that agrees with it in gender, number, case and animateness and followed by two words that do not inflect in the MWU. Two syntactic structures are possible:
   a. **ANPrepNp** - A noun preceded by an adjective and followed by a prepositional phrase (as in case 4b); e.g. *gravitacijska koncentracija u vodi* 'gravity concentration in water'.
   b. **ANNgiNgi** - a noun preceded by an adjective and followed by two adjectives/nouns in the genitive case or in the instrumental case (a 4a case); e.g. *površinska eksploatacija mineralnih sirovina* 'surface exploitation of mineral resources'.
10. **2XAXN** - an adjective followed by a noun that agrees in all four grammatical categories and preceded by a word that does not inflect in the MWU; e.g. *magmatsko-eruptivni masiv* 'magmatic-igneous massif'.

FST for extraction of MWUs of type AXN with two paths from one of the subgraphs that illustrate the agreement between adjectives and nouns is depicted in Figure 5. Dictionary variable used for FST output in the form $a.LEMMA$ retrieves a lemma of recognized word form a thus performing the simple word lemmatization.

Due to high homography of word forms it may happen that the same sequence of words is recognized by two or more graphs; naturally, only one recognition may be correct. For instance if the MWU *bager kašikar* (case 6, NXN) is detected in the analyzed text in the genitive case *bagera kašikara* it may be erroneously interpreted as a MWU of a form NNg (case 3) in the genitive case. Consequently, all NNg constructions in an analyzed text that appear in the genitive case (which happens very frequently) will be interpreted also as a NXN case. For that reason, in the case of ambiguous recognition we always give precedence to the more probable case. For instance, for 2-component MWUs the precedence is: AXN, 2XN, N2X, NXN.

As a rule, we are looking for the longest match for a MWU, that is, if a text matches an

AXAXN pattern, than we will ignore the match AXN that is subsumed. However, in certain cases we take into consideration the shorter matches as well. For instance, a sequence recognized as NNgNgNg, may well not be a multi-unit term, but rather consist of two multi-unit terms of the form NNg or contain as its part a AXAXN term; e.g. *sprečavanje zagađenja životne sredine* 'prevention of environmental pollution' may not be considered a term, while *zagađenje životne sredine* '*environmental pollution*' is. For that reason, the order of term candidate extraction is:

1. AXAXN, 2XAXN, AXN2X, AXN4X, AXN
2. N6X
3. N4X
4. 2XN, N2X, NXN

At the end of each round duplicates are eliminated according to the priorityand the union of all results is performed.

The output of processing by transducers is the initial version of the normalized MWU that consists of simple word lemmatization — inflected parts of a MWU are replaced by their lemmas, as they are recorded in e-dictionaries. The list of produced normalized MWUs is then additionally processed by a new set of transducers in order to obtain correct MWU lemmas. The following adjustments have to be performed:

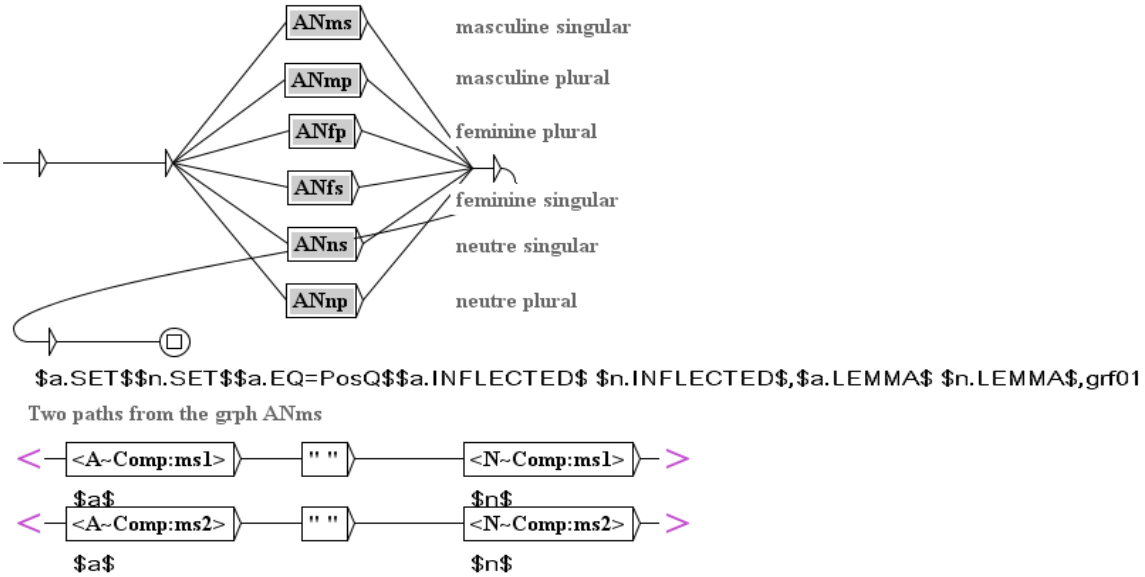


Figure 5. An FST for extraction of MWUs

1. For MWUs with syntactic structure AXN, AXAXN, AXN2X, AXN4X, and 2XAXN the form of the adjectives has to be corrected so that the right gender is selected to correspond to the gender of the noun (simple word lemmas are always in the masculine gender). For example, when simple word lemmatization offers a lemma $minski_m$ $bušotina_f$ 'blasting boreholes' it has to be corrected to $minska_f$ $bušotina_f$.
2. For all MWUs, the right number of the MWU has to be selected: if it appeared in a text only in singular form or only in plural form, then the lemma will be in the respective form (e.g. only singular form *jamski vazduh* 'air in the underground mine', only plural form *atmosferske padavine* 'atmospheric precipitation'); if it appeared in both plural and singular forms, then both forms of lemmas will be offered.

Production of correct MWU lemmas is a prerequisite for the successful evaluation. Moreover, entries for morphological e-dictionary of MWUs can be produced only from correct MWU lemmas. Finally, as a byproduct of the whole process MWU inflectional classes for newly retrieved MWUs are obtained – they are derived directly from local grammars used for their extraction.

4.3 Evaluation of performance of MWU extraction

In order to evaluate our approach, we applied it to a collection of 74 papers in Serbian from the journal Infotheca.[6] The size of the corpus is

[6] Infotheca - Journal for Digital Humanities (http://infoteka.bg.ac.rs/index.php/en/infoteca)

272,557 simple word forms. Our procedure extracted from it 65,279 MWUs, 86.9% of them occurring only once, 7.9% occurring twice, 3.8% occurring 3 to 5 times and 1.9% with more than 5 occurrences.

The graph 3 (N2X) extracted 31% of all MWUs with frequency greater than 1. It is followed by graph 6 (NXN) with 26% MWUs, graph 4 (N4X) with 22%, graph 1 (AXN) with 16%, and the remaining six graphs with 6%. As to MWUs with frequency greater than 5, graph 1 (AXN) covers 31%, graph 3 (N2X) 25%, graph 6 (NXN) 22%, graph 4 (N4X) 17%, and the remaining six graphs 5%.

Extracted MWUs were manually evaluated on a subset of 690 entries. The evaluators checked 1) whether proposed lemmas were grammatically correct and 2) whether MWU terms belong to domain terminology, in this case library and information science, or to the general lexica.

For candidate ranking three measures were used: frequency, C-Value (Franzi et al., 2000) and log-likelihood (Dunning, 1993; Gelbukh et al., 2010).

For grammatical correctness best precision at rank *n* (P@n) measure is very high (figure 6) and independent of the ranking (the trend is flat).

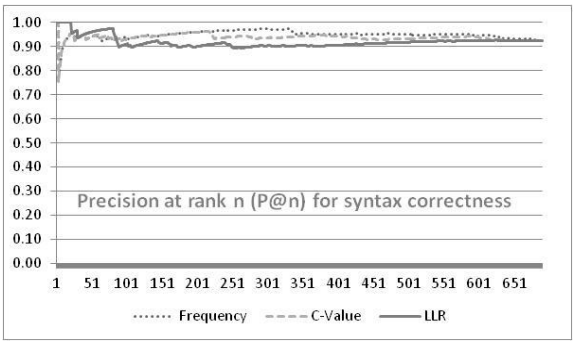


Figure 6: Precision at rank *n* for all evaluated term candidates for grammatical correctness.

In order to calculate the log-likelihood measure we used an excerpt from the general Corpus of Contemporary Serbian[7] that consists of 22 million simple word forms.

Figure 7 presents the precision at rank *n* for 690 evaluated term candidates for domain affiliation, where log-likelihood gave best results for precision at rank *n* (P@n) measured on a sorted list of candidates.

[7] The Corpus of Contemporary Serbian (http://www.korpus.matf.bg.ac.rs/)

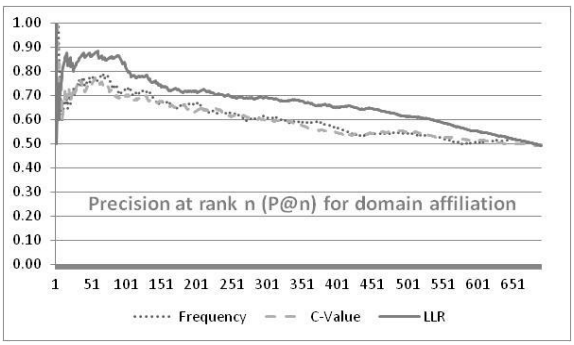


Figure 7: Precision at rank *n* for all evaluated term candidates for domain affiliation.

5 Discussion and Conclusion

The research outlined in this paper tackles the extraction of domain terminology and its integration into terminological dictionaries using lexical resources and local grammars. Results obtained by following this approach justify its basic assumption that the task of term extraction, both in the case of simple words and multi-word units, can be successfully accomplished combining existing e-dictionaries and FSTs. Moreover, lexical resources and local grammars alleviate the task of integrating the newly discovered terms into terminological dictionaries by simplifying the task of defining the proper inflectional class for new terms, a task extremely complex in case of morphologically rich languages such as Serbian. By implementing the procedure proposed within this paper we have considerably sped up the development of terminological dictionaries for Serbian.

Further research will address the integration of inflectional class prediction in existing software tools used for handling dictionaries developed at University of Belgrade and creation of a web tool that would support the entire procedure described in this paper. Production of dictionary entries in DELA format for verbs, akin to the one described for nouns, is also under consideration. A detailed evaluation will follow with the aim of further refinement of the presented procedure in order to reduce to the least possible extent the necessity for human intervention within the process of terminology acquisition and description. Our future work will be oriented towards usage of Web sites for evaluation of new term candidates (Robitaille et al., 2006).

Acknowledgement. This research was supported by the Serbian Ministry of Education and Science under the grant #47003 and Parseme COST action IC1207.

References

Ammar, C., Haddar, K., & Romary, L. (2015). Automatic Construction of a TMF Terminological Database Using a Transducer Cascade. *Proc. of Recent Advances in Natural Language Processing*. (pp. 17-23).

Baldwin, T., & Kim, S. N. (2010). Multiword expressions *Handbook of Natural Language Processing, second edition*. (267-292): CRC Press.

Cerbah, F., & Daille, B. (2007). A Service Oriented Architecture for Adaptable Terminology Acquisition. In Z. Kedad, N. Lammari, E. Métais, F. Meziane & Y. Rezgui (Eds.), *Natural Language Processing and Information Systems* (Vol. 4592: 420-426): Springer Berlin Heidelberg.

Courtois, B., Silberztein, M. (1990). Dictionnaires électroniques du français. Larousse, Paris.

Daille, B. (2000). Morphological rule induction for terminology acquistion. *Proc. of the 18th conference on Computational linguistics*- (Volume 1: pp. 215-221).

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist., 19*(1), 61-74.

Enguehard, C., & Pantera, L. (1995). Automatic natural acquisition of a terminology. *Journal of quantitative linguistics, 2*(1): 27-32.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries, 3*(2): 115-130.

Gelbukh, A., Sidorov, G., Lavin-Villa, E., & Chanona-Hernandez, L. (2010). Automatic Term Extraction Using Log-Likelihood Based Comparison with General Reference Corpus. In C. Hopfe, Y. Rezgui, E. Métais, A. Preece & H. Li (Eds.), *Natural Language Processing and Information Systems* (Vol. 6177, pp. 248-255): Springer Berlin Heidelberg.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1* (01): 9-27. doi:10.1017/S1351324900000048

Krstev, C. (2008). *Processing of Serbian. Automata, Texts and Electronic Dictionaries*: Faculty of Philology of the University of Belgrade.

Nakagawa, H., & Mori, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology, 9*(2), 201-219.

Przepiórkowski, A., Degórski, Ł., & Wójtowicz, B. (2007). On the evaluation of Polish definition extraction grammars. *Proc. of the 3rd Language & Technology Conference*.

Quochi, V., Frontini, F., & Rubino, F. (2012). A MWE Acquisition and Lexicon Builder Web Service. *Proc. of COLING 2012* (pp. 2291-2306).

Ramisch, C., De Araujo, V., & Villavicencio, A. (2012). A broad evaluation of techniques for automatic acquisition of multiword expressions. *Proc. of ACL 2012 Student Research Workshop (1-6)*.

Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., & Utsuro, T. (2006). Compiling French-Japanese Terminologies from the Web. *Paper presented at the 11th Conference of the European Chapter of the Association for Computational Linguistics - EACL*.

Rodrıguez, F. M. B., Noya, E. D., Otero, P. G., Martınez, M. L., Mato, E. M. M., Rojo, G., Docıo, S. S. (2007). A Corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a Minority Language. *Proc. of 3rd Language & Technology Conference*.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP *Computational Linguistics and Intelligent Text Processing* (1-15): Springer.

Savary, A. (2009). Multiflex: A Multilingual Finite-State Tool for Multi-Word Units. In S. Maneth (Ed.), *Implementation and Application of Automata* (Vol. 5642, pp. 237-240): Springer Berlin Heidelberg.

Savary, A., Zaborowski, B., Krawczyk-Wieczorek, A. & Makowiecki, F (2012). SEJFEK—a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. *Proc. of Cognitive Aspects of the Lexicon (COGALEX-III)*. (pp. 195-214).

Zhang, Y., Kordoni, V., Villavicencio, A., & Idiart, M. (2006). Automated multiword expression prediction for grammar engineering. *Proc. of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. (pp. 36-44).

Helping term sense disambiguation with active learning

Pierre André Ménard
Centre de recherche informatique
de Montréal, Canada
pamenard@gmail.com

Caroline Barrière
Centre de recherche informatique
de Montréal, Canada
caroline.barriere@crim.ca

Jean Quirion
École de traduction
Université d'Ottawa, Canada
jquirion@uottawa.ca

Abstract

Our research highlights the problem of term polysemy within terminometrics studies. Terminometrics is the measure of term usage in specialized communication. Polysemy, especially within single-word terms as we will show, prevents using term corpus frequencies as appropriate statistics for terminometrics. Automatic *term sense disambiguation*, as a possible solution, requires human annotation to feed a supervised learning algorithm. Within our experiments, we show that although being polysemous, terms have a strong in-domain sense bias, making random sampling of annotation data less than optimal. We suggest the use of active learning and implement it within an annotation platform as a way of reducing annotation time.

1 Introduction

In our research, we investigate the measure of term usage in specialized communication, called Terminometrics (Section 2). The studied *terms* are provided by a *term bank*, and we are interested in the fact that many of these terms are polysemous, creating difficulties for our terminometrics study. We therefore investigate the presence of polysemy in term banks (Section 3). Polysemy found in term banks is problematic since it leads to term occurrences in corpora being possibly polysemous, preventing simple corpus frequencies to provide proper statistics. We confirm such polysemous occurrences within our specific terminometrics experiment in the nanotechnology domain (Section 4), as we analyse term sense human annotation results for a set of nanotechnology terms.

Results also show that terms, although polysemous, have a very strong bias toward their in-domain sense. In such biased case, a random sampling of annotation data is far from optimal, wasting much human effort. We therefore introduce active learning (Section 5) and implement it within an annotation platform (Section 6), to obtain a sense-annotated dataset in less time.

2 Terminometrics

Terminometrics is the measure of term usage in different types of communications (Quirion, 2006). Its purpose is to determine, for a particular concept, the relative corpus frequencies of its competing terms.

The protocol of terminometrics, as defined in Quirion (2003), consists in first deciding on a domain of interest and selecting its set of concepts (most often all) from a term bank. Then, for each particular concept, the individual number of occurrences of all its competing terms is counted within different corpora from the same domain gathered by terminologists to represent different communicative settings. Acknowledging the possible polysemy of competing terms, the protocol includes a human expert, to actually disambiguate a randomly selected subset of occurrences, and obtain better estimates of real frequencies.

A good example of this would be the concept of a *atomic cluster* within the *nanotechnology* domain. According to the term bank used, such notion can be expressed by the following 6 terms *atomic cluster*, *atom cluster*, *atomic aggregate*, *atom aggregate*, *cluster* and *aggregate*. In terminometrics, comparative studies of use of terms in specialized communications, government liter-

ature, specialized media, and general media are of interest, as they might reveal how some terms are used by the general public, while others are used by more official government documents.

Studying the occurrence in text of different synonyms of concepts would not be problematic if each one was monosemous. But unfortunately, that is not the case. For example, referring to Table 1, the term *cluster* is a competing term for multiple concepts, and simply counting its occurrences in text, without disambiguation, would not be indicative of its usage for any of them.

Obviously, human annotation is costly, and the possibility of performing automatic term sense disambiguation is quite appealing. In terminometrics, concepts are evaluated one at a time, reducing the disambiguation task to a binary decision. The annotation is not a selection among N senses, but rather a yes/no decision on whether the current instance represents the current concept or not. Furthermore, term disambiguation within terminometry cannot be dealt with similarly to more typical word-sense disambiguation or even term-sense disambiguation relying on knowledge contained in an external resource (Barrière, 2010) since the annotator, or the algorithm, is likely to only have access to the context of occurrences to perform term disambiguation.

3 Polysemy of specialized terms

Terms for the terminometrics studies are provided by term banks. Such repositories of terms are not often investigated for the study of polysemy. In Natural Language Processing, a typical task of word sense disambiguation requires a lexicographic resource, such as WordNet (Miller, 1995), to provide a repository of possible word senses in order to disambiguate words in texts (Pantel and Lin, 2002). No doubt that words are polysemous, even in specific domains (Chroma, 2011; Vogel, 2007), but less studies show and discuss the polysemy of terms.

Terms are single-word or multi-word expressions denoting particular concepts within particular domains. A term bank is organized by domains (e.g. biology, automotive, etc) and contains records corresponding to concepts. Each record contains at least one term, and often competing terms (synonyms) denoting that concept, possibly in more than one language. Examples of records for the term *cluster*, as found in the Grand Dictionnaire Terminologique (GDT)[1] are shown in Table 1.

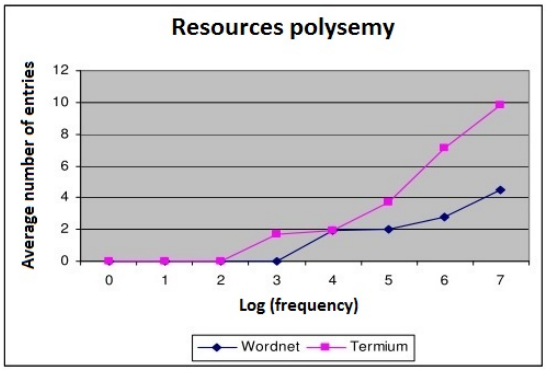


Figure 1: Degree of polysemy in Wordnet and Termium per term length

There might be a misconception that specialized language is less ambiguous, and would then not provide a proper challenge for word-sense disambiguation. A study by Barrière (2007), shows the contrary, as Wordnet and Termium[2] (the actual resource used in this experiment) were compared along different criteria. One criteria of comparison was coverage, and another one, more of our interest in this research, is the degree of polysemy in relation to word specificity. Word specificity was approximated by "hit counts", as found in a very large corpora (Waterloo Terabyte Corpus, used by Terra and Clarke (2003)), with words occurring from 1 to millions of times. Figure 1 shows their results. We see how for common words (hit counts in the $log_{10}(freq) > 3$), the degree of polysemy in the term bank is even larger than in WordNet.

In our study, we wished to further characterize this degree of polysemy in terminological resources. We used a small set of 164 terms from the current experiment (presented in Section 4.1), and looked at the number of senses in two term banks: Termium and GDT. Figure 2 shows that specialized terms, especially short ones (1 to 3 words) can have many senses (records) and span many domains. This trend generally diminishes as the term length increases.

[1]The GDT can only be accessed via a web interface at *http://www.granddictionnaire.com* .

[2]Termium term bank can be accessed online at *http://www.btb.termiumplus.gc.ca* or downloaded at *http://open.canada.ca/data/en/dataset/94fc74d6-9b9a-4c2e-9c6c-45a5092453aa*

Domain	Terms
nanotechnology	atomic aggregate, **cluster**, aggregate, atom aggregate, atom cluster, atomic cluster
	molecular aggregate, **cluster**, aggregate, molecule aggregate, molecule cluster
	nanoaggregate, **cluster**, aggregate, nanocluster, nanometer-size cluster, nanoscale aggregate, nanoscale cluster
seafood	crab section, section, crab cluster, **cluster**
software	**cluster**, document cluster
mining	vein system, vein set, cluster of veins, mining cluster, **cluster**
internet	service cluster, cluster of service, **cluster**
nanotechnology	scanning tunneling electron microscope, **microscope**, scanning tunneling microscope, STM
	atomic force microscope, **microscope**, AFM, SFM, scanning force microscope
	magnetic force microscope, **microscope**, MFM, SMM, scanning magnetic microscope
	scanning probe microscope, **microscope**, SPM, scanned-probe microscope

Table 1: Different records for "'cluster"' and "'microscope"'.

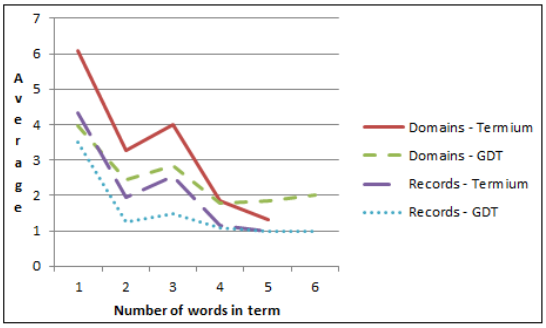


Figure 2: Degree of polysemy in Termium and GDT per term length

4 Experiment - Terminometrics in nanotechnology domain

Our current terminometrics study focuses on term usage in the nanotechnology domain within Canadian French. This domain, within the GDT term bank, contains 1,035 records (concepts)[3], each with its competing terms. This set of terms is what we call our *nanotechnology term base* covering "'the science of working with atoms and molecules to build devices that are extremely small"' (Merriam-Webster dictionary).

To study the competing terms for the nanotechnology concepts, a corpus was built using documents from corporative, educational, news medias and government websites. These documents were retrieved first by selecting most of the organizations originating from the province of Québec, Canada, and whose core activities dealt with nanotechnology. This list was then vetted by an expert. Next, the websites of these organizations were downloaded. After such process, the corpus might still be noisy, but it does contain a majority of nanotechnology-related documents.

All terms in the nanotechnology term base are searched for in the corpus. For each of their occurrences, a window spanning 90 characters each side of the term is extracted. This text span becomes a contextualized instance to be annotated. Table 2 shows examples of these instances.

4.1 Human annotation process

For our current annotation experiment, a total of 164 terms taken from 29 records (among the 1,035 mentioned earlier) were selected along with the complete set of instances found in the nanotechnology corpus. Each term occurred between 75 to 2100 times in the corpus for a total of 17,227 instances for the whole term sample. This dataset was divided into two parts distributed between 2 PhD students in terminology. As shown in Table 2, annotators were presented text sample with a targeted term and were asked to indicate "yes" if the term was used in the correct nanotechnology sense and "no" otherwise. Prior to the annotation effort, the dataset was sorted by terms, as this was considered easier to annotate compared to an annotation by document order, which would ask the annotator to constantly switch between term definitions. They took a total of 82 hours (41 hours each) to annotate all the instances of the selected dataset. Each text sample was composed of the 90 characters prior to a term occurrence, the term occurrence as is, and another 90 characters following the term occurrence. The 90 characters window was adjusted to avoid word truncation.

The annotators were also asked to indicate the difficulty level of the provided answer: standard,

[3]As the GDT expands everyday, this number might not represent its current status.

Annotation	Instance
Yes	... une technologie d'intégration par laquelle plusieurs nanostructures sont intégrées sur un même **substrat**. L'interface entre les dispositifs et d'autres systèmes (oxyde, verre) sera aussi étudiée. *(... an integration technology for which many nanostructures are integrated on a substrate. The interface between the components in other systems (oxyde, glass) will also be studied.)*
Yes	... dollars à Bromont dans une petite usine qui allait employer 200 personnes pour la production de **substrats**, que le dictionnaire définit comme un matériau sur lequel sont réalisés les éléments d'un ... *(...dollars at Bromont in a small factory which was going to employ 200 people for the production of substrates, which dictionary define as a material on which are realized elements of...)*
No	... et valoriser les boues de station d'épuration. L'investigation des possibilités d'acquérir ces **substrats** requiert l'inventaire des industries de la région, les quantités et les caractéristiques des ... (*... and valorize the epuration station's muds. Investigating the possibility of acquiring these substrates requires to inventoriate the region's industries, the quantity and features of ...)*
Yes	... MNT Définition : Fabrication mécanique et contrôlée de structures moléculaires, par une **approche ascendante** qui consiste à les assembler, étape par étape, molécule par molécule, en se servant d'appareil ... (*... MNT Definition : Mechanical and controled fabrication of molecular structures by a bottom-up approach which consist of assembling, step (by step, molecule by molecule, by using tool ...*
No	... Quand il est possible de le faire, l'analyse de la demande d'énergie est fondée sur une **approche ascendante** agrégeant les demandes par usage, par secteur d'activités économiques, par région et par ... (*When it is possible to do it, the energy request analysis is founded on a bottom-up approach aggregating the requests by use, by economic activity sector, by regions and by ...)*
No	... que beaucoup de problèmes rencontrés en pratique ne sont pas adressés par ces processus. L'**approche ascendante** de l'amélioration du processus consiste donc, selon ces mêmes auteurs, à implanter une équipe ... (*... that many issues encountered in practice are not adressed by these processes. The bottom-up approach of process improvement consist of, for these same authors, implanting a team ...)*

Table 2: Instances for the terms *substrat* (substrate) and *approche ascendante* (bottom-up approach)

hard, hardest. Results showed that 626 instances (3.6%) needed a little more analysis while 222 instances (1.3%) were much harder to annotate with only the presented context. All the other instances were judged of standard difficulty meaning that the textual contexts of the term occurrences were sufficient for the disambiguation task. In anticipation of an automatic disambiguation algorithm which would only have access to the immediate context of the term, this confirmed that for most cases, it should be possible to disambiguate with a $\pm$90 characters window[4].

4.2 Observations and results on polysemy

Analysis of the annotated instances reveals that 84.31% (14,524) of them occur in the correct nanotechnology sense of the term, and the remaining 15.69% (2,703 instances) are used with other meanings. To measure the overall polysemy in our dataset, we use the notion of entropy. Entropy is defined as a summation of all possible event probabilities multiplied by the log of their probability. In our current experiment, there are only two possible events, first the occurrence of a term in a correct sense, let us call that x, and second, the occurrence of a term in a different sense. If $P(x)$ is the probability of the correct sense, then $1 - P(x)$ is the probability of another sense. Then, we have the entropy, shown in Equation 1, as a sum over two possible events.

$$E(x) = P_x log_2 P_x + (1 - P_x) log_2 (1 - P_x) \quad (1)$$

The resulting function is at its maximum, a value of 1, with a probability of 50% and is equal to 0 with probabilities of either 0% or 100%. In our case, x is the rate of occurrence of an anticipated term sense in a corpus. A term with an entropy of 0 would mean it is not ambiguous, either all or none of the term's instances use the correct sense, and a term with an entropy of 1 would mean 50% of its instances are used in the correct sense, the remaining 50% of the instances using other meanings.

For example, the term *STM* (acronym of *scanning tunnelling microscope*) counts as a single-word term occurring a total of 341 times. Among those, 104 instances (104/341=0.30499) have the nanotechnology sense, which gives an entropy of 0.8873 as shown in Equation 2. This is a relatively high entropy level as it nears the 50% maximum. If the case would have been less ambiguous, for example 5 out of 341 instances, the entropy would have been 0.1103.

[4]This claim disregards the fact that humans certainly have much apriori knowledge which they use during the disambiguation task. Nevertheless, trigger of this apriori knowledge would still come from the limited context window.

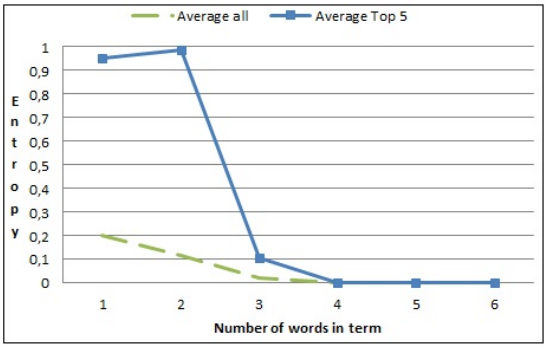


Figure 3: Average entropy per length on gold corpus

$$E(STM) = 0.30499 \times log_2(0.30499) + (1 - 0.30499) \times log_2(1 - 0.30499) = 0.8873 \quad (2)$$

The bottom dashed line (Figure 3) shows the average entropy over all terms having a particular word count. The top full line shows the average entropy for the 5 terms with the highest enthropy (and thus the highest degree of ambiguity) of each length, emphasizing how a few terms account for much of the corpus polysemous instances. Examples of these very polysemous terms are *tunnelling*, *substrat*, or *top-down*.

These corpus results, showing an overall tendency for entropy to decrease with term length, are in line with our previous results presented in Figure 2 relating term length to the polysemy level within term banks. Nevertheless, these corpus results also show that the in-domain sense is much more likely than all other senses. This leads us to think that we should take advantage of the particularity of our task in selecting the annotation dataset, as we further describe in the next section.

5 Active learning for term sense annotation

The strong in-domain sense bias results shown in the previous section, indicate that random sampling, suggested by the terminometrics methodology, could lead to collecting a biased sample and provide a distorted analysis. Traditional machine learning algorithms trained on these unbalanced samples would suffer the same bias, as less information would be available to classify the minority class. This type of algorithm would likely produce

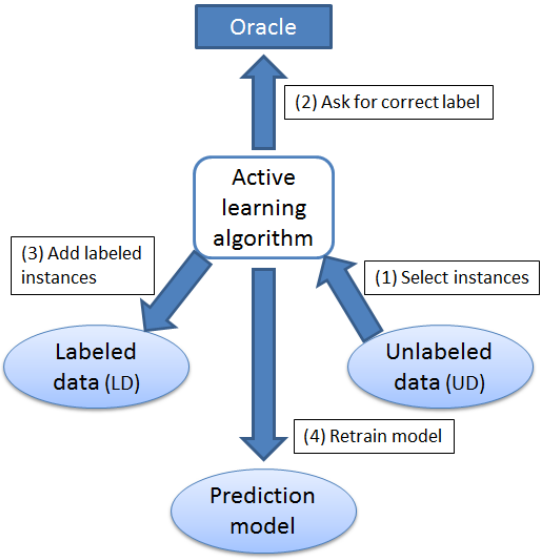


Figure 4: High level view of the active learning process.

a prediction model which would only target the majority class, overlooking instances potentially useful for terminometrics experts.

To sidestep this risk, we lean toward a learning approach called *active learning* which defines an iterative annotation process in order to reduce the risk of producing a biased prediction model. As shown in Figure 4, this four-step process implies the interaction with an oracle, typically a human annotator who needs to be familiar with the domain's terminology and concepts being studied.

The active learning process starts with a set of unlabelled data (UD) containing, in the current context, individual occurrences of a term in a corpus, described by a group of features (e.g. a bag-of-word made of its co-occurring words in context). At this point, the labeled dataset (LD) is empty and there is no prediction model available. The active learning algorithm starts by selecting a group of instances, called the seed S, from UD. For each instance of S, the oracle is queried to specify a label, and the labeled example is then stored in LD. The oracle annotates the instance using one value of a predefined class label set, in this case $\{yes, no\}$, *yes* meaning the instance is used in the targeted sense, *no* if another other sense is used. When all instances in S are labeled, the active learning algorithm uses them to create a prediction model. It is important to note that there is no ideal size for the seed, but it should be suf-

ficient to enable the algorithm to train a relevant prediction model.

Once a prediction model is available, the process takes place in the same order, but with a variant. Instead of a seed, the algorithm superficially applies the prediction model to instances in UD (without labeling them or changing them to the labeled set) and pick an instance for which the model does not provide a sufficient level of confidence for its classification. It then submits this instance to the oracle who applies a label. Then, the newly labeled example is added to LD. The prediction model is then retrained and the process continues until the algorithm reaches an overall level of confidence for all instances in UD.

When this stopping criteria is reached, the active learning process is complete and the prediction model can be used to annotate the remaining instances in UD, if needed, or another similar dataset. Again, the level of confidence used as the stopping criteria must be empirically defined, as there is no ideal value. Of course, a higher confidence level might increase the annotation effort needed to produce the final prediction model, while a lower value might produce a less effective prediction model using fewer instances. Fine-tuning the confidence level helps to reduce the risk of training a biased prediction model on a predominant class in a dataset.

In our current implementation of active learning, we select a seed of 20 instances with random sampling which is then processed with RandomForest (Tin Kam Ho, 1995) as the prediction model. The oracle is then asked to annotate other blocks of 20 instances until the algorithm reaches its parametered confidence level. If this level is not reached after a total of 200 instances (including the seed), a final prediction model is trained and applied on UD in order to limit the effort to annotate each expression. The features for the classification process are extracted from the 90 characters window, which was judged as sufficient during the experiments (Section 4.1).

At this stage in our research, the current implementation provides a baseline on which we can later improve using different alternative models presented in the literature. Certainly, other research in word sense disambiguation has explored the empirical behaviour of active learning (e.g. (Chen et al., 2006)). Specific issues associated with active learning range from feature selection for particular disambiguation tasks (Palmer and Chen, 2005), model adaptation when changing domain between the training and application of the model (Chan and Ng, 2007), class imbalance problem (Zhu and Hovy, 2007) or deciding when the prediction algorithm stops asking for additional annotation (Zhu et al., 2008).

6 Terminometrics active-learning platform

We developed an annotation platform, shown in Figure 5, to facilitate terminometrics studies with an active learning component for term disambiguation. The platform implements the interactive active learning process described above to control and optimize the active learning between the prediction module and the human annotator. The platform will also enable future experiments within the field of terminometrics in which both the active learning algorithms and the human interaction can be further explored.

The user of this platform (typically the oracle in the active learning process) can create a corpus of documents, use this corpus to create an annotation project by defining a set of concepts, related terms and variations (plural, gender) and participate in the active learning process. At the end of the active learning process, the platform annotates the remaining instances in UD (see Figure 4) in order to estimate the distribution of occurrences of competing terms of a concept. This is used for the terminometrics analysis.

Aside from the $\{yes, no\}$ classification, the interface offers two other choices; *undecided* and *reject*. The first choice allows the user to skip an instance and go to the next, while being able to later return to provide an answer. This could happen when the user wishes to see a larger context to perform the disambiguation. In fact, to help this process, the platform also provides an option to view an instance within its original document. The second choice, *reject*, removes the instance entirely from the unlabeled and labeled datasets. This is used typically when the user considers that the instance should not be used for the terminometrics final analysis.

In order to further reduce the annotation effort needed to perform a terminometrics study, other features, unrelated to active learning, were added

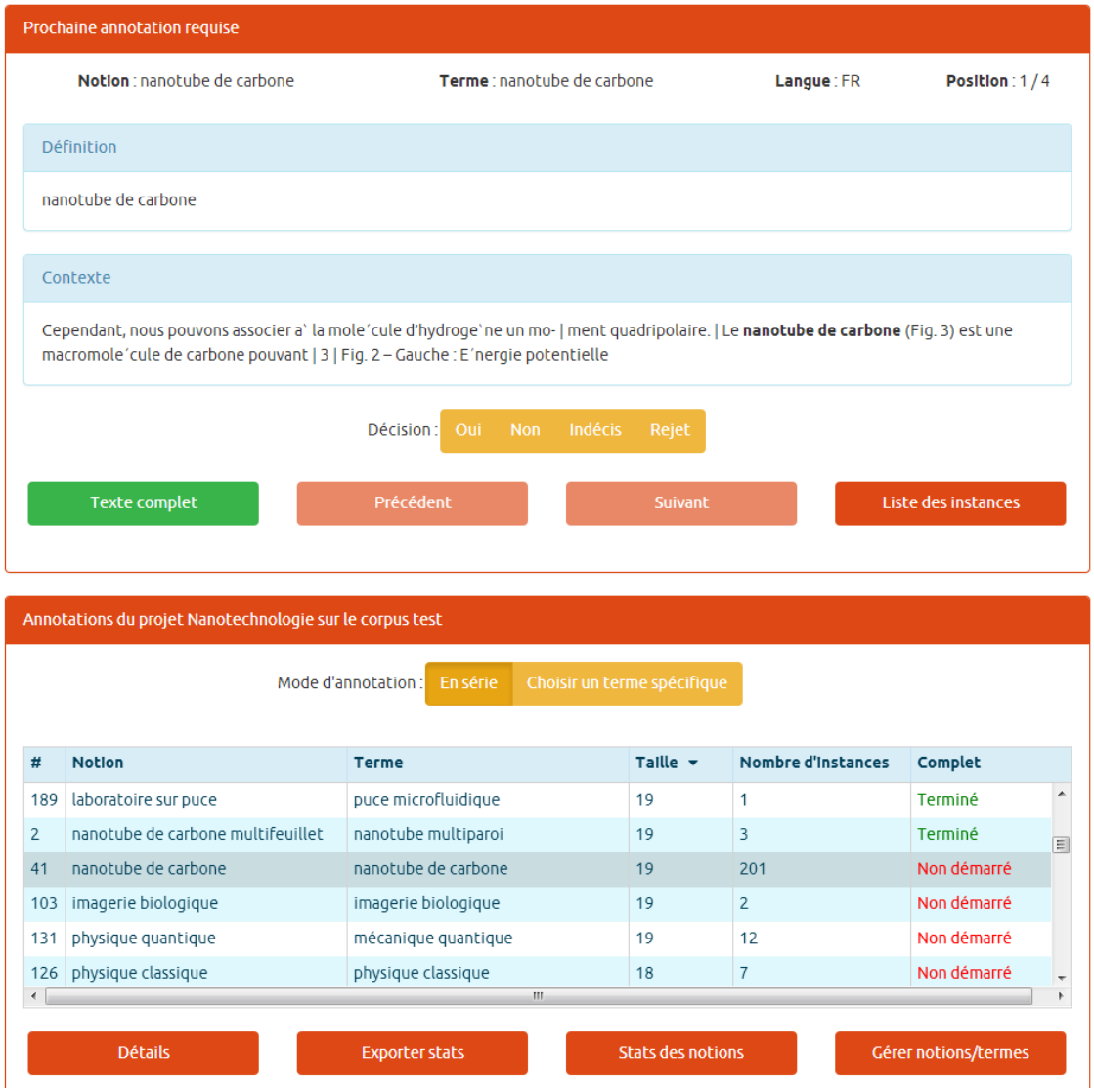


Figure 5: Annotation interface for terminometrics.

to the platform. The first is a language-based document filter which can be applied during the corpus creation to try to remove documents which are not suited for the targeted analysis. Each document is analysed with a language detection algorithm to extract a confidence level associated with its deduced language. It then enables the user to keep only the documents which are above a specific threshold and exclude the remaining from the corpus to be annotated. Of course, documents with no text, such as files containing only images, are also removed.

Another effort reduction feature is the duplicate context detection which takes place at the creation of an annotation project. The source issue is that a sentence or a whole paragraph (or sometimes complete documents) can be found in several locations within a corpus created from web sites. While each occurrence of a term (or its variations) is stored and kept for an accurate assessment of its rate of occurrence in a corpus, only unique contexts (the term occurrence and a ± 90 characters window) are used for the active learning process. For example, if the first context of "'substrat"' shown in Table 2 was found with the same prior and post context in five documents in a corpus, the oracle would be asked at most once to annotate this instance (if it is selected for annotation by the

algorithm), but it would count as five occurrences in the terminometrics analysis.

The platform also facilitates the management of terminometrics studies by providing many features: an integrated storage and search capability on domain-specific corpora, a user interface specifically designed to facilitate annotation by providing in-context display of a term to validate, an access to a term list with the possibility for addition and removal of terms, and so on. This is an improvement over the traditional manual handling of documents and term lists, instances generation and annotation, traditionally done with folders and spreadsheets. While the upper limits of the platform have not been tested explicitly, the current experiment was done with a term list of 1,036 entries on a corpus of over 220,000 documents. As far as the sizes of the corpora and vocabulary are concerned, the platform is mainly limited by the speed and capacity of the computer that runs it.

7 Conclusion and future work

In this article, we introduced *term sense disambiguation*, a close cousin to word sense disambiguation, but much less studied within the NLP community. We showed how terms, especially single-word terms, are polysemous, both in term banks and in specialized corpus.

We presented the idea of using active learning within our terminometrics application, in which the in-domain sense bias is quite strong. So far, we have implemented a simple active learning algorithm, and will move toward more complex ones in the near future. The annotation platform, ready for experimentation, will allow terminologists to further complete, in less time, the annotation process of the nanotechnology domain and other domains. This will provide test data, on which we can measure the different gains in terms of time and accuracy of our current and future active learning approaches.

Furthermore, we plan to push further our exploration of term disambiguation. In fact, although lexicographic and terminological resources are organized differently, the distinction between terms and words is not always that "clear-cut". Many single-word terms exist also as common words. Some specialized terms also migrate from specific domains to the general language (Meyer, 2000) when a specialized domain becomes more part of the day-to-day life of people (e.g. computer domain). We believe there is much room to further study term polysemy in term banks, in specialized corpus and also in more general corpus where both specialized and common senses might be present.

One of the envisioned experiments is to annotate semi-automatically a whole corpus to be able to compare the current approach to a supervised learning method. This will enable us to evaluate the contribution of active learning on the raw performance of disambiguation and time reduction of the annotation task. A new dataset related to a domain different than nanotechnology will also be defined for this experiment to avoid evaluating the approach on the dataset used for development.

Acknowledgments

We thank the annotators Julián Zapata and Barış Bilgen.

References

Caroline Barrière. 2007. La désambiguïsation du sens en traitement automatique des langues (TAL): l'apport de resources terminologiques et lexicographiques. In Marie-Claude L'Homme and Sylvie Vandaele, editors, *Lexicographie et Terminologie: compatibilité des modèles et méthodes*, pages 113–140. Presses de l'Université d'Ottawa.

Caroline Barrière. 2010. Recherche contextuelle d'équivalents en banque de terminologie. In *Traitement Automatique des Langues Naturelles 2010*.

Yee Sang Chan and Ht Ng. 2007. Domain adaptation with active learning for word sense disambiguation. *Acl*.

Jinying Chen, Andrew Schein, Lyle Ungar, and Martha Palmer. 2006. An empirical study of the behavior of active learning for word sense disambiguation. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 120–127.

Marta Chroma. 2011. Synonymy and Polysemy in Legal Terminology and Their Applications to Bilingual and Bijural Translation. *Research in Language*, 9:31–50.

Ingrid Meyer. 2000. Computer Words in Our Everyday Lives : How are they interesting for terminography and lexicography ? In *Euralex'2000, International Congress on Lexicography*, pages 39–58, Stuttgart, Germany.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

M Palmer and Jy Chen. 2005. Towards robust high performance word sense disambiguation of English verbs using rich linguistic features. *Natural Language Processing - Ijcnlp 2005, Proceedings*, 3651:933–944.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 613–619, New York, NY, USA. ACM.

Jean Quirion. 2003. Methodology for the design of a standard research protocol for measuring terminology usage. *Terminology*, 9(c):29–49.

Jean Quirion. 2006. Terminometrics - an Evaluation Tool of/for Term Standardization. In *TSTT'2006 - International Conference on Terminology, Standardization and Technology Transfer*, pages 19–24, Beijing, China.

Egidio Terra and C.L.A. Clarke. 2003. Frequency Estimates for Statistical Word Similarity Measures. In *Proceedings of the NAACL 2003*, page 165.

Tin Kam Ho. 1995. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282.

Radek Vogel. 2007. Synonymy and polysemy in accounting terminology: fighting to avoid inaccuracy. In *Proceedings of the English for Specific Purposes Terminology and Translation Workshop, Košice 13-14 September 2007*. Univerzita P.J. Šafárika.

Jingbo Zhu and EH Hovy. 2007. Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem. *EMNLP-CoNLL*.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. 2008. Learning a Stopping Criterion for Active Learning for Word Sense Disambiguation and Text Classification. *International Joint Conference on Natural Language Processing*, pages 366–372.

Syntagmatic Behaviors of Verbs in Medical Texts : Expert Communication vs. Forums of Patients

Ornella Wandji Tchami, Natalia Grabar
STL UMR 8163 CNRS, Université Lille 3
59653 Villeneuve d'Ascq, France
ornwandji@yahoo.fr
natalia.grabar@univ-lille3.fr

Ulrich Heid
IWIST, Universität Hildesheim
Germany
heidul@uni-hildesheim.de

Abstract

In this paper, we propose an automatic contrastive analysis of the behavior of verbs, with regard to the semantic features of their arguments (subject, direct object, indirect object), within and across medical subcorpora. We compare four medical subcorpora with texts whose authors and intended readership have different levels of expertise. The semantic annotation of the subcorpora is based on semantic information provided by a medical terminology. Our results indicate that the proposed procedures and tools could be used for the automatic detection of different ways of expressing medical concepts and conceptual relations, according to the types of texts.

1 Introduction

Research has shown that despite the growing body of literature available to patients, communication between medical practitioners and patients is not always easy and successful. This situation is to some extent due to linguistic complexity in medical care texts (Putz (2008)). Indeed, the availability of medical information does not guarantee its readability and correct understanding. Standard medical language contains specific terminology and specialised phraseology which is hard to understand for non-expert users (McCray (2005), Zeng-Treiler et al. (2007)), and which can therefore render the communication difficult (Jucks and Bromme (2007), Tran et al. (2009)). Research into this issue has been conducted in sociology (Kharrazi (2009), Chy et al. (2012)), in Medical Informatics (Kokkinakis and Toporowska Gronostaj (2006), Smith and Wicks (2008)) and in Natural Language Processing (Zeng-Treiler and Tse (2006), Chmielik and Grabar (2011)) in order to identify the specificities of this communication. As one could expect, these studies suggested the simplification of the medical doctors' vocabulary. Researchers in NLP went further, proposing the creation of lexicons which relate expert terminology with expressions used by lay people (Zeng-Treiler and Tse (2006), Deléger and Zweigenbaum (2008), Grabar and Hamon (2014)).

In line with the studies mentioned above, we are interested in the written communication between medical experts and non-experts. We propose a comparative analysis of the distributions of argument structures (and semantic patterns) in French medical texts which have been classified and grouped according to their discursive specificity (Pearson (1998)) and the respective level of expertise of the target public. More specifically, we compare verbal arguments in four types of subcorpora, focusing on lexical preference and making different hypotheses. We assume that medical experts use more specific and specialised verbal configurations (frames, co-occurrences, collocations (i.e preferred co-occurrences)) in order to express medical concepts and the relations between them, while non-experts tend to use less specific configurations. Also we verify to which extent the semantic categories of the Snomed terminology allow to distinguish these different configurations. Our study is an extension to a previous work where we looked at the syntactic and semantic features of the elements surrounding the verbs in the expert and forum subcorpora, without taking into consideration the intermediary subcorpora and the dependency relationships between the verbs and their arguments. This work is intended to highlight the relationship between verbal argument structures and the different ways of expressing specialised concepts in texts written by people who have different levels of specialised medical knowledge. In fact, lexical preferences, collocations, semantic category preferences and verb frames share the ability to express concepts and/or relations between concepts.

2 Studies of argument structures in corpora

Investigations into the distribution of argument structures of verbs have helped describe and understand the relationship between the verbs, the argument structures they occur in and the semantic classes to which they belong. These studies have shown the tendency of particular verbs to select a particular type of arguments, and the attraction of certain argument structures for particular verbs (Gries and Stefanowitsch (2004), Gries and Stefanowitsch (2010)). Some studies focusing on verb valency patterns and their frequencies have revealed that verbs show certain preferences with respect to their valency schemes and alternations (Köhler (2005), Engelberg (2009), Cosma and Engelberg (2013)). Other researchers have automatically induced verb classes from data on the distribution of valency patterns (Schulte im Walde (2003), Schulte im Walde (2009)).

Quantitative data on argument structures are also used for the construction of lexical classes, or to build a lexical organisation which predicts much of the behaviour of a new word by associating it with an appropriate class. As far as English is concerned, several studies were conducted for the acquisition of subcategorisation information from raw corpora (Briscoe and Carroll (1997); Preiss et al. (2007)). Some of these studies like Korhonen and Briscoe (2004) use subcategorisation frames for the extension of lexical-semantic classifications. Others use them as main features for the classification of verbs in specialised texts from the biomedical domain (Korhonen et al. (2008)). Only recently, French has become the target of such research. Chesley and Salmon-Alt (2006) carried out an exploratory study of 104 common verbs that allowed them to identify 27 subcategorisation schemes. More recently, Messiant et al. (2010) have implemented a method to automatically acquire a syntactic lexicon of subcategorisation frames for French verbs from large corpora.

It has been shown that the neighborhood of a verb can be different according to the type of text in which the verb appears (Helbig (1985), Wandji Tchami et al. (2013), Wandji Tchami and Grabar (2014)). Roland and Jurafsky (1998) analyse how the frequency of verb subcatgorisation schemes is affected by corpus choice. This study has revealed that verb senses are closely related to types of discourse, in such a way that both determine the frequency of the different subcatgorisation schemes of the verbs in the corpora.

Although they all look at verbal argument structures within different types of texts, none of the above-mentioned studies proposes the kind of approach we are trying to develop. We propose a study of subcategorisation schemes in medical corpora that are differentiated according to their levels of specialization, and we use a medical terminology for the semantic annotation of the texts, to detect selectional restrictions and lexical preferences.

3 Material

The study is based on two types of material: corpora distinguished by the levels of expertise of their authors and intended readers (section 3.1) and a semantic resource (section 3.2), used for the semantic annotation of the corpora.

3.1 Corpora

The corpus is made up of a set of four medical subcorpora of written French, which are distinguished by their discursive specificities (Pearson, 1998) and the respective levels of expertise of their readership. The first three subcorpora come from the portal CISMeF[1], which indexes medical texts according to three different categories: texts for medical experts, texts for medical students, texts for patients or non-experts. The fourth subcorpus is made of texts written by non-experts. It contains discussions between patients and/or persons participating in a forum called *Doctissimo, Hypertension, Problèmes Cardiaques (Doctissimo, Hypertension, heart problems)*[2].

Corpus	Size	Verb occ.	pron. occ.	description
C_1 / *expert*	1,285,665	52529	1349	scientific publications and reports
C_2 / *student*	384,381	22092	920	didactic supports created for students
C_3 / *patient*	253,968	19421	1176	documentation and brochures
C_4 / *forum*	1,588,697	184843	8261	forum messages from participants

Table 1: Size of the subcorpora used

Table 1 indicates the size of the four subcorpora (number of tokens) and the number of verbal oc-

[1] *http://www.cismef.org/*

[2] *http://forum.doctissimo.fr/sante/hypertension-problemes-cardiaques/liste_sujet-1.htm*

currences per subcorpus; the rightmost column indicates how many verbal occurrences per subcorpus have pronominal arguments (which will not be resolved and thus not counted in this study). As can be seen, the expert and forum corpora are almost equal in size, while the student and the lay persons' corpora are much smaller, but also similar in size. We make the assumption that the authors of the four subcorpora represent actors of the medical domain, who have different levels of expertise as far as the use of specialised medical language is concerned.

3.2 Semantic resource

We use the *Snomed International Terminology* (Côté (1996)) which groups medical terms into eleven semantic categories, of which nine are considered in this study[3]. This terminology was chosen because it is one of the largest medical terminologies available for French.

$\mathcal{T}$: Topography or anatomical locations (*e.g.*, *coeur (heart), cardiaque (cardiac), digestif (digestive), vaisseau (vessel)*);

$\mathcal{S}$: Social status (*e.g.*, *mari (husband), soeur (sister), mère (mother), ancien fumeur (former smoker), donneur (donor)*);

$\mathcal{P}$: Procedures (*e.g.*, *césarienne (caesarean), transducteur ultrasons (ultrasound transducer), téléexpertise (tele-expertise)*);

$\mathcal{L}$: Living organisms, such as bacteries and viruses (*e.g.*, *Bacillus, Enterobacter, Klebsiella, Salmonella*); plants (*e.g.*, *fougère (fern), pomme de terre (potato)*), but also animals (*e.g.*, *singe (monkey), chien dalmatien (dalmatian dog)*);

$\mathcal{J}$: Professional occupations (*e.g.*, *équipe de SAMU (ambulance team), anesthésiste (anesthesiologist), assureur (insurer), magasinier (storekeeper)*);

$\mathcal{F}$: Functions and dysfunctions of the organism (*e.g.*, *pression artérielle (arterial pressure), métabolique (metabolic), protéinurie (proteinuria), détresse (distress), insuffisance (deficiency)*);

$\mathcal{D}$: Disorders and pathologies (*e.g.*, *obésité (obesity), hypertension artérielle (arterial hypertension), cancer (cancer), maladie (disease)*);

$\mathcal{C}$: Chemical products (*e.g.*, *médicament (medication), sodium, héparine (heparin), bleu de méthylène (methylene blue)*);

$\mathcal{A}$: Physical agents and artefacts (*e.g.*, *cathéter (catheter), prothèse (prosthesis), tube (tube)*).

In our approach, the semantic categories of the Snomed International terminology are considered as ontological categories used for the characterisation of the verbal arguments. The used version of Snomed contains 144 267 entries (mainly French nouns, noun phrases and adjectives). We used it for the semantic annotation of our corpus. The Snomed entries may not necessarily cover all domain notions in our texts (Chute et al., 1996). For this reason, in a previous study, we attempted to complete the coverage of the terminology in relation with the corpus used (Wandji Tchami and Grabar (2014)). We computed the plural forms of Snomed's single word terms, and we tried to detect misspellings of the terms by means of the string edit distance (Levenshtein, 1966). In both cases, the computed forms inherit the semantic type of the terms from the Snomed. In this way, 14 035 entries were added to the terminology.

4 Method

The method applied in this study aims at describing and comparing the argument structures of verbs in different types of subcorpora, with a particular focus on selectional restrictions and lexical preferences. The tools and procedures used allow us to detect collocations and different ways of expressing concepts and conceptual relations. In order to achieve our aim, we follow 3 main steps: the corpus pre-processing and annotation (syntactic and semantic) (section 4.1), the extraction of verbal argument structures and co-occurrence data (section 4.2), both performed automatically and followed by a manual analysis (section 4.3) which aims at contrasting and interpreting the automatically extracted data.

4.1 Corpus pre-processing and annotation

The subcorpora have all been downloaded from the above-mentioned online sources, converted into plain text and recoded in UTF-8 format. The syntactic analysis of sentences is performed with the Cordial dependency parser (Dominique et al., 2009). Its output contains sentences in a tabulated

[3]The two semantic classes containing modifiers are not taken into consideration in this study.

format similar to the CONLL format (Buchholz and Marsi, 2006). In this format, a sentence consists of one or more tokens, each one annotated with thirteen fields, separated by a tab character. Among these fields, the *syntactic function* and the *pivot* verb are the main information that allow us to extract the verbs and their arguments.

The syntactically annotated sentences are then processed with Perl programs that perform the semantic annotation by projecting the resource described in Section 3.2 onto the lemmatised sentences. The categories of the terminology add semantic information to the syntactic patterns of verbs. Hence, at the end of this stage, each verb argument appearing in the terminology is labeled with a semantic category, in addition to its syntactic function; such pair constitutes what we call *a specialised configuration or frame* while a pair whose argument has no Snomed categories is considered as a *non specialised configuration or frame*.

4.2 Extraction of verbal argument structures and of verb+noun co-occurrence

The sets of sentences annotated at the previous step are processed with Perl programs that extract argument structures involving the Snomed categories of terms, when provided by Snomed, as in Table 2 (*V+Su/Scat+DO/Scat, V+Su/Scat+DO/Scat+IO/Scat*) and pairs of *V+Su/Scat*, *V+DO/Scat* and *V+IO/Scat*[4].

For each verb, the most frequent cooccurring objects are automatically extracted and their corresponding frequencies are computed from all subcorpora. Indeed, in 5.1 and 5.2, we focus particularly on direct objects, except with the verb *exposer* for which we have considered the subject (*patientS+exposer*) and the indirect object (*exposer un risque*) (Table 2).

For a given verb A, after extracting its most frequent objects from the corpora, we automatically extract further verbs that frequently combine with A's objects, most particularly those which are semantically close to A, and we compute the frequency of all verb+Object pairs (see Tables 3 and 4). These data function as indicators of the phenomena observed on the medical language of experts and non experts. Indeed, this experiment helps to identify semantic groups of verbs expressing similar concepts and conceptual relations between the verb arguments.

After processing all the verbs found in the different subcorpora, 11 verbs were selected for a more detailed case study : augmenter (*add*), évaluer (*evaluate*), exposer (*expose*), subir (*undergo*), prescrire (*prescribe*), provoquer (*provoke*), accompagner (*accompany*), suivre (*follow*), causer (*cause*), baisser (*lower*), and entraîner (*lead to*). These verbs were selected according to two main criteria:

- Frequency: the verbs should have at least 20 occurrences each, in at least two of the subcorpora;
- Types of verbs: we tried to choose not only verbs that intuitively tend to have specialised usages in specialised domain texts, but also general language verbs like *accompagner*, *baisser*, *suivre* etc.; The tendency to co-occur frequently with particular terms was also taken into consideration, since we focus on lexical preference and collocation.

4.3 Comparative analysis of verbal behaviors

The comparative analysis is done manually and aims at highlighting the differences and similarities of the subcorpora with regard to selectional restrictions and lexical preferences. We compare the frequency of verbal configurations (pairs of verb+argument or frames) across the subcorpora. This analysis addresses different aspects : the arguments (terms) cooccurring with verbs, the verbs cooccurring with those arguments, the different frames verbs frequently appear in, and argument structures expressing similar conceptual relations. The results are discussed in Section 5.1.

5 Results and Discussion

5.1 Terms cooccurring with verbs

The data provided in Table 2 lead to several observations. Some verbs frequently select terms from a particular Snomed category, mostly specific terms, in a particular subcorpus, while in the other subcorpora this co-occurrence never happens or only happens scarcely. This phenomenon is particularly striking with verbs like *prescire* and *subir*. In the forum and sometimes in the lay subcorpus, these verbs frequently combine with

[4] V=verb, Su=sujet, DO, direct Object, IO=indirect Object Scat=Snomed category

Verbs	Nominal cooccurrents				
	Arguments	exp	stu	lay	for
prescrire	*traitement* $\mathcal{P}$	3	0	0	7
	examen $\mathcal{P}$	0	0	2	7
	médicament $\mathcal{C}$	0	0	7	26
subir	*ablation* $\mathcal{P}$	0	0	0	39
	intervention $\mathcal{P}$	6	0	1	30
	AVC $\mathcal{D}$	0	0	2	12
augmenter	*tension* $\mathcal{F}$	0	0	7	14
	risque/risque de $\mathcal{F}$	26	8	5	7
baisser	*tension* $\mathcal{F}$	0	0	4	18
exposer	*à+risque* $\mathcal{F}$	14	8	0	3
	patient $\mathcal{S}$	23	5	1	0
suivre	*apparition de symptômes* $\mathcal{F}$	5	0	0	0
	patient $\mathcal{S}$	6	0	0	0
	régime $\mathcal{F}$	1	0	0	5
	conseil	0	0	4	10
	traitement $\mathcal{P}$	2	2	1	13
évaluer	*patient* $\mathcal{S}$	7	0	0	0
	indication	6	0	0	0
	risque $\mathcal{F}$	9	2	0	1

Table 2: Most frequent verb/arg pairs: capital letter=the Snomed category, no capital letter=no category provided

terms belonging to category P (procedures); more specifically, *prescrire* seems to have an attraction for the terms *traitement* and *examen*, while *subir* has a strong attraction for *intervention* and *ablation* (which refers to a type of medical intervention (hyponym)). *Prescrire* also combines frequently with names of chemical products (C) and shows a particular attraction for the term *médicament*, while *subir* prefers terms referring to disorders and diseases (D), and more precisely the term *AVC* (*stroke*). These are preferred co-occurrences which are therefore seen as collocations.

Such collocations may involve polysemous verbs and their different readings. For example, in the expert subcorpus (and sometimes in the student subcorpus), *évaluer* and *suivre* tend to appear frequently with terms referring to functions of the organism (F) or to Social status (S). *Évaluer* seems to be attracted by *risque*, *indication* and *patient*. *Évaluer*+F means *to mesure, determine, calculate, gauge, quantify*, while *évaluer*+S means *to examine*.

The differences in verb/arg pair frequencies can lead to different interpretations. First of all, when the frequency difference is very important from the forum subcorpus to the expert subcorpus, this may signal some specificities of the laypersons' language. Indeed, while health care specialists share foundational domain knowledge based on formal education and professional experience, the patients' or non experts' medical language is characterised by the use of common expressions and collocations, sometimes involving technical medical terms (*prescrire un médicament*, *subir une ablation*, *subir un AVC*, *suivre un régime*) borrowed from the medical experts' language. According to researchers in Consumer Health Literature, such mixed phraseology is the result of social and cultural influence on language and they are acquired from formal and informal sources such as the internet (Zeng-Treiler et al. (2006), Zeng-Treiler and Tse (2006)). The frequent use of these expressions makes them progressively become part of everyday language. This could be a plausible explanation for the high frequency of expressions like *prescrire un médicament*, *subir une ablation* or *subir un AVC*, in the forum texts.

Secondly, looking at the results from the expert subcorpus to the forum subcorpus, we notice that sometimes the frequency difference is not very important. The explanation given above could once more apply here. Indeed, medical technical terms are quite often used by non-experts to describe medical concepts. On the other hand, when a verbal combination involving a particular Snomed category is very frequent in the expert subcorpus like *exposer + name of a medication* (*votre patiente est exposée au ramipril*), *évaluer + function* (*évaluer un risque*) while the verb is totally absent or very rare in the other subcorpora, we might deal with a highly specialised (expert) or expert language-specific usage of the verb.

5.2 Lexical preferences of the arguments for verbs

The results of Section 5.1 give an account of the lexical preferences of the verbs within and across the subcorpora. In this section, we investigate the lexical preferences of nominals in the expert and forum subcorpora. Tables 3 and 4 give the results of this experiment. These data were obtained as described in Section 4.2. The blue color represents the processed verb, the entries in the column *Arguments* are the most frequent arguments of the processed verb, and the red color represents a semantic group of verbs frequently combining with the corresponding argument in the given corpus. The numbers in bracket show the frequency of each pair verb+arg.

Depending on the corpus, certain terms frequently combine with particular verbs, in order to express a particular concept. For instance, as we can see in Table 2, the terms *médicament* and *traitement* are *prescrire's* favourite cooccurrents

Arguments	Verbal cooccurrents	
	Expert	Forum
médicament	indiquer(3), recommander(2) proposer(2)	prescrire
traitement	proposer(8), envisager(7) recommander(3), imposer(3)	
examen	imposer(1), proposer(1) recommander(1), autoriser(1)	
intervention	-	subir
ablation	faire(1)	
AVC	prsenter(4), faire(2), avoir(2)	
tension	-	baisser
régime	-	suivre
conseil	considrer(1)	
traitement	recevoir(12), bénéficier(6) faire(6), poursuivre(3),	
tension	-	augmenter

Table 3: Lexical preferences of arguments in the expert subcorpus.

Arguments	Verbal cooccurrents	
	Forum	Expert
patient	traiter(1), voir(1)	suivre
apparition de	-	
symptome	expliquer (5)	
risque	mesurer(1), juger(1), exposer(23)	évaluer
patient	-	
indication	apprécier(1)	
risque	accroître(3), multiplier(2)	augmenter
	élever(1),	

Table 4: Lexical preferences of arguments in the forum subcorpora.

in the forum and sometimes in the lay subcorpus, while in the expert subcorpus, the terms frequently co-occur with the verbs *indiquer*, *recommander*, *proposer*, and *envisager*, *recommander*, *proposer*, *imposer*, respectively.

1) Ces médicaments ne sont plus recommandés en première intention dans le traitement de l'hypertension (*These drugs are no longer recommended as first-line in the treatment of hypertension*)

Although the two groups of verbs combine with the same terms, in the professionnal language, these verbs are not semantically equivalent, they correspond to different levels of evidence. Indeed, they are used by medical experts to express the relevance of prescribing a given drug or treatment for a given disease. In contrast, patients just know about the drug or treatment they have been prescribed for their disease but do not necessarily know about these distinctions. These examples highlight a very relevant difference in the way experts and non-experts use verbal configurations : the first choose very specific and technical configurations while the others use more general ones.

In the expert subcorpus, several sentences are in the passive voice with an omitted agent, as in Example 1. This applies to some of the above-mentioned verbs and is quite reccurrent with other verbs.

The lexical choice difference within subcorpora does not only concern terms. Verbs also select particular terms to combine with, depending on the subcorpora. For example, in the forum subcorpus, the verb *suivre* frequently co-occurs with the term *conseil*, while in the expert subcorpus, the term *conseil* does not combine with this verb. Instead, *suivre* combines with *indication*. The latter and mainly the term *recommandation*, which are semantically close to *conseil*, are very frequent in the expert subcorpus. They appear in positions where *conseil* could appear. For example, *recommandation* is combined with verbs like *proposer (4), appliquer (4), actualiser (8), publier (4), élaborer (2)* and *faire (3)*. This seems to show that the experts prefer to talk about *recommandations* and *indications* which have specific and technical meanings, while laypersons are more familiar with the term *conseil* which is a common word.

Another observation was made based on the experiment carried out. In the forum subcorpus *baisser* and *augmenter* frequently co-occur with the term *tension* (*augmenter la tension* (*increase blood pressure*), *baisser la tension* (*reduce blood pressure*) (see Table 2)), expressing different states of the blood pressure. In the expert subcorpus, none of these collocations were found. In addition, among the verbs combining with *tension* in the expert subcorpus, none is semantically related to the two verbs. However, we have noticed the presence of verb based nominalisations, constructions requiring support verbs or relational adjectives, which are synonymous with the two above-mentioned collocations : *élévation tensionnelle (4)*, and *hausse de tension (1)* correspond to *augmenter la tension*, while *réduction tensionnelle (2)*, *abaissement tensionnel (2)* and *baisse de tension (4)* have the same meaning as *baisser la tension*.

This phenomenon is consistent with the results obtained in a previous study (Wandji Tchami and Grabar (2014)) and with Condamines and Bourigault (1999)'s findings which confirmed the fact that nominal entities tend to be more frequent in expert texts than in non-expert texts. The above data demonstrate that the difference between the expert and forum texts does not lie in verbs alone, but mostly in the different types of constructions

the verbs are involved in (support verb, paraphrase, verb-based nominalisation, etc.).

5.3 Verbal frames and conceptual relations

Table 5 shows frames which represent different ways of expressing the cause-effect conceptual relation. The data were extracted from the subcorpora, through the analysis of frames of *accompagner*, *causer*, *provoquer*, and *entraîner* which are causative verbs. We are aware of the fact that some of the numbers presented in this table are not high enough to draw conclusions. However, we found it important to report them because they might highlight phenomena that could be further analysed in future work, with more data.

verbs	accompagner		causer		provoquer		entraîner	
frames	pro	for	pro	for	pro	for	pro	for
$C_s\ D_$do	1	0	2	1	0	11	3	1
$C_s\ F_$do	1	0	0	1	1	5	3	0
$D_s\ D_$io	5	3	0	0	0	0	0	0
$C_s\ D_$do	3	0	3	5	0	10	6	0
$D_s\ F_$do	5	1	1	1	3	8	3	0
$F_s\ F_$do	4	3	4	5	0	32	3	2
$F_s\ D_$do	1	0	1	0	0	12	5	1
$F_s\ P_$do	0	2	1	0	3	0	2	1
$P_s\ F_$io	5	0	0	0	0	0	0	0
$P_s\ D_$do	2	0	0	0	2	2	4	7
$P_s\ F_$do	0	0	0	0	1	0	5	0
$P_s\ P_$do	0	1	0	1	0	0	5	0
$F_s\ F_$io	6	3	0	0	0	1	0	0

Table 5: Frames: *s*=subject, *do*=direct object, *io*=indirect object; capital letters=Snomed semantic categories.

Many frames were identified, Table 5 shows the most frequent ones which are : $F\ D$, $P\ F$, $F\ F$, $P\ D$, $D\ F$, $F\ P$, $F\ D$, $P\ P$, $C\ D$, $D\ D$. These frames are all found in the four subcorpora but they tend to choose specific verbs depending on the subcorpus. The difference mostly lies on the lexical level with the choice of verbs. In the above-mentioned frames, the left side semantic class provokes or entails an effect or consequence that is expressed by the right side category. Let us take for example the relation *Functions-Functions* ($F\ F$), where a function of the organism has an effect on another function of the organism.

2) Exp: la prise de poids$_F$ s'accompagne d'une élévation de la pression artérielle$_F$ (*weight gain is followed by a rise in blood pressure*$_F$)

3) For: une diaphorèse$_F$ intense accompagne souvent la douleur$_F$ (*the pain*$_F$ *is often followed by an intense diaphoresis*$_F$)

4) For: le stress$_F$ provoque des spasmes vasculaires$_F$ (*stress*$_F$ *causes vascular spasms*$_F$)

As we can see from the data provided in Table 5, in the expert subcorpus, this conceptual relation is frequently expressed with the verbs *accompagner* and *entraîner* while in the forum texts the verbs *provoquer* and *causer* are the most used. This remark also applies for the other above-mentioned frames. Collocational differences between expert and forum verb use also involve differences in valency and syntactic construction. In Example 2, the verb *accompagner* is in a pronominal form with a reflexive pronoun *se/s'*; this construction is the most used one in the expert subcorpus, and in the table, it is represented by the presence of the indirect object in the frame.

Another tendency observed in the expert subcorpus is the frequent use of the passive voice with a syntactically omitted agent, while in the forum subcorpus, the active voice is the most used. This observation was already underlined in Section 5.2 with *recommander, indiquer* and *proposer*.

6 Conclusion and Perspectives

In this study, we have proposed a method for the comparative analysis of verbal argument structures in medical subcorpora whose authors and intended readership have different levels of expertise, with a focus on lexical preference. The main difference observed is that medical experts tend to choose verbal configurations with very specific and technical meanings which apply to specific situations, while non-experts use more generic and common verbal configurations. Lexical choice differences often come with differences in the syntactic constructions used. Indeed, medical expert writings are characterized by the frequent use of a passive form with an omitted agent. The analysis of the two intermediary subcorpora shows that the expert and student subcorpora are close to each other while the lay subcorpus is close to the forum. As far as the method is concerned, the use of a dependency parser seems to improve the results. However, a detailed evaluation of the parsing quality is still to be done. We are also planning to carry out the analysis exemplified here on more verbs.

References

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *In Proceedings of the ACL*, pages 356–363.

Sabine Buchholz and Erwin Marsi. 2006. Conll-xshared task on multilingual dependency parsing. In *In Proc. of CoNLL*, pages 149–164.

Jolanta Chmielik and Natalia Grabar. 2011. Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques. *TAL*, 51(2):151–179.

Christopher G. Chute, SP Cohn, KE Campbell, DE Oliver, and JR Campbell. 1996. The content coverage of clinical classifications. for the computer-based patient record institute's work group on codes & structures. *J Am Med Inform Assoc*, 3(3):224–33.

Anne Condamines and Didier Bourigault. 1999. Alternance nom/verbe : explorations en corpus spécialisés. In *Cahiers de l'Elsap*, pages 41–48, Caen, France.

Ruxandra Cosma and Stefan Engelberg, 2013. *Subjektsätze als alternative Valenzen im Deutschen und Rumänischen*.

Roger A. Côté, 1996. *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

Louise Deléger and Pierre Zweigenbaum. 2008. Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *AMIA 2008*, pages 146–50.

Laurent Dominique, Sophie Nègre, and Patrick Séguéla. 2009. L' analyseur syntaxique Cordial dans Passage. *Actes de TALN*, 9.

Natalia Grabar and Thierry Hamon. 2014. Automatic extraction of layman names for technical medical terms. In *ICHI 2014*, Pavia, Italy.

Stefan Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis. a corpus-based perspective on "alternation". *IJCL*, 9(1):97–129.

Gerhard Helbig. 1985. Valenz und kommunikation (ein wort zur diskussion). *Deutsch als Fremdsprache*, 22:153–156.

Regina Jucks and R. Bromme. 2007. Choice of words in doctor-patient communication: an analysis of health-related internet sites. *Health Commun*, 21(3):267–77.

Hadi Kharrazi. 2009. Improving healthy behaviors in type 1 diabetic patients by interactive frameworks. In *AMIA*, pages 322–326.

Reinhard Köhler. 2005. Quantitative untersuchungen zur valenz deutscher verben. *Glotometrics*, 9:13–20.

Dimitrios Kokkinakis and M Toporowska Gronostaj. 2006. Comparing lay and professional language in cardiovascular disorders corpora. In James Cook University Pham T., editor, *WSEAS Transactions on Biology and Biomedicine*, pages 429–437.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 2008. The choice of features for classification of verbs in biomedical texts. In *Proc. of COLING*, pages 449–456.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, 707(10).

Alexa McCray. 2005. Promoting health literacy. *J of Am Med Infor Ass*, 12:152–163.

Cédric Messiant, Kata Gábor, and Thierry Poibeau. 2010. Acquisition de connaissances lexicales à partir de corpus: la sous-catégorisation verbale en français. *TAL*, 51(1):65–96.

Jennifer Pearson. 1998. *Terms in Context*. John Benjamins, Amsterdam/Philadelphia.

Judita Preiss, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL*, volume 45, page 912.

Magdalena Putz. 2008. Approaching linguistic complexity in medical care. *International Journal of Anthropology*, 23(3-4):275–284.

Douglas Roland and Daniel Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of ACL*, Montreal, Quebec, Canada.

Schulte im Walde. 2003. Experiments on the automatic induction of german semantic verb classes. Technical report, Universität Stuttgart.

Catherine Smith and PJ Wicks. 2008. PatientsLikeMe: Consumer health vocabulary as a folksonomy. In *Proceedings of the AMIA 2008 Symposium*, pages 682–686.

Thi Mai Tran, H Chekroud, P Thiery, and A Julienne. 2009. Internet et soins : un tiers invisible dans la relation médecine/patient ? *Ethica Clinica*, 53:34–43.

Ornella Wandji Tchami and Natalia Grabar. 2014. Towards automatic distinction between specialized and non-specialized occurrences of verbs in medical corpora. In *Proceedings of Computerm*, pages 114–124, Dublin, Ireland, August.

Ornella Wandji Tchami, MC L'Homme, and Natalia Grabar. 2013. Discovering semantic frames for a contrastive study of verbs in medical corpora. In *TIA*, Villetaneuse.

Qing Zeng-Treiler and T Tse. 2006. Exploring and developing consumer health vocabularies. *JAMIA*, 13:24–29.

Qing Zeng-Treiler, Tony Tse, Guy Divita, Alla Keselman, Jon Crowell, and Allen C Browne. 2006. Exploring lexical forms: first-generation consumer health vocabularies. In *AMIA 2006*, pages 1155–1155.

The time factor as an associative concept relation in modelling post-liver transplant management complications

Paul Sambre
MIDI
Multimodality, interaction and discourse
University of Leuven, Belgium
paul.sambre@ku-leuven.be

Cornelia Wermuth
MIDI
Multimodality, interaction and discourse
University of Leuven, Belgium
cornelia.wermuth@ku-leuven.be

Hendrik J. Kockaert
QLVL
Quantitative lexicology and variational linguistics
University of Leuven, Belgium & University of The Free State, South-Africa
hendrik.kockaert@ku-leuven.be

Abstract

We propose a first termontological analysis of temporal parameters and relations applied to the case of medical complications in post-liver transplant management (PLTM). Medical complications contribute to different degrees of morbidity and mortality in the process of medical follow-up after transplant surgery. Understanding the full ontological and conceptual complexity of such complex spanning (SPAN) time events time is a central issue in drawing an implementable semantic map of the potential causes of early and long term complications, their diagnosis and the potential effects due to medical treatment. The analysis is usage-based and relies on linguistic utterances for complications in a concise medical review article.

1 Intro: medical complications and associative relations in termontology

This paper deals with medical complications and their termontological description. Termontology (Roche, 2007) combines insights from terminology and applied ontology, combining them with a linguistic dimension: whereas terminology is basically interested in complex tree representations between terms and their normalized definitions, applied ontology models (visual) representations of complex concept systems, including semantic relations that dynamically connect static entities in a terminological database. This paper combines both preoccupations, starting from descriptions of medical complications, more specifically those occurring in post-liver-transplant management (PLTM). In the area of liver transplantation, termontology improves understanding of conceptual structures based on the lexico-grammatical structures retrieved from scientific literature about complications following transplantation. So far, medical complications have not been the subject of much analysis in terminological or applied ontology. Yet, they represent a rich resource for investigating cause-effect relations that constitute a specific subcase of causal events. In a broader sense, a complication is any adverse, undesired and unintentional result of disease management. More technically, medical complications are iatrogenic (i.e. disease-related) subsets of cause-effect relations and refer to the (negative, adverse) morbid consequence of a disease or disorder resulting from unsuccessful disease management. Complications, occasionally known as suboptimal (negative) outcome, are not to be confused with sequelae of previous acute medical conditions and therapies. Negative outcomes can be subdivided into failure to cure (pre-existing conditions that remain unchanged after the transplantation procedure), negative sequel and complication (Clavien et al., 2006). Complications differ from the commonly expected sequelae resulting from an anterior acute medical condition and therapy. Understanding and dealing efficiently with postsurgical medical complications is a crucial issue for the improvement of healthcare quality, since complications require longer and more expensive treatment, and, more importantly,

will negatively affect patient prognosis. In extreme cases, they may lead to severe co-morbidity and even death.

Concept relations are a key concept in knowledge representation, because they interconnect the different concepts or entities in a given knowledge domain. Despite extensive work on concept relations (e.g. Khoo and Jin-Cheon, 2006), a descriptive, usage-based account of associative relations is lacking to date (Sambre and Wermuth, 2010). In the termontological approach, the focus has been on (static) hierarchical concept relations such as type-token and/or meronymic relations.

Biomedical terminology uses these vertical relations for unique designations of medical concepts and their terminological variants that result in a compendium of several controlled vocabularies such as the Unified Medical Language System (UMLS) and the controlled thesaurus of Medical Subject Headings (MeSH) (Grabar et al., 2012). These classifications display conceptual and so-called static snap shots of medical events. Biomedical ontologies, conversely, aim to explore snap shot relations against a dynamic background of temporal unfolding, which results in so-called SPAN (or spanning time) relations. This approach reflects the true conditions as medical concepts simultaneously express both hierarchical and associative (i.e. time-based) relations. The dynamicity of medical concepts is a phenomenon that deserves further study. In this paper, we therefore investigate in greater detail dynamic associative relations, focusing on the temporal underpinnings of causality in medical complications in post-liver transplant management (PLTM). The rationale is that causes, by definition, precede effects both from a logical and experiential perspective. Thus, the medical complications under investigation can be assumed to entail both causal and time-related relations. Our primary objective is to set out temporal parameters to be used in a conceptual model and to inventory specific associative time elements inherent to the concept of medical complication in PLTM.

2 Time in the conceptual map of medical complications

Time, in our view, is the underlying conceptual basis or background against which causality of (un)intentional medical events occur, as these events are or are not triggered by instrumental actions performed by medical teams. The following example taken from a specialized journal on transplantation in medicine should help to illustrate our objective. The example discusses a complication in PLTM, i.e. hepatic artery thrombosis (HAT): "[…] hepatic artery thrombosis (HAT) is the second main cause of liver graft failure. Moreover, HAT is the most common vascular complication in orthotopic liver transplantation (OLT). It is associated with a marked increase in morbidity, being the leading cause of graft loss (53%) and mortality (33%) during the immediate postoperative period. […] the consensus definition for early HAT was an arterial thrombosis detected during the first month after OLT. Late HAT was also defined as the event detected ≥ 1 month after OLT. […] The true incidence of early HAT is unknown, but it varies between 0% and 42%. […] improvements in postoperative care have led to a marked reduction in its incidence." (Pareja et al., 2010)

As can be seen from the example, there is clear linguistic evidence of both hierarchical and associative relations. The following observations can be made:

1. The linguistic expression (italics) designating the most common vascular complication, *the hepatic arterial thrombosis*, refers to a hierarchical type of thrombosis (kind_of relation).
2. Causal information is provided by the linguistic expressions *cause of liver graft failure, cause of graft loss and mortality*.
3. Implicit instrumental reference is made to the medical treatment in (complex) nouns such as *liver transplantation, graft failure* and *postoperative care*.
4. Temporal relations are set up between the medical treatment and the post-surgery adverse effect by means of prepositions expressing a relation of time such as *during the immediate postoperative period*, *during the first month after OLT*, or *≥ 1 month after OLT*.

From the above we can conclude that in medical discourse recurrent, grammatically complex linguistic patterns are used in order to connect (1) the entities and types, (2) the causes and (unintended) effects of health care, as well as (3) the instrumental treatment administered by medical doctors against a sequential background of (4) time. The interplay between these four kinds of relations may serve as a lexico-grammatical starting point for drawing up a conceptual map of the medical subdomain under investigation.

Such conceptual maps are fundamental for improving general procedures (in English) regarding

complications for medical care in **multi**disciplinary medical treatment. PLTM occurs in hospital teams, where physicians and nurses specialized in complementary fields of haematology, radiology, internal medicine, surgery and intensive care etc., collaborate in order to improve survival rates and reduce the impact of complications. Patient-centered healthcare implies collaborative settings that call for efficient IT support systems to monitor patient status and share information on the needs of patients. These maps are to be shared between the healthcare team's actors, providing the scenarios they share and the input of each team section or member based on their individual knowledge levels, their educational training, and the different services and platforms these persons work in.

3 A usage-based account of the conceptual structure of time

Our goal is to pinpoint the major dimensions of time in the description of complications in PLMT. For reasons of concision, we base our depiction on an often-quoted review article about PLTM (Moreno and Berenguer, 2006), a common genre in the medical scholar tradition, that summarizes available data for a given medical phenomenon. The article under investigation provides an overview of allograft dysfunctions and surgical complications following liver-transplant and discusses the state-of-the art concerning their medical follow-up. Strikingly, the article conceptually opposes immediate and long-term complications. These complications are of a different nature: they can be strictly medical (think of respiratory changes, renal dysfunction or hemodynamic complications), or technical (complications due to h(a)emorrhage or vascular complications resulting, for example, in infections or draft dysfunctions (major complications in this particular case are acute cellular rejection or recurrent viral hepatitis)). Both short and long-term complications are rather heterogeneous as well, due to the fact that the liver interacts with very different subsystems of the organism, whereby any dysfunction may cause diseases such as chronic rejection, arterial hypertension, obesity or bone complications, just to name a few. Our investigation is limited to the temporal aspects of the different complications, leaving aside other conceptual aspects such as anatomical location or severity.

3.1 Peri- and post-operative time

A first important time factor is the distinction between treatment peripheral to the central (intra-operative) transplant intervention that can have a direct impact on the reduction of post-surgical complications (Junttila et al., 2005), and the time lapse proper to the complication itself (the so-called post-operative time). Peri-operative time has to do with non-problematic follow up of surgery, before, during and after surgery. The difference between this peri- and post-operative time is minimal, given the fact that some complications such as infections may arise due to improperly performed medical actions:

(1) The prophylaxis of bacterial infection includes the following strategies: a) selective intestinal decontamination; b) administration of systemic antibiotics peri-operatively, c) antibiotic prophylaxis before invasive explorations of the biliary tract, and d) personnel hand washing together with strict asepsis in all invasive procedures.

The following observations can be made about example (1): strategies a), b) and c) in the example refer to such perioperative precautions. As a part of medical prevention, peri-operative treatment contributes to building the temporal barrier *ab quo* the time sequence of complication starts to run. Preventive, pre-operative treatment is an important time issue that should be taken into account in modelling medical complications: a distinction is needed between pre-symptomatic (example 2) and post-symptomatic treatment of complications. Prophylaxis aims at non-invasive avoidance of complication outbreak (example 3) and therefore reduces medical cost (example 4).

(2) Another form of prevention, mainly targeted to avoiding the development of clinically manifest CMV disease, is the treatment of infection in the pre-symptomatic stage. [81; note of the authors: CMV refers to cytomegalovirus, the most frequent micro-organism in liver transplantation]

(3) Universal prophylaxis is useful mainly in high-risk patients […] and can be done effectively and safely with oral drugs […]. [81]

(4) Anticipated treatment is also an effective and probably most cost-effective strategy.

The distinction between peri- and post-operative time as classification parameter is a central issue in the temporal (when do problems occur?), causal (what effects are produced?) and instru-

mental (what therapy positively affects such effects?) format of medical decision-making. In medical discourse, peri-operative techniques are explicitly juxtaposed to post-operative care, as the following example shows.

(5) The results of liver transplantation have improved due to advances in perioperative technique, a better understanding of the course and prognosis of several [sic] liver disease improved immunosuppressive therapy and more effective postoperative care. [77]

In fact, knowledge about postoperative complications is actively used in optimal peri-operative treatment.

A second time factor at the interface between peri-operative management and post-operative complications is the moment of detection of (early) complication symptoms. Detection and diagnosis of new symptoms clearly marks the distinction between prophylaxis and post-operative care. The following excerpt illustrates this distinction:

(6) Thus, knowledge of complications that emerge during follow up period, early and accurate establishment of diagnosis, and prompt institution of appropriate interventions are essential for optimal patient and graft outcome [77]

An important part of the state-of-the-art consists in describing so-called early detection methods such as in the following example:

(7) Methods for early detection of viral infection, in the case of cytomegalovirus, are periodic determination of CMV antigenemia in peripheral blood leukocytes and PCR techniques to detect the blood viral genome.

Defining medical states is a central issue in liver transplant surgery. The literature defines the normal state in the intensive care unit, after transplantation that shows increasing degrees of recovery as illustrated in the following example by the different modified verbal and nominal phrases:

(8) When the transplant evolves favorably, the patient is awake, hemodynamically stable, with spontaneous respiration, preserved renal function, and with progressively improving liver activity.

3.2 Time of occurrence: immediate and long-term

Complications are possible alterations to the desired optimal condition. The most frequent in PLTM are complications that can be expected during this early post-transplant period are hemodynamic alterations, and respiratory, renal and neurological complications.

A global distinction is the one between early and late complications, or more correctly, between immediate and long-term complications. Late complications are gaining importance as survival rates during the early-postoperative period increases. The distinction between complication subtypes is based on the point of their occurrence in time:

(9) The complications occur either immediately post-transplantation or in the long-term. The main complications in the immediate postoperative period are related to the function of the graft (dysfunction and rejection), the surgical technique, infections (bacterial, fungal, and viral), and systemic problems (pulmonary, renal, or neurological). In the long term, the complications are typically a consequence of the prolonged immunosuppressive therapy, and include diabetes mellitus, systemic arterial hypertension, de novo neoplasia, and organ toxicities, particularly nephrotoxicity.

An important note is that the underlying pathology causing the transplantation is not considered a complication, though the causal trigger may persist (or reemerge at some later point in time).

The definition or discursive description of complications typically contains the designation of the medical phenomenon, a general characterization in terms of immediate or late occurrence, followed by a more precise time label for the time span within which complications arise (in terms of hours, as in (10), days, or, in the case of late complications, months).

(10) A hemorrhage in the immediate postoperative period is another potential complication […].It is typically diagnosed within the first 48 hours post-transplantation (hemorrhagic abdominal drainages, hemodynamic instability, serial determination of the hematocrit/hemoglobin).

In the above example, the time label is followed by a summary of medical actions performed during this time lapse. Modelling complications then may entail two different time lines: one for occurrence of complications as such (snap and span, notions defined in the introduction of this paper), and one mapping the full (linear or cyclical) scenario of common medical actions associated with postoperative care.

The review article also mentions a shift in the historical evolution of PLTM complications: as surgical techniques and immunosuppressive treatment improve, prognosis and survival chances increase, correspondingly extending the time span of (late) medical post-surgical follow-up (cfr. (11) and (12)).

(11) The main barriers to overcome in the first period were immediate post-surgical survival together with prevention of acute rejection.

(12) With greater survival of patients, new problems have arose that basically affect transplant recipients with long-term follow up

Long-term complications are rather flexible conceptual notions. Their emergence is connected with the specific complication (such as chronic renal failure, systemic arterial hypertension, diabetes mellitus, etc.), but at the same time differs accordingly. Generic time information is commonly expressed as occurring at a random moment (13), or by means of unspecified post-operative time (14).

(13) […] malignant tumors can appear at any time after transplantation […].

(14) A variable percentage of patients, 4-20% according to the series, will develop diabetes mellitus following transplantation (de novo DM).

Apart from these general time labels, also different discursive strategies are used for indicating a precise moment in time, specifically in late complications. Here, different scenarios may occur. In the first one a precise numeric cut-off point after transplantation is expressed (*not until, not before*):

(15) Chronic rejection is usually not evident until at least 6 months after transplant. The pathogenesis is still unclear.

In the second scenario, the complication is associated with a risk decreasing in time, without mentioning an endpoint:

(16) Arterial hypertension (AHT) is a frequent complication in liver transplant recipients. Its prevalence varies between 50-70% in the first post-transplantation months but decreases thereafter probably due to the reduction of the immunosuppressive doses.

In the third scenario, the full span of time is expressed (e.g. a one-year period) during which the complication is most pronounced:

(17) Obesity is a very frequent complication in transplanted patients […] one year after transplantation, the period when the greatest weight gain is seen.

3.3 Timing of occurrence and diagnosis

An interesting issue is the fact that even within the subgroup of immediate complications, both the nature of the complication and its treatment depend on the moment of discovery: similar sets of symptoms may cause different kinds of pathology. This is the case, for example, for the most frequent complication in pediatric cases:

(18) Symptoms are highly variable and depend on the timing of development and diagnosis.

(19) When the thrombosis occurs at an early stage, it typically leads to ischemia/necrosis of the graft; in contrast, when it occurs at a later time point, it generally leads to biliary complications (intrahepatic biliomas and biliary stenosis) but with preservation of the graft function.

This evolutionary, dynamic nature of medical conditions needs to be taken explicitly into account in concept modelling of complications. In its earlier or later diagnostic establishment, a complication takes up different forms and therefore requires different treatment types. Consequently, in the case of thrombosis mentioned in the previous example, the therapy in the acute or late form consists of different medical actions:

(20) In the acute form, thrombolysis can be accomplished by surgical radiology. Arterial thrombectomy may be an alternative that can be done either by interventional radiology or surgical intervention. In patients where these options fail, urgent re-transplantation may be required. In the late form, treatment is mainly focused to prevent/treat biliary complications derived from the thrombosis.

Note that in the late form, treatment does not zoom in on the primary complication, but on the derived one. This example clearly shows that one complication may trigger another one. The same complication may be connected with or caused by different intentional operations, leading to these unintended side effects at different stages in post-operative care:

(21) Biliary fistula can occur initially in the first month in relation to anastomotic dehiscence secondary to technical errors or biliary tract ischemia. It is also a common complication in the third month when the T-tube is withdrawn.

A central issue in this dynamic picture is the notion of *lead time to diagnosis*: given the fact that complications constantly evolve, this time factor

contributes to variable medical decisions, taking transplantation as a starting point:

(22) The clinical picture is variable and depends on the time of development, lead time to diagnosis, and existence of a T-tube.

We finish this part of our investigation with two specific subcases of timing. First, some complications can occur both early, or emerge only late, after a so-called normal postoperative course. The following example refers to liver graft dysfunction:

(23) Dysfunction of the graft may occur in the immediate postoperative period (early dysfunction) or late during the follow-up of the patient {typically related to the recurrence of the original disease (viral hepatitis, primary biliary disease, sclerosing cholangitis, alcohol or autoimmune liver disease) or chronic rejection}.

This later complication's manifestation has a very distinct causal origin: it is overtly tied to the original disease, which urged for liver transplantation. Second, different complications (as the neurological states) sometimes occur simultaneously: the first complication is continuous *(disorientation)*, and then punctuated by episodes of the second one *(agitation and confusion)*.

(24) The most frequent neurological alterations are disorientation with episodes of agitation and confusion.

A very specific feature of generic late complications such as malignant tumor is the correlation between duration (of immunosuppression) and specific cancer subtypes *(Kaposi's sarcoma, skin tumors, carcinomas of vulva and perineum):*

(25) Although malignant tumors can appear at any time after transplantation, Kaposi's sarcoma followed by lymphoproliferative disorders are the earliest that usually develop. The later ones are skin tumors and carcinomas of the vulva and perineum.

3.4 Duration of complication: rejection from (hyper-) acute to chronic

Some complications occur either early or late. A specific case is graft rejection. An interesting time distinction in this respect is the one between hyperacute and acute. There is not only a difference in terms of time itself (hyperacute rejection occurring within minutes or hours), but also in terms of the complications' nature. The prefix *hyper-* refers as well to the severity of the rejection reaction and the fact that antibodies reject the graft in an irreversible way. Hyperacute rejection is humoral, whereas acute rejection has a cellular origin. This is the primary difference with *acute* rejection, for which drugs are available. The primary difference with the third subtype is that

(26) […] Chronic rejection generally occurs over a span of months, can be unresponsive to current therapy, and contributes to be a source of graft loss.

An interesting, yet unsolved issue is the difference between chronic and repeated acute rejection. Some studies report that acute rejection generally occurs in the first few weeks following transplantation, whereas chronic rejection "typically occurs several months to a year posttransplantation" (Batts, 1999) and requires additional more histological fine-tuning by means of liver biopsy.

Determining correct diagnosis for each of these complications is a highly complex issue, because of the many clinical parameters shared by different complications.

3.5 Frequency and prevalence

Frequency and prevalence are commonly used terms when describing the epidemiological status of a complication (Greenberg et al., 2005, chapter 2). Whereas incidence refers to the number of new cases occurring in a given period of time, prevalence indicates the actual number of cases alive either during a period or at some point in time.

The prevalence can be addressed in different ways: on average and based on variation (either within the different complications of a subtype or as a sample within the complication population). The following three examples illustrate this characterization in three progressive steps.

(27) The prevalence of technical complications is on average 26%.

(28) Arterial complications, particularly the thrombosis of the hepatic artery (prevalence ranging from 1.5 to 25%) are the most frequent ones.

(29) Hepatic artery thrombosis is a complication that develops more frequently in the pediatric population.

Apart from overall values (27-28) for the population, samples typically address age, such as children (29), specific pre-surgical diseases triggering grafting (30), and retransplant patients (31):

(30) Portal vein thrombosis is an infrequent complication with an overall prevalence of 2-3%.

(31) Globally, 20-40% of liver transplant recipients present atraumatic bone fractures; this

prevalence rises to 65% in patients transplanted due to cholestatic disease and in retransplant patients.

Typical long-term spans include mention of one and five year periods.

(32) Currently though, survival rates of over 90-95% and 70% at one year and five years post-transplantation, respectively are expected.

Prevalence spans reach from 100% to values as close as 1.5%, as in (33) and (28).

(33) Pleural leakage, predominantly on the right, is the most frequent complication with a prevalence reported to be as high as 100% in some series.

Different diagnostic criteria are used to characterize a (chronic) complication. Variable prevalence is therefore a common measure:

(34) The prevalence is variable, depending on the criterion used to define it and to the method used to assess renal function. Indeed, serum creatinine measurement may underestimate the presence of renal failure.

Particularly in late complications, an important comment on the relation between time and treatment has to be made: a complication may display decreasing prevalence over time. As mentioned before, complications are sometimes caused by immunosuppressive drugs. Since drug administration may be decreased over time, this has an impact on its frequency. Conceptually, there is again a correlation (or even causal relation) between treatment method and time, independent of a specific complication, as the two following examples (concerning hypertension and diabetes, respectively) show.

(35) Arterial hypertension (AHT) is a frequent complication in liver transplant recipients. Its prevalence varies between 50-70% in the first post-transplantation months but decreases thereafter probably due to the reduction of the immunosuppressive doses.

(36) A variable percentage of patients, 4-20% according to the series, will develop diabetes mellitus following transplantation (de novo DM). The prevalence depends on the time elapsed since transplantation and particularly on the immunosuppressive drugs. In the initial post-transplantation period, DM is very frequent, probably due to the use of high CNI and steroid doses.

Typically, measures for prevalence and risk may compare patients with complications to the healthy population (37), taking into account the moment of grafting (38) and stage of drug development.

(37) The natural history of malignant tumors in the transplant patient tends to be different from that of the normal population; they appear at an earlier age, tend to be in a more advanced stage when diagnosed, and their evolution is more aggressive, causing high mortality directly related to the tumor.

(38) Some data suggest that in patients undergoing liver transplantation in recent years, there is a higher incidence of hematological neoplasms with de novo internal neoplasms developing at earlier time-points than in those transplanted years ago.

A relevant measure and objective of PLTM management is the reduction of frequency of complications to values for the general population, as in the case of bacterial infections:

(39) After the sixth month, with the transplanted organ functioning normally and minimum immunosuppressive doses, the frequency of bacterial infections is reduced to figures similar to those of the general population and the causes are pathogenic bacteria of the community.

Liver transplantation entails serious risks. Prevalence therefore is not only coined in terms of morbidity, but also of (decreasing) global mortality.

(40) The global mortality in this early posttransplantation period is approximately 5-10%.

Particularly in retransplantation, mortality rises. Retransplantation entails two subsequent time sequences: the management of the first graft, leading to an incurable complication (such as graft rejection mentioned before) within the first 48 hours following surgery, and a second one, with reduced survival prognosis.

(41) However, if regression of the clinical situation is not observed after 24-48 hours, retransplantation must be considered as soon as possible to avoid the development of multiorgan failure, in which case the mortality associated with retransplantation is very high.

4 Conclusion

This paper proposes a first sketch of time factors as an associative relation relevant for modelling the temporal unfolding of PLTM complications. We list the time factors useful for a model of time.

1. A boundary separates peri-operative care and management of complications. Prophylaxis is

a peri-operative issue in avoiding complications [3.1], and should be integrated in the time model.

2. Early complications are tested and detected against normal states.
3. These complications are temporally decomposed in immediate and late complications [3.2]; complications, however, do not belong unequivocally to one of these.
4. Relevant time notions are the cut-off point for occurrence and
5. the time span of (early) complications.
6. A distinction is needed between underlying pathologies not affected by transplantation [see section 3.1], and diseases that may affect complications and patient prognosis.
7. There is a variable correlation between the snap/span time (Munn and Smith, 2008) line of complication occurrence and associated medical diagnostic and therapeutic actions [see section 3.3].
8. Complications take different symptomatic forms causing different pathologies and therefore require various diagnostic tools.
9. Lead time to diagnosis is an important time dimension in this respect.
10. A distinction is needed for duration in graft rejection, between a (hyper-) acute and chronic disease status (without treatment options or not) [see section 3.4].
11. Complications involve some measures [3.5] such as prevalence of morbidity and mortality.
12. Prevalence may decrease based on the reduction of immunosuppressive drugs in the long run.

Future work will develop in three directions, The two first steps expand the corpus, taking into account, first the most quoted review articles on PLTM complications after 2006 and second, for specifying such time relations, in specialized research articles. The final goal is to transform this descriptive conceptual research into a logical entity-relation model, which can be useful in the clinical decision-making process. Describing such a model is clearly beyond this exploratory paper. Time in PLTM provides the conceptual basis for modelling subsequent associative relations: causal relations (what is the consequence of which (un)intended aspects of transplantation management, against which kinds of complications and/or sets of symptoms,?), as well as instrumental relations (what therapeutic tools, such as devices, drugs, surgical techniques, are used in order to prevent, block or reduce what kind of complications?).

References

Batts, K.P. 1999. Acute and chronic hepatic allograft rejection: pathology and classification. *Liver Transplantation and Surgery* 5(44 Suppl 1): S21-9.

Clavien, Pierre.-Alain, Carlos A. Camargo, Ruth Croxford, Bernard Langer, Gary A. Levy & Paul D. Greig. 2006. Definition and classification of negative outcomes in solid organ transplantation. Application in liver transplantation. *Annals of Surgery* 220(2): 109-120.

Dindo, Daniel, Nicolas Demartines & Pierre-Alain Clavien. 2004. Classification of Surgical Complications. A New Proposal with Evaluation in a Cohort of 6336 Patients and Results of a Survey. *Annals of Surgery* 240(2): 205-213

Grabar, N., Hamon, T., Bodenreider, O. 2012. Ontologies and terminologies: continuum or dichotomy? *Applied Ontology* 7(4): 375-386.

Greenberg, Raymond, S., Stephen R. Daniels, W. Dana Flanders, John William Eley & John R. Boring III (eds.). 2005. *Medical Epidemiology*. New York: McGraw-Hill.

Junttila, K., S. Salantera & M. Hupli. 2005. Developing terminology for documenting perioperative nursing interventions. *International Journal of Medical Informatics*. 74(6): 461-471.

Khoo, C. & N. Jin-Cheon. 2006. Semantic Relations in Information Science. *Annual Review of Information Science and Technology* 40: 157–228.

Moreno, R. & M. Berenguer. 2006. Post-liver transplantation medical complications. *Annals of Hepatology* 5(2): 77-85.

Munn, Katherine & Barry Smith. 2008. *Applied ontology. An Introduction.* Frankfurt and Paris: Ontos.

Pareja, E., M. Cortes, R. Navarro, F. Sanjuan, R. López & J. Mir. 2010. Vascular complications after orthotopic liver transplantation: hepatic artery thrombosis. *Transplantation Proceedings* 42: 2970-2972.

Roche, Christophe. 2007. Terme et concept: fondements pour une ontoterminologie. *Proceedings TOTh 2007 Terminologie & Ontologie: Théories et Applications*. Annecy, 1-22.

Sambre, Paul & Cornelia Wermuth. 2010. Causal framing for medical instrumentality: applied ontology and frame-based construction grammar. *Belgian Journal of Linguistics* 24: 163–191.

Tracing Research Paradigm Change Using Terminological Methods
A Pilot Study on "Machine Translation" in the ACL Anthology Reference Corpus

Anne-Kathrin Schumann[†] **and Behrang QasemiZadeh**

[†]Applied Linguistics, Translation and Interpreting
Universität des Saarlandes
Campus A2.2, 66123 Saarbrücken, Germany
anne.schumann@mx.uni-saarland.de
behrang.qasemizadeh@insight-centre.org

Abstract

This paper explores the use of terminology extraction methods for detecting paradigmatic changes in scientific articles. We use a statistical method for identifying salient nouns and adjectives that signal these paradigmatic changes. We then employ the extracted lexical units for discovering terms that are assumed to be central in characterising paradigm shifts. To assess the method's performance, in this pilot study, we work on "machine translation" (MT) research articles sampled from the ACL anthology reference corpus. We analyse this corpus to check whether the proposed approach can trace the dramatic changes that machine translation research has experienced in the last decades: from transformational rule-based methods to statistical machine learning-based techniques.

1 Introduction

Research in computational terminology traditionally focuses on static models of knowledge acquisition and representation. Corpus-based approaches have led to an increased interest in the automatic extraction and semantic categorisation of terms with many successful applications. However, progress in the empirical description and computational modelling of *terminological dynamics* has been rather slow.

This paper suggests that terminological methods and principles can be employed in empirical investigations of *diachronic knowledge evolution*. In particular, terminological methods can provide new insights into problems of diachrony since they can be used to trace (a) how terminologies come into being, and (b) how they develop over time as the scientific field itself evolves. Empirical work on the creation and development of terminologies is especially relevant for investigations into the history of science. Furthermore, studies of this kind are also likely to benefit terminology as a discipline, since they might provide insights into the driving forces of terminological development and knowledge organization.

The method proposed here identifies lexical units the importance of which increases or decreases upon the transition from an earlier period to a more recent one. In other words, we approach history of science in the form of a trend analysis task. Formally, this task consists of two sub-tasks, namely:

(a) the detection of those periods in time when a paradigm change is taking place (e.g., as signalled by terminological dynamics in a domain);

(b) the extraction of terms that are indicative of a declining or rising paradigm.

The pilot study described in this paper relates only to the extraction of terms signalling paradigm shift (i.e., sub-task (b)). The material for our analysis consists of research articles dealing with "machine translation". These articles are sampled from the ACL Anthology Reference Corpus (ACL ARC)—introduced in Bird et al. (2008).

Linguistically, the proposed method is inspired by studies on *register*.[1] Register linguistics approaches linguistic variation as the description of

[1]See Cabré (1998) for an elaboration of terminological aspects of register. Also, see Teich et al. (2015) for an applied perspective.

changing configurations of linguistic features on the textual level. One of the relevant dimensions for this type of study certainly is the lexicon. Accordingly, we hypothesise that paradigmatic changes in a field of knowledge are the cause of terminological dynamics. These dynamics are expressed in the form of the rise or decline of not just isolated terms but whole groups of terms.

We conclude that terms extracted by our method are salient if they are able to depict the paradigmatic change that the MT field has undergone in the last decades—that is, the advent of statistical methods in contrast to symbolic approaches that were in use earlier. The remainder of this paper is structured as follows. Section 2 briefly summarises relevant previous work. Section 3 outlines our extraction method. Section 4 reports the results of our pilot study, followed by an evaluation in Section 5. Section 6 discusses obtained results and concludes this paper.

2 Related Work

The term "paradigm" in the sense intended here goes back to Kuhn (1962). According to Kuhn, a paradigm emerges from a generally acknowledged scientific contribution to a research field. The significance of the paradigm consists in its ability to propose research problems and solutions to these problems to the relevant community. Some of Kuhn's arguments can be traced back to Fleck (1935). Fleck describes scientific communities as communities of thought ("Denkkollektive") who share habits in their way of perceiving and solving scientific problems ("Denkstil", literally "style of thought"). What is important here for our research question is that paradigms are coupled not only with specific types of problems and research methods, but also with terminologies: they constitute the inventory of lexical units used to refer to concepts that are central for a given paradigm. Consequently, they are subject to change whenever the conceptual outline of the discipline changes.

Terminological dynamics have been approached by terminology proper from various perspectives. Relevant to our study are the articles by Kristiansen (2011) and Picton (2011). Kristiansen (2011) provides a detailed account of external motivating factors of conceptual and, eventually, terminological dynamics. Picton (2011) elaborates a typology for the description of *short-term* term evolution patterns such as *neology–necrology* (i.e., appearance–disappearance of terms), term migration, and topic centrality–disappearance. Both papers, unfortunately, do not provide any methodology for the automatic detection of these dynamics.

In computational linguistics, trend analysis is usually approached by computing topic centrality and/or community influence measures and plotting them on a timeline. An example is the work by Hall et al. (2008) who try to trace the "development of research ideas over time". They employ the standard Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003)—a term-by-document model—for identifying "topic clusters". The method involves manual selection of relevant topics and seed words in multiple runs of the LDA algorithm. Probabilities derived from the LDA model are then used for the identification of rising and declining topics. Similar to our work, the authors report experiments over the ACL ARC, using publications from 1978–2006.

A term-based approach to topic and trend analysis is proposed by Mariani et al. (2014). The analysis is conducted on the *ELRA Anthology of LREC publications* starting in 1998. A term extraction method, namely TermoStat (Drouin, 2004), is employed to extract "topic keywords". For each year, terms and their variants are grouped into synsets and the most frequent terms are found. Finally, the authors study the rank development for the 50 most frequent terms in order to extract information on whether topics designated by these terms have risen, declined, or stayed stable over the period under analysis. Relevant co-occurrences of terms are also listed.

Gupta and Manning (2011) stress that for the purpose of detailed investigations into the history of science "... an understanding of more than just the 'topics' of discussion ..." is necessary. They extract semantic information for the categories FOCUS (i.e., the main contribution of an article), TECHNIQUE, and DOMAIN from the title and abstract sentences of research papers using a set of bootstrapped patterns. They then identify *communities* using the LDA algorithm. An influence measure is defined and calculated for communities based on the number of times their FOCUS, DOMAIN, or TECHNIQUE have been adopted by

other communities. Finally, results obtained from the ACL ARC are projected onto a timeline.

The work listed above has a number of shortcomings, amongst them are:

- Approaches based on topic modeling do not always provide readily interpretable topics. While many of the induced topics are convincing in terms of their lexical outline, we believe that the use of terminology, as proposed by Mariani et al. (2014), can provide more targeted information.
- For any detailed understanding of the history of a given discipline, it is insufficient to measure how "central" or "popular" certain topics were at different periods in time. Instead, the internal, fine-grained dynamics of the field such as paradigms and paradigm shifts need to be understood. To our knowledge, the work by Mariani et al. (2014) is the only one that includes a study of the lexical context of terminological units; however, this analysis is not carried out systematically. We believe that a systematic study of how groups of terms change over time can provide rich information for users that are interested in the history of a given scientific discipline (e.g., see Figure 2).

3 Detection of Lexical Rank Shifts: The Method

Our work differs from previous studies in that we exploit the notion of rank shifts for detecting fine-grained shifts rather than measuring topic centrality or popularity. The comparison of rank shifts between two lists of sorted lexical items is an established research method in the field of quantitative historical linguistics (e.g., c.f. Arapov and Cherc (1974)) and we believe that it can be adapted to our purposes.

In essence, our approach to the detection of terminological dynamics revealing a paradigm change is two-fold. *Firstly*, we extract lemmas that experience a change in their ranks upon the transition from older publications to more recent ones. We believe that these lemmas are either paradigmatic terms themselves or can be used to extract paradigmatic terms. We restrict word classes to nouns and adjectives since we believe that they are the most characteristic units for a given research paradigm. *Secondly*, we use extracted lemmas for identifying paradigmatic terms.

The first step (i.e., extraction of lemmas) consists of three sub-processes:

1. extraction of frequency per document information for all nouns and adjectives in the two sub-corpora under analysis and removal of strings containing non-alpha-numeric characters;
2. ranking of lexemes obtained for the two time periods using the method explained below;
3. comparison of the two ranked lists in order to identify those lexemes that have undergone relevant rank-shifts.

Frequency and document-related information is extracted using the IMS Open Corpus Workbench (CWB) loaded with our data (Evert and Hardie, 2011). For ranking, we employ the measure for calculating *domain consensus* proposed by Sclano and Velardi (2007). This measure—$DC_{D_i}(t)$—is defined as follows:

$$DC_{D_i}(t) = -\sum_{d_k \in D_i} nf(t, d_k) \log(nf(t, d_k)), \quad (1)$$

where d_k denotes the kth document in domain D_i, and nf is the normalised frequency of term t in $d_k \in D_i$. $DC_{D_i}(t)$ goes beyond the use of raw frequencies (e.g., as used by Mariani et al. (2014)). Instead, $DC_{D_i}(t)$ favors lexemes that are evenly distributed over all the texts in the two sub-corpora as opposed to candidates that are frequent just in a small number of texts. The process results in ranked lists of lexemes for the two time periods that we want to compare. Each lexeme either occurs in only one of the two lists or in both of them. To detect major rank shifts RS for a lexeme t that occurs in both lists, we use the following formula:

$$RS(t) = \frac{1}{R_{New}(t)} - \frac{1}{R_{Old}(t)}, \quad (2)$$

where $R(t)$ denotes the rank of t in the two ranked lists New (recent publications) and Old (early publications).

In the next step, the lemmas with highest rank shifts are employed to build partly lexicalised term extraction patterns for identifying paradigmatic terms. PoS sequence patterns are taken from the

Pattern	CWB query
adjective + noun	[pos="JJ.*"] [lemma="lexicon"]
past participle + noun	[pos="VVN"] [lemma="lexicon"]
noun + noun	[pos="N.*"] [lemma="lexicon"]
noun + noun + noun	[pos="N.*"] [pos="N.*"] [lemma="lexicon"]
noun + preposition + noun	[pos="N.*"] [pos="IN"] [lemma="lexicon"]
adjective + adjective + noun	[pos="JJ.*"] [pos="JJ.*"] [lemma="lexicon"]

Table 1: Examples of partly lexicalised term extraction patterns.

ONLY_NEW	ONLY_OLD
alignment	periphrasing
tag	cannonical
annotation	transcodage
database	transcoded
baseline	pidgin
ontology	sjstem
threshold	descri
monolingual	ption
multilingual	versinn
learning	periphrasin
architecture	paragrapher
engine	subroutine
n-gram	Noninclusive
decoder	inclusiveness
tagger	quelques

(a)

UP	DOWN
word	language
translation	sentence
corpus	structure
model	analysis
result	rule
text	form
method	problem
information	semantic
feature	grammar
system	computer
approach	program
set	theory
training	way
pair	possible
source	dictionary

(b)

Table 2: The result obtained from processing and comparing the *Old* and *New* sub-corpora. Note that dues to the presence of noise in pre-processes (e.g., OCR), the extracted lists of lexemes also contain invalid lexical units such as in Table 2a.

multilingual term extraction tool *TTC TermSuite* (Daille and Blancafort, 2013)[2]. Table 1 provides examples of these patterns.

4 Experiment

As stated earlier, we used the ACL ARC as a dataset. The corpus contains research articles on the topic of human language technology dating back as far as 1965. In our experiments, we use the preprocessed segmented version of the ACL ARC (i.e., the ACL RD-TEC) provided by QasemiZadeh and Handschuh (2014). Our pilot study is limited to the research publications in the domain of MT. Given our knowledge that MT research has undergone a major paradigm shift since the late 1980s, we want to examine whether our method is able to capture and characterise this paradigm shift.

[2] http://code.google.com/p/ttc-project/

Up terms	Down terms
machine translation	natural language
language model	deep structure
translation system	phrase structure
word sense	transformational rule
training datum	syntactic analysis
test set	surface structure
mt system	sentence structure
translation model	physics problem
sentence pair	semantic theory
statistical machine translation	transformational grammar
machine translation system	phrase structure grammar
bleu score	average number
parallel corpus	linguistic theory
training set	conversion rule
english word	source language

Table 3: Most frequent paradigmatic term candidates extracted using the proposed lexicalised PoS sequence patterns. We consider *Up terms* and *Down terms* as indicators of topics that are *trending* and *un-trending*, respectively.

To prepare the data for experiments, we extract nouns and adjectives from papers containing either the string "machine translation" or "automatic translation". We divide the corpus into two sets of articles: Old (1960s–70s) and New (1980s onwards). Since New is substantially larger than Old, we randomly reduce the size of the New set in order to make it more comparable to Old. Despite this effort, the two sub-corpora still have a different size and structure—New contains 290,337 nouns and adjectives whereas Old contains only 79,247.

The extracted lemmas are weighted using Equations 1 and 2. Consequently, four sets of words are generated:

- words that occur only in *New* (ONLY_NEW);
- words that occur only in *Old* (ONLY_OLD);
- words whose rank increases upon the transition from *Old* to *New* (UP);
- words whose rank decreases upon the transition from *Old* to *New* (DOWN).

The first set—items that occur only in *New*—is comparatively large and contains 14,347 adjectives and nouns. *Old*, on the other hand, has

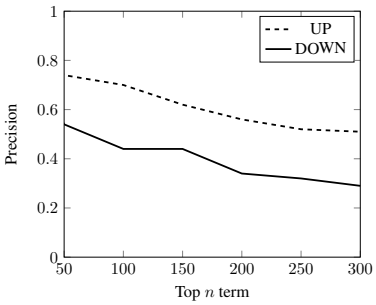


Figure 1: Precision at n for the extracted list of terms using the lexicalised patterns for the Up and Down lemmas.

7,094 unique adjectives and nouns. 1,023 lemmas have an increased rank over time, and 2,880 words are subject to rank decrease. Table 2 details the results by showing the top 15 items in each set of generated words. Table 2a shows words that occur only in *New* or only in *Old*. Table 2b, however, shows common words with the largest rank shifts. Note that ONLY_NEW and ONLY_OLD have been ranked by their assigned DC score (Equation 1), whereas Up and Down are sorted according to the score computed using Equation 2.

In the *second step*, we select the top 30 plausible noun lemmas from the UP list (shown in Table 2b) and use them for building term extraction patterns (as exemplified in Table 1). This process is also repeated for the top 30 nouns from the DOWN list. The two obtained sets of patterns are employed to extract terms from the *New* and the *Old* sub-corpora, respectively. Table 3 provides an overview over the 15 most frequent candidate terms extracted by this method. Figure 1 reports the precision for the first 300 Up and Down paradigmatic term candidates obtained by automatically comparing them to terms annotated in the ACL RD-TEC by QasemiZadeh and Handschuh (2014).

5 Evaluation

The 15 lemmas listed in Table 2b (i.e., $DC_{D_i}(t)$-ranked lemmas) are presented to 5 researchers in the area of machine translation. The evaluators are asked whether

(a) the individual lemmas in Table 2b are salient for the period they are supposed to represent (*New* and *Old*); and,

Up	**Down**
training	transformational
corpus	routine
score	force
probability	picture
target	location
pair	numeral
evaluation	title
task	reverse
statistical	geometric
source	physics
performance	decimal
bilingual	personal
feature	intension
error	Russian
sense	storage

Table 4: The baseline lemma list: top 15 lemmas sorted by frequency and rank shifts.

(b) the lists as a whole contain words that are typical for the mainstream research paradigms in the respective periods.

To investigate (a), participants make binary distinctions (i.e., in each of the Up and Down lists, a lemma is marked either as relevant or irrelevant). To investigate (b), participants are asked to provide a grade indicating the relevance of the lists of terms on a scale from 1 (“list is irrelevant”) to 5 (“relevant”).

In order to assess whether the $DC_{D_i}(t)$ ranking mechanism proposed in this paper (i.e., Equations 1 and 2) outperforms simpler ranking methods, we also construct a baseline data-set: nouns and adjectives in New and Old are sorted by their frequency and then evaluated by the differences in their ranks. The resulting baseline data is given in Table 4. Evaluators are asked to repeat the abovementioned assessment also for this baseline without being aware of how both data-sets were produced. Table 5 summarizss the results of this evaluation.

Each row of the Sub-Tables 5 summarises the input from each of the expert evaluators. The first and the second column in each sub-table show the sum of positively marked Up and Down items—that is, the sum of those lemmas (out of 15) that were found salient for either the 1960s–1970s or the 1980s–2000s (sub-task (a)). The third column presents the overall evaluation of the lists (i.e., sub-task (b)). Table 5a provides the results for the list of lexical items that are ranked using the $DC_{D_i}(t)$ score (i.e., listed in Table 2b). Table 5b provides the assessments for the

Up	Down	Overall
12	10	4:5
12	10	4:5
13	12	4:5
10	10	3:5
3	4	2:5

(a)

UP	DOWN	Overall
15	5	3:5
11	2	3:5
14	11	3:5
11	6	4:5
6	2	3:5

(b)

Table 5: Each row of these tables summarises the assessment of each of the evaluators. Table 5a shows the results for the sets of lexical items ranked by $DC_{D_i}(t)$ (listed in Tables 2b). Table 5b, in contrast, provides the result for the sets of lexical items that are sorted by their raw frequencies (listed in Tables 4).

baseline list (i.e., listed in Table 4).

As can be observed in Table 5, the evaluators tend to prefer the $DC_{D_i}(t)$-ranked lexical items over the baseline data-set. Except for one of the annotators who suggests that the baseline method provides more informative output (i.e., the last row of Tables 5a and 5b), the evaluators consistently prefer the ranking mechanism proposed in this paper, assigning an overall grade of 3–4 (out of 5) points to the output. However, the difference remains but slight.

Table 6 shows the 15 most frequent terms in the *Old* and the *New* corpus, respectively. These terms were collected using the manual annotations in the ACL RD-TEC by QasemiZadeh and Handschuh (2014). By comparing these terms to the output of our method (Table 3), we observe considerable differences. Evidently, for the detection of paradigm shifts, terms extracted using semi-lexicalised part-of-speech (PoS) patterns based on our $DC_{D_i}(t)$ method are better indicators of the paradigm shift than terms ranked by their raw frequencies.

Figure 2 exemplifies some of the dynamics detected by our method. For each year, the plot shows the frequencies of terms normalised by the sum of all term frequencies extracted from the publications in that year. All plotted terms were among the top items in our Up and Down lists. Up paradigmatic terms are given in blue whereas Down paradigmatic terms are plotted in black.

Figure 2 illustrates what types of information can be drawn from the analysis conducted here. For example, we observe that "automatic evaluation" rises synchronously with "Bleu score"

Sub-Corpus Old	Sub-Corpus New
natural language	machine translation
machine translation	natural language
computational linguistics	language processing
data base	translation system
artificial intelligence	target language
language processing	computational linguistics
phrase structure	natural language processing
syntactic analysis	training data
translation system	source language
automatic translation	test set
natural languages	information retrieval
information retrieval	machine translation system
noun phrase	language model
language understanding	training corpus
noun phrases	noun phrase

Table 6: 15 most frequent terms (two tokens or longer) in the Old and the New sub-corpora. This list was collected using the manual annotations in the ACL RD-TEC and from the documents in the two Old and New sub-corpora.

and is only slightly preceded by "statistical machine translation" itself. We also find that, during the 1980s, references to "linguistic'theory" were rather frequent, but they have largely vanished since 1990. Themes such as generative grammar or phrase structure grammar were not dominant even in the earlier decades, but they exhibit a constant decline at least since the 1990s. Evidently, the plot confirms that our attribution of terms to the categories Up and Down is justified. Moreover, this plot supports our hypothesis that paradigm shifts are lexically expressed by dynamics of whole groups of related terms.

6 Discussion and future work

For a detailed understanding of the dynamics of science, it is insufficient to measure how "central" or "popular" certain topics are at different periods of time. Instead, those groups of terms that signal paradigm changes must be detected—this is the key idea that motivates the research presented in this paper. The pilot study described here, therefore, aims at showing that terminological methods can be employed to serve this purpose, and to provide information for understanding what is going on in a scientific field at a given moment in time.

An inspection of our method's output indicates that the renewal of vocabulary (happening by some words falling from use and others being in-

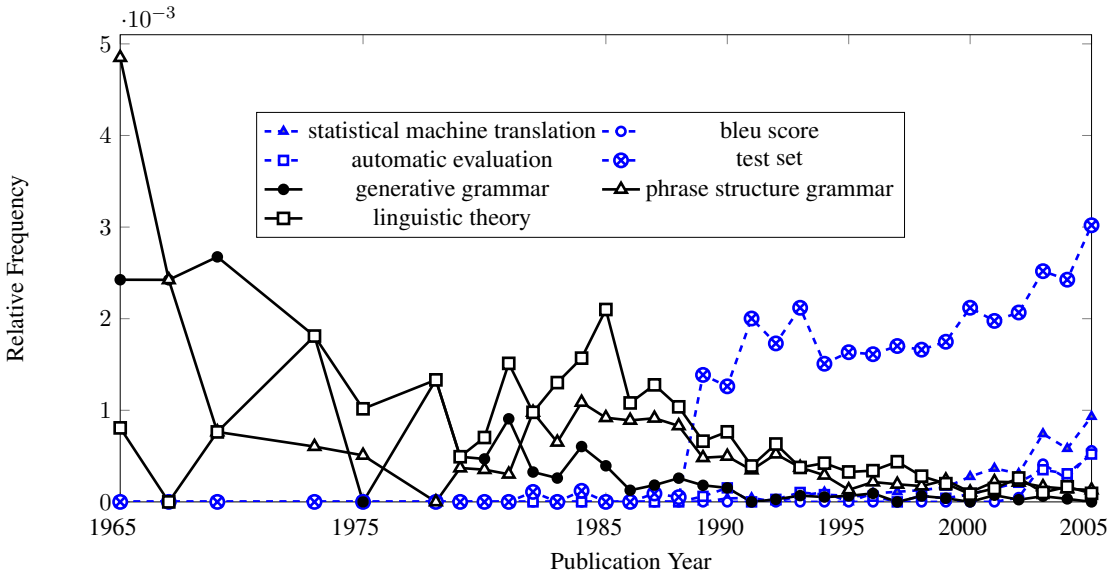


Figure 2: Terms mapped onto a timeline: For each year, the y-axis shows the frequencies of terms normalised by the sum of the frequencies of all the terms extracted in that year.

troduced) is considerable given the relatively short time span under analysis in our experiments. We observe that the content words shared by the two data sets are, in fact, a minority. However, we also observe that *Only_New* (Table 2a) clearly contains items that are indicative of more recent MT research such as "alignment", "n-gram" or "decoder". The items that are specific to *Only_Old*, on the other hand, seem to be rather spurious and low-frequent. These lexical units, rather unsurprisingly, disappear upon the transition from *Old* to *New*.

Our evaluation also indicates that the lemmas extracted by our method (Table 2b) are indicative of the respective time periods, at least as far as the top ranks are concerned. MT experts prefer the output of our proposed method over the output of the baseline method, perhaps due to the improved coverage of the relevant Down lemmas.

Moreover, the terminological evaluation of the extracted paradigmatic terms (Figure 1) shows that Up lemmas indeed help to extract valid computational linguistics terms. Performance for Down lemmas, however, is consistently worse. This difference in performance, in our opinion, is related to the higher productivity of the Up lemmas from Table 2b: Up lemmas are used in a growing number of more specific and more frequent terms, whereas Down lemmas do not experience a similar increase in frequency and specificity. That is, it is harder to distinguish irrelevant collocations containing Down terms from collocations with terminological value. Hence, term extraction performance for Down terms is worse. We believe that, if this property can be shown to hold in general, it is highly relevant as it can be used for the extraction of emergent and semantically related terms. Term extraction performance itself can be further improved by integrating standard practices such as stop-word filtering.

Last not but not least, a timeline plot of Up and Down paradigmatic terms indicates that Down terms, as expected, do not exhibit the same exponential growth as Up paradigmatic terms. However, what we also observe is that many relevant terms do not simply fall from use (e.g., the term "linguistic theory"). They may even increase their absolute frequency or become salient again in new or unforeseen contexts.

The local context of terms therefore remains an unexplored factor in trend analysis research. If we look more closely into our data, we find unexpected formulations such as "the language model in the human" or "translation model based on semantic interpretation". Future work will need to address these kinds of dynamics in superficially identical terms that are even more fine-grained than the rank shifts observed in this pilot study.

Several measures can be taken into consideration for improving our current evaluation method. Future work will also strive for a comparison of multiple sub-corpora that represent time slices of different granularity, perhaps of more similar size and structure. The detection of time periods in which paradigm shifts occur and a more precise modelling of their interplay with terminological dynamics are also important topics for future research.

Finally, we would like to mention that an important observation about the dependence of lexical dynamics on frequency has already been made by Arapov and Cherc (1974) who explicitly refer to Zipf:

> The speed of decay ... can, in a way, be understood as the probability of decay. The higher the ordinal number (rank) of a [word] group ..., the lower the frequency of the words belonging to that group, the higher is the speed of decay of this group.[3]

It is no surprise that term frequency does play a role in term necrology. However, the formula that we currently use for rank comparison (i.e., Equation 2) does not account for this aspect. Furthermore, the question how to compare terms the frequencies of which differ by sizes of magnitude is also yet unresolved. Future work will address these shortcomings.

Acknowledgements

We thank Mihael Arcan, Iacer Calixto, Peyman Passban, Liling Tan and colleagues for evaluating our data. We would also like to thank Prof. Elke Teich for her comments and advice. This research has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Cluster of Excellence 'Multimodal Computing and Interaction'.

References

M. V. Arapov and M. M. Cherc. 1974. *Matematičeskie metody v istoričeskoj lingvistike*. Nauka.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of LREC'08*, Marrakech, Morocco, may. ELRA.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (3).

Maria Teresa Cabré. 1998. Do we need an autonomous theory of terms? *Terminology*, 2(5).

Béatrice Daille and Helena Blancafort. 2013. Knowledge-poor and Knowledge-rich Approaches for Multilingual Terminology Extraction. In *Cicling*.

Patrick Drouin. 2004. Detection of Domain Specific Terminology Using Corpora Comparison. In *LREC*.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics*.

Ludwik Fleck. 1935. *Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv*. Schwabe.

Sonal Gupta and Christopher D. Manning. 2011. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *IJCNLP*.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the History of Ideas Using Topic Models. In *EMNLP*.

Marita Kristiansen. 2011. Domain dynamics in scholarly areas: How external pressure may cause concept and term changes. *Terminology*, 17(1).

Thomas S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago.

Joseph Mariani, Patrick Paroubek, Gil Francopoulo, and Olivier Hamon. 2014. Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis. In *LREC*.

Aurelie Picton. 2011. Picturing short-period diachronic phenomena in specialised corpora: A textual terminology description of the dynamics of knowledge in space technologies. *Terminology*, 17(1).

Behrang QasemiZadeh and Siegfried Handschuh. 2014. The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics. In *Computerm*.

Francesco Sclano and Paola Velardi. 2007. Termextractor: A Web Application to Learn the Shared Terminology of Emergent Web Communities. In *Enterprise Interoperability II: New Challenges and Approaches*. Springer.

Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2015. The Linguistic Construal of Disciplinarity: A Data-Mining Approach Using Register Features. *J. Assoc. Inf. Sci. Technol.*

[3] Translated from Russian.

Evaluating noise reduction strategies for terminology extraction

Johannes Schäfer[1], **Ina Rösiger**[1], **Ulrich Heid**[2], **Michael Dorna**[3]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]University of Hildesheim, Germany, [3]Robert Bosch GmbH, Germany
{schaefjs|roesigia}@ims.uni-stuttgart.de
heid@uni-hildesheim.de, michael.dorna@de.bosch.com

Abstract

We present work on the task of reducing noise in nominal terminology extraction. Based on a comparative evaluation of statistical measures aimed at capturing domain specificity, we propose strategies to increase the typically quite low accuracy of classical hybrid nominal multi-word term extraction. Our experiments on a set of German do-it-yourself instruction texts show that using linguistic filters that determine the right span of the MWE before applying a suitable combination of statistical measures improves results.

1 Introduction

The automatic extraction of terminology from domain-specific text is a task that has gained interest in the research community over the last twenty years. It is an important prerequisite for applications such as ontology creation or knowledge extraction from texts.

The work presented here is part of a project that deals with knowledge extraction and ontology creation from German texts from the do-it-yourself domain. As a first step, we aim at high quality terminology extraction of nominal candidates, as these describe the objects of the domain, followed by the extraction of verbal items and verb+complement patterns before we bring them together to build up partial ontologies of the domain. This paper describes strategies to reduce noise in nominal term extraction.

We consider single-word terms and multi-word terms, but focus on the latter because the extraction of multi-word terms (MWTs) is more difficult. As they are of variable length it is in many cases nontrivial to ensure the correct span of the term. We aim at extracting noun phrases (NPs) of different levels of complexity, such as adjective and noun combinations as well as NPs containing a genitive or prepositional modifier.

Nominal terms may contain embedded prepositional phrases (PPs), such as in example (1).

(1) *Bohrer mit Diamantspitze*[1]
(drill with diamond bit)

However, we do not want to extract PPs that are not syntactically attached to a term, e.g. because they are verb-dependent, such as in example (2).

(2) *die *Oberfläche mit Leinölfirnis bedecken*
*(cover the *surface with linseed oil varnish)*

Thus, one of the noise reduction steps is to ensure that the extracted nominal candidates are syntactically valid, and do not cover too long spans.

There are also cases where the term extraction may return too short candidates. Sometimes, the extracted terms only occur as part of bigger terms, and are not valid on their own, such as in example (3).

(3) *elektromagnetisch *angetriebene Spritzpistole*
*(electromagnetically *operated spray gun)*

There are both statistical and hybrid approaches to term extraction (Cabre and Vivaldi Palatresi, 2013). Association measures (cf. e.g. Evert (2005)) are designed to extract collocations ("unithood", cf. Kageura and Umino (1996)) and have been used for terminology extraction, e.g. by Couturier et al. (2006). However, Roche et al.

[1]Extracted term candidates are underlined. The * here denotes wrongly extracted MWT candidates.

(2004) investigated the use of association measures for this task and came to the conclusion that these standard measures are outperformed by more sophisticated approaches. They do not focus on domain-specificity ("termhood") and thus do not perform better at terminology extraction than mere frequency based approaches (Pazienza et al. (2005), confirmed by our own experiments, where the maximal F_1-score obtained with association measures is 0.45).

Termhood is addressed through statistical measures that only use the candidate's frequency in a domain-specific corpus (e.g. Frantzi et al. (2000)), as well as by measures based on a comparison of a candidate's frequency in a domain and in general-language corpora (cf. Ahmad et al. (1992) and section 2.3). However, among others due to data sparseness in small size specialized corpora, both approaches perform much better on single-word terms (SWTs) than on MWTs.

Most hybrid systems (linguistic pattern-based search for candidates plus statistical ranking) often do not address variable length and syntactic validity satisfactorily (with a few noteworthy exceptions, cf. for example Chen et al. (2008)): part-of-speech (POS) sequence patterns are typically flat and cannot identify phrase boundaries and grammatical functions (cf. example (2)). As the POS patterns do not model phrase structure, they may cut off essential parts from a multi-word, returning unattested candidates (cf. example (3)).

We address the above issues by means of a three-step approach which modifies and extends the classical hybrid scenario: (i) nominal candidates are selected via part-of-speech patterns; (ii) they are filtered wrt syntactic validity and embedding and finally (iii) ranked according to statistical measures that involve a comparison between specialized and general-language corpus. The system with which we experiment extracts lemma combinations, morphosyntactic properties and text-specific metadata. Our method is evaluated on a 2.7 million word corpus of German do-it-yourself (DIY) instruction texts against a gold standard.

The main contributions of this paper are a study on the suitability of standard statistical measures for the extraction of nominal single- and multi-word terms, as a basis for further adaptations and methods to improve the noise-silence ratio, based on experiments with linguistic filters (we use parsing information to ensure the MWE is a valid nominal phrase (NP)), and experiments on the combination of statistical measures. We believe that the methods we propose are generalizable to other domains of specialization and to other languages. More experiments will however be needed to confirm this.

2 Improving term extraction quality

In the present work, we only deal with nominal candidates. To maximize recall on (comparatively) small specialized corpora, we use POS patterns that account for basic terms[2] (N, Adj N, N P N, N D $N_{genitive}$) and for their potential variants (step (i) in the summary above). The set of patterns is described by the regular expressions given below.

– (Adv? Adj? Adj)? N
– (N D)? (Adv? Adj)? N P D? (Adv? Adj)? N
– (Adv? Adj)? N D (Adv? Adj)? $N_{genitive}$

These patterns are flat and thus do not adequately represent syntactic structure. In particular, they cannot distinguish between cases (a) where NP and PP are sister nodes vs. (b) where the PP is embedded in the NP (cf. examples (1) and (2) above). Thus, step (ii) is added to remove noise: we exclude items from our candidate set which are syntactically invalid (too long ones) by checking phrase boundaries and we use the C-value score (Frantzi et al., 2000) to remove too short items, i.e. those occurring only embedded in other candidates. In step (iii) we combine statistical measures to rank the selected candidates by domain specificity.

2.1 Ensuring syntactic validity

Candidates covering too long spans typically occur when part of the extracted MWT is actually attached to the verbal phrase, e.g. in example (4) and example (5).

(4) *die *Schablone mit Farbe besprühen*
*(spray the *template with paint)*

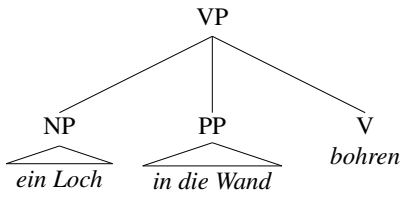
(5) *ein *Loch in die Wand bohren*
*(drill a *hole into the wall)*

[2]POS tags: N-noun, Adj-adjective, P-preposition, Adv-adverb, D-determiner.

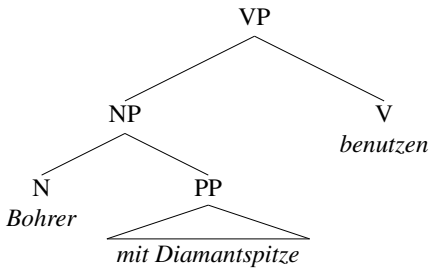
We filter these by using the dependency parser *mate* (Bohnet, 2010) to find start and end points of NPs.[3] This parser was chosen because, in the long run, we aim at relation extraction based on syntactic functions. Moreover, *mate* has been shown to produce the highest accuracy in a recent evaluation of the currently available dependency parsers (Choi et al., 2015). We are aware that *mate* has not been optimized to solve the PP attachment problem and that there is no evaluation on specialized text available yet (cf., however, Zollmann et al. (2016) on a partial evaluation).

The boundary violation filter works as follows: If an instance of a MWT candidate comprises two sister phrases, i.e. if the POS sequence identified goes beyond the end point of an NP, it is not counted as a valid occurrence of the respective lemma sequence. As an example for a violation, consider the following MWT candidate where '*ein Loch*' and '*in die Wand*' are sister phrases ('*ein Loch in die Wand bohren*' (drill a hole into the wall), (5)), whereas in the example (6) ('*Bohrer mit Diamandspitze benutzen*' (use a drill with diamond bit)) there is no violation.

(5)



(6)



The candidate sequence is not removed from the list of possible candidate terms, as other occurrences might not have been analyzed as violating syntactic boundaries. The filter is thus a "soft" one as it only affects the frequency of the lexeme combination candidate. We also experiment using a "hard" filter, where the lexeme combination candidate is removed altogether.

[3] In the current experiments only for subject and object phrases.

2.2 Filtering out invalid embedded phrases

An example to show the necessity for an accurate treatment of nested terms was found in our extraction result: '*zugängliche Stelle*' (accessible place) and '*schlecht zugängliche Stelle*' (poorly accessible place). It should be obvious that from occurrences of the latter term we do not want to extract the former.

Thus, with nested MWEs, not all fragments of a longer expression might be suitable candidate terms. C-value (Frantzi and Ananiadou, 1996) identifies embedded sequences as valid units under the following condiditons: (a) the embedded sequence also occurs on its own; (b) the embedded sequence occurs in lexically diverse longer sequences. The C-value for a candidate term a is defined as in formula 1.

$$\mathrm{C}(a) = \begin{cases} \log_2 |a| * f(a) & \text{if } a \text{ not nested} \\ \log_2 |a| * f(a) - \frac{\sum_{b \in T_a} f(b)}{P(T_a)} & \text{otherwise} \end{cases} \quad (1)$$

$|a|$ = term length of a (number of words)
$f(.)$ = frequency in the domain corpus
T_a = set of longer candidate terms that contain a
$P(T_a)$ = number of these longer candidate terms

Furthermore, C-value reflects the idea that longer sequences have a tendency to be more (domain-) specific than shorter ones.

Based on this, we consider German noun compounds as pseudo-MWEs and compute the term length $|a|$ from formula 1 by using the result of compound splitting as produced by the compound splitting tool COMPOST (Cap, 2014) (cf. formula 2).

$$\text{termlength}(a) = \sum_{w \in a} (1 + \log(\text{cslen}(w))) \quad (2)$$

w = a word
$\text{cslen}(w)$ = number of compound elements in w

Frantzi and Ananiadou (1996) propose C-value as a termhood measure using only the frequencies in the domain corpus; thus, most general-language noise cannot be filtered out. We therefore suggest to use C-value as a corrected frequency and to combine it with further statistical measures.

2.3 Ranking by domain-specificity

In section 4, we will compare the following statistical measures designed to rank candidate terms by domain-specificity. A detailed description of these measures is given in Schäfer (2015). As these measures place general-language candidates

at the bottom of the list, a selection of top candidates shows a reduced amount of noise. The measures are defined as follows using the domain frequencies f, the general-language frequencies F as well as the sizes of the corpora: s for the domain corpus and S for the general-language corpus.

- Weirdness ratio for domain specificity (DS) (Ahmad et al., 1999): Identifies domain-specific terms by the ratio of the relative frequencies in the domain and in general language as in formula 3.

$$\text{Weirdness} = \frac{f/_s}{F/_S} \tag{3}$$

- Corpora-comparing log-likelihood (LL) (Rayson and Garside, 2000): Identifies units with significant frequency differences between the two corpora by formula 4[4]. Note that this version of LL differs from the standard log-likelihood collocation measure.

$$LL = 2\left(f * \log\left(\frac{f}{E_f}\right) + F * \log\left(\frac{F}{E_F}\right)\right) \tag{4}$$

With $E_f = \frac{s*(f+F)}{s+S}$ and $E_F = \frac{S*(F+f)}{S+s}$.

- Contrastive Selection via Heads (CSvH) (Basili et al., 2001): Computes the domain-specificity of a multi-word candidate (ct) using a contrastive filter based on the general-language frequency of its head ($h(ct)$) by formula 5.

$$cw_{ct} = \log(f_{h(ct)}) * \log(\frac{S}{F_{h(ct)}}) * f_{ct} \tag{5}$$

- Term Frequency Inverse Term Frequency (TFITF) (Bonin et al., 2010): Combines the term frequency in the domain corpus with the inverse term frequency in the general-language corpus as in formula 6.

$$w_t = \log(f(t)) * \log\frac{S}{F(t)} \tag{6}$$

[4]As the LL formula is obviously symmetric in the two corpora, we multiplied the result for candidates by -1 if their relative frequency in the domain is smaller than the one in general language, in order to place candidates with a significantly high general-language frequency at the bottom of the list.

- Contrastive Selection of multi-word terms (CSmw) (Bonin et al., 2010): Applies a contrastive filter using the general-language frequency including an $\arctan$ scaling to reduce variation in low-frequency candidates as in formula 7.

$$\text{CSmw}(t) = \arctan(\log(f(t)) * \frac{f(t)}{F(t)/S}) \tag{7}$$

3 Evaluation setup

Tool. We used the TTC[5] (Terminology extraction, translation tools and comparable corpora (2010-2012)) term extraction research prototype (Gojun et al., 2012), a standard hybrid tool that combines linguistic preprocessing with statistical measures, which has recently proven to outperform SDL MultiTerm[6], a purely statistical commercial state-of-the-art tool (George, 2014).

The pipeline involves the following components:

Figure 1: Term extraction pipeline

- Preprocessing:
  - Tokenization: sentence and word form delimitation and markup;
  - Word class tagging and preliminary lemmatization: annotation by means of the RFTagger (Schmid and Laws, 2008), including an annotation as "unknown" of word forms absent from the tagger lexicon;
  - Lemmatization: specific treatment of the word forms absent from the tagger lexicon, with a view to guessing their lemma and part of speech, by use of word form similarity, inflection-based rules and compound splitting.
- Pattern-based term candidate extraction: use of simple as well as extended POS-based patterns to identify term candidates;for the patterns used see section 2.

[5]TTC-project: `http://www.ttc-project.eu/`

[6]`http://www.sdl.com/de/cxc/language/terminology-management/multiterm/`

# tokens	text
62,131	do-it-yourself handbook
6,868	encyclopedia entries
5,150	list of FAQs with answers
15,104	tips and tricks for do-it-yourselfers
35,302	marketing texts
2,160,008	user generated project descriptions
444,381	user generated wiki content
2,728,944	total DIY corpus

Table 1: Number of tokens in the domain corpus

- Ranking:
  sorting of the candidate lists produced by the preceding step, according to different measures (cf. section 2.3).

Domain corpus. We use a corpus of expert and user-generated German texts from the DIY-domain, consisting a.o. of manuals, practical tips, marketing texts and project descriptions (cf. table 1). This corpus is highly heterogeneous since the domain texts were acquired by various methods resulting in fundamentally different types of texts. The texts also differ with regard to the level of expertise of the author and the intended reader. Some texts are written by a domain expert as instructions for users and some are user-generated context.

As the texts differ wrt authorship and text style, several statistical measures are implemented in order to identify different properties of terms providing multiple lists of term candidates. A domain expert can then select the most relevant lists for the construction of a terminological representation of the domain language. For the experiments presented in this work we treated the corpus as a single unit. A source identifier was included as meta data annotation. In future work on a larger version of the corpus, subsets by text type and author/intended reader may be analyzed separately.

General-language corpus. We use the SdeWaC corpus (Faaß and Eckart, 2013) as a general-language corpus. It consists of 880 million tokens. This corpus was chosen since its sentences are a broad collection of German web texts supposed to provide a statistically representative distribution of words in general language. [7]

[7] An alternative source would be Wikipedia, as it covers a broad variety of specialized topics.

POS pattern	number	example
N	4,238	*Kreissäge (circular saw)*
Adj N	604	*thermische Zersetzung (thermal decomposition)*
N P N	148	*Bohren von Dübellöchern (drilling of dowel holes)*
N D N_{gen}	107	*Viskosität der Farbe (viscosity of the paint)*

Table 2: Terms in the gold standard

Gold standard. A gold standard (GS) has been developed for the basic POS patterns (cf. section 2) which we take to capture the core terminology of the domain. Lemma sequences with a minimum frequency of four were extracted from the domain corpus matching these patterns. The gold standard contains those terms which were marked as terms by at least two out of three independent annotators carefully following defined guidelines (George, 2014). This process of creating a gold standard is based on the concept of monolingual reference lists (cf. Loginova et al. (2012)). Our gold standard contains 4,238 SWTs (nouns including compounds) and 826 MWTs (cf. table 2). This distribution is due to the fact that we derived the gold standard from the available text data and not e.g. from a test suite. Moreover, the frequency cut-off of four removed many MWT from being considered for the gold standard. As German compounds "count" as SWT, the MWT number is comparatively low. Our statistical methods are however also applied to compounds (cf. section 2.3). The inter-annotator agreement[8] ranges between moderate and substantial agreement, depending on the pattern. For multi-words, the kappa is 0.59 (moderate agreement) which is satisfactory considering the imbalanced distribution of categories.

4 Evaluation of noise reduction steps

4.1 Comparative evaluation of statistical measures

We compare the suitability of the measures mentioned in section 2.3, as a basis for further adjustments. The measures DS and CSmw seem to be the most suited overall (cf. figure 2), while TFITF

[8] We compute Fleiss' kappa (Fleiss, 1971). Interpretation according to (Landis and Koch, 1977).

term	freq	f-rank	DS-rank	TFITF-rank	CSmw-rank
Drehmomentvorwahl (torque pre-selection)	33	2,344	65	314	67
Bohrer (drill)	1,094	44	158,094	40	2,554
Mutter (screw nut/mother)	510	133	216,341	2,276	38,036

Table 3: Ranked candidate term examples

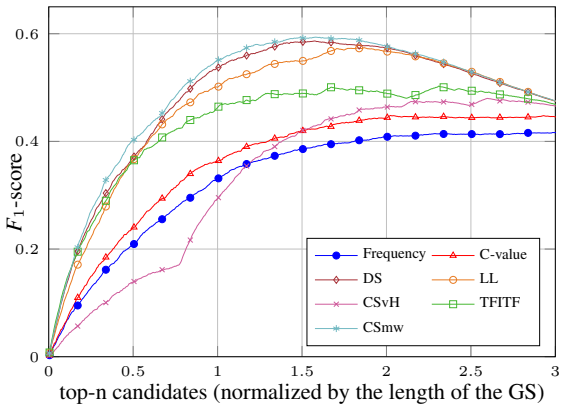


Figure 2: Statistical measures for term extraction

produces superior results only for very short candidate term lists. CSmw achieves a maximum F_1-score of 0.59 (48% precision and 77% recall) which is an improvement of 20 % over the simple frequency baseline. For very short lists TFITF achieves a precision of above 80%, for example 84% in the top 150 extracted candidates where other measures barely reach 70%. In the following we analyze these observations in detail and illustrate the reasons with examples from the extraction result as presented in table 3. This table shows in columns from left to right: extracted terms, their frequency in the domain corpus and their rank in the result lists according to the measures frequency, DS, TFITF and CSmw.

Sorting the candidates by their **DS**-value shows a high density of highly domain-specific terms at the top. For example, '*Drehmomentvorwahl*' (torque pre-selection) on rank 65 (out of 226,715 candidates). DS seems to strongly focus on very specialized terms of the domain texts which do not occur in general language, or only with very low frequency. However, as a consequence it misses important domain terms which also occur with a moderate frequency in general language. For example the term '*Bohrer*' (drill) which is essential for the domain (domain corpus frequency: 1,094) is only on DS-rank 158,094. This shows that the DS-value approach is not suitable to provide a list of important terms in the domain, but rather to identify its very specialized terms.

The **TFITF**-measure determines termhood by including the domain corpus frequency of a candidate term logarithmically (cf. formula in section 2.3) - unlike the DS-measure which uses it relatively. As a result several top candidate terms of the TFITF list are of a different kind than the best DS terms. For example, the above-mentioned term '*Bohrer*' (drill) is now at rank 40. TFITF even puts the candidate '*Mutter*' (screw nut/mother) which, due to its homography, is hard for a terminology extractor to identify as a term, at rank 2,276; this is acceptable, considering that there are 5,097 terms listed in the gold standard. The measure puts a stronger emphasis on the domain corpus frequency producing a considerable amount of noise in form of general-language candidates, which explains its rather mediocre overall performance. We thus suggest to use TFITF to extract a relatively small set of terms with a very high precision, for example for bootstrapping approaches or for an ontology learning which not only focuses on special technical terms of the domain but rather on its key topics.

The results acquired by the measure **CSmw** are in between TFITF and DS. It also gives more emphasis to the domain corpus frequency of a candidate term than DS, however not as much as TFITF. In the statistical analysis this approach reached the best overall F-scores. The CSmw-measure identifies the same top terms as the DS-measure, namely those which are highly domain specific and rare in general language, for example: '*Drehmomentvorwahl*' (torque pre-selection) on rank 67. Furthermore, it ranks comparatively high the essential objects of the domain which are also used in general language, for example: '*Bohrer*' (drill) is on rank 2,554. Consequently, the measure also produces some noise (as TFITF does) which is why it does not outperform the DS-measure. The ranking by CSmw turns out to be the most recommendable for a general terminology extraction which focuses on

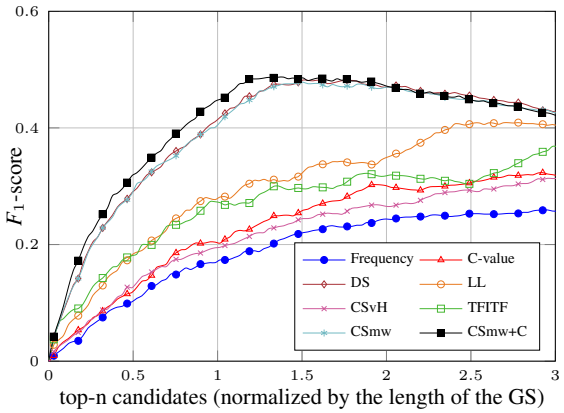


Figure 3: Statistical measures for MWT extraction

technical terms as well as on essential objects of the domain.

While **LL** outperforms the frequency baseline, for our domain it has proven to be one of the weaker measures and we could not identify any further useful characteristics.

The unsatisfactory result of the ranking by **CSvH** is based on its strong emphasis on MWTs which have a head with a very high domain corpus frequency. Thus, the result lists show many general-language candidates at their top which is why the measure underperforms the baseline.

The above results were obtained from an evaluation of a mixed set of SWTs and MWTs. A separate analysis of the MWT extraction (cf. figure 3) shows that the two best-performing measures (DS and CSmw) achieve an F_1-score of only about 0.49 (recall 64%, precision 40%). In comparison, the maximum F_1-score for the extraction of single-word terms was about 0.65. The low performance of the MWT extraction is due to the fact that this task also includes the determination of the right length of the MWT and therefore leads to more noise (in total approximately 80% noise in the MWT candidates selected by the basic POS patterns). Thus, a further filtering of the multi-word candidates is necessary.

4.2 Effect of C-value

Out of the 226,715 items that follow our extended patterns (frequency ≥ 1), C-value successfully removes 58,491 cases of noise that only occur embedded (25.8%). In our GS-based evaluation, C-value outperforms mere frequency, as shown in figure 2 in the extraction of the basic term patterns.

candidate term	freq	C-value
Band	301	296.50
Klebeband	376	707.33
doppelseitiges Klebeband	117	342.00

Table 4: Comparison of frequency to C-value

The positive effect of C-value is illustrated with a few examples in table 4. The frequency of occurrence in the domain corpus of the first candidate term '*Band*' (tape) is relatively similar to the one of the second candidate '*Klebeband*' (adhesive tape). They are both single-word nouns and thus would be considered almost equally as terms for the domain. After applying the C-value approach however their termhood values differ clearly with '*Klebeband*' having a value twice as high as the value of '*Band*'. This mainly follows from the term length computation based on the number of components of compounds ('*Klebeband*' is a compound with two components: '*kleben*' (to glue) and '*Band*'). Furthermore, the frequency of '*doppelseitiges Klebeband*' (double-sided adhesive tape) is only approximately a third of the frequency of the single noun '*Klebeband*'. The C-value method here also rewards the length of the multi-word and computes a value that is half of the C-value of the single noun despite the much lower absolute frequency of '*doppelseitiges Klebeband*'. Note that the termhood value of '*doppelseitiges Klebeband*' is also greater than the value for '*Band*' even though it has a lower frequency. This shows that the fine-grained measurement of the length characteristic of candidate terms including a special treatment for compounds is beneficial for terminology extraction.

However, it has to be noted that the ranking of candidate terms by C-value alone is not sufficient for term extraction, as extracted top lists with a recall of greater than 50% still contain a considerable amount of noise (at least 78%), mostly in the form of general-language candidates.

We found that CSmw, one of the best-performing measures in the comparative evaluation, improves when domain-specificity is computed on C-value instead of frequency (*CSmw+C*-plot in Figure 3, maximum F_1-score 0.51). This is due to C-value's sensitivity for nested terms which is combined with the domain-specificity filter from CSmw.

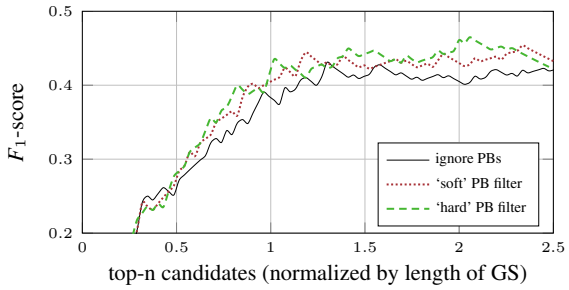


Figure 4: Syntactic validity filter for N P N extraction

4.3 Effect of phrase boundaries

As the syntactic filter only affects POS sequences with prepositions whose proportion in the GS is rather small, the effect of phrase boundaries (PBs) cannot be shown in Figure 3 and was thus tested in two other settings. First, we evaluated against the 107 N P N terms in the gold standard. Figure 4 shows the effects of applying the "soft" filter (frequency adjustments, section 2.1) and the "hard" filter (candidate removed altogether) on the F_1-score. Both filters clearly improve the standard extraction based on CSmw. However, the filter affects more than just the terms in the GS (17,4% of all NP+PP candidate occurrences affected) and we would like to observe the effects on all variants of prepositional patterns. Thus, in a precision-based evaluation, we ranked the MWT candidates by the number of times they violated the syntactic filter and manually checked, for the top 500 removal candidates, whether the removal was justified. The result, as shown in Table 5, indicates that the quality of the parser output is sufficient to predict syntactic validity: the overall precision for these top 500 candidates was 83%.

Top n	50	100	150	200	250
Precision	0.76	0.75	0.78	0.81	0.81
Top n	300	350	400	450	500
Precision	0.82	0.83	0.82	0.82	**0.83**

Table 5: Top-n manual plausibility check for "hard" filter

5 Conclusion and future work

We presented three steps to remove noise and to increase performance in nominal terminology extraction. We also suggested a combination of statistical measures that is particulary suitable for this task: Our best setting has proven to be a combination of C-value and CSmw, together with a syntactic validity check. A qualitative analysis of the extraction results showed that different termhood measures emphasize different characteristics of terms, as their top lists differ. Therefore, a combination of statistical measures can also be considered for further improvements, instead of only focusing on one single best performing measure. One could for example think of ways to combine the top lists of a set of best-performing measures, or try an approach that combines or ranks different scores of certain measures in one formula. Future work will also be based on English data where we will evaluate further steps to improve term extraction results, e.g. by combining the termhood measures also with association measures and by further improving the syntactic analysis through the use of an additional constituency parser. A further objective of this work will be to assess the generality of the approach on different domains.

References

Khurshid Ahmad, Andrea Davies, Heather Fulford, and Margaret Rogers. 1992. What is a term? The semi-automatic extraction of terms from text. *Translation Studies: An Interdiscipline: Selected papers from the Translation Studies Congress, Vienna, 1994*, pages 267–278.

Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In *Text REtrieval Conference*.

Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2001. A contrastive approach to term extraction. In *Terminologie et intelligence artificielle. Rencontres*, pages 119–128.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China*, pages 89–97. Association for Computational Linguistics.

Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2010. A contrastive approach to multi-word term extraction from domain corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, Malta*, pages 19–21.

Maria Teresa Cabre and Jorge Vivaldi Palatresi. 2013. Acquisition of terminological data from text: Approaches. In *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguis-*

tics and Communication Science (HSK) 5/4, pages 1486–1497. DeGruyter Mouton.

Fabienne Cap. 2014. Morphological processing of compounds for statistical machine translation. PhD thesis, Institute for Natural Language Processing (IMS), University of Stuttgart, `http://elib.uni-stuttgart.de/opus/volltexte/2014/9768/`.

Chaomei Chen, Fidelia Ibekwe-SanJuan, Eric SanJuan, and Michael Vogeley. 2008. Identifying thematic variations in sdss research. In *6th International Conference on Language Resources and Evaluation Conference (LREC-08)*, pages 319–330. Presses Universitaires de Lyon.

Jinho D. Choi, Joel R. Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 387–396.

Jean-François Couturier, Sylvain Neuvel, and Patrick Drouin. 2006. Applying Lexical Constraints on Morpho-Syntactic Patterns for the Identification of Conceptual-Relational Content in Specialized Texts. *Language Resources and Evaluation Conference (LREC-06), Genoa, Italy*.

Stefan Evert. 2005. The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD thesis, Universität Stuttgart, `http://elib.uni-stuttgart.de/opus/volltexte/2005/2371`.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – A Corpus of Parsable Sentences from the Web. In *Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, Language processing and knowledge in the Web: 25th International Conference, GSCL 2013, Darmstadt, Germany, volume 8105 of Lecture Notes in Computer Science*, pages 61–68. Springer.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Katerina T. Frantzi and Sophia Ananiadou. 1996. Extracting nested collocations. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 41–46. Association for Computational Linguistics.

Katerina T. Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Tanja George. 2014. Comparing a commercial term extraction tool with a research prototype: an evaluation study on DIY instruction texts. Bachelor thesis, ms. Institute for Natural Language Processing (IMS), University of Stuttgart.

Anita Gojun, Ulrich Heid, Bernd Weißbach, Carola Loth, and Insa Mingers. 2012. Adapting and evaluating a generic term extraction tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Elizaveta Loginova, Anita Gojun, Helena Blancafort, Marie Guégan, Tatiana Gornostay, and Ulrich Heid. 2012. Reference lists for the evaluation of term extraction tools. In *Proceedings of the Terminology and Knowledge Engineering Conference (TKE'2012)*.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge Mining*, pages 255–279. Springer.

Paul Rayson and Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9*, WCC '00, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mathieu Roche, Jérôme Azé, Yves Kodratoff, and Michele Sebag. 2004. Learning interestingness measures in terminology extraction. a roc-based approach. In *"ROC Analysis in AI" Workshop (ECAI 2004), Valencia, Spain*, pages 81–88.

Johannes Schäfer. 2015. Statistical and parsing-based approaches to the extraction of multi-word terms from texts: implementation and comparative evaluation. Bachelor thesis, ms. Institute for Natural Language Processing (IMS), University of Stuttgart.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 777–784. Association for Computational Linguistics.

Marie Zollmann, Ina Rösiger, Ulrich Heid, and Michael Dorna. 2016. Nutzen von Parsing in der Termextraktion: eine qualitative und quantitative korpusbasierte Untersuchung. In *Poster session of the DGfS conference 2016 (to appear)*.

OMTAT annotation tool: semantical enrichment for legal document search *

Sylvie Szulman
LIPN, Université Paris 13
Sorbonne Paris Cité & CNRS
France
ss@lipn.univ-paris13.fr

François Lévy
LIPN, Université Paris 13
Sorbonne Paris Cité & CNRS
France
fl@lipn.univ-paris13.fr

Eve Paul
LuapLab
France
eve.paul@luaplab.com

Abstract

This paper describes a help system for legal document searching. The proposed approach relies on creating specific annotations over a corpus of documents. A tool has been built which implements the visualization of annotations, texts and semantic resources, the creation of annotations and their collation in resources. A search engine has been implemented as well to query the set of annotated documents in order to answer user questions.

1 Introduction

This paper presents a system intended to facilitate the access to French legal documents. Given the huge amount of existing legal documents, we developped a document search system for non lawyer users (trade unionist, human resources manager, etc.) to find relevant information in the overabundance of documents. Even if there are legal databases and law data Banks for the French law, most of them target professional users, so they are hard to access for the non lawyer user. Our system should be a base to make the documentation more accessible to non lawyers. The domain of experiences is restricted to the *Code du Travail* (labour code) and collective labour agreements.

The proposed approach relies on creating specific annotations over the reference documents. "Annotation" is used here in an NLP sense (a specific mark in the text) and not in its legal sense (a comment mainly based on jurisprudential cases). The work relies on the textual annotation standoff format defined by the Brat tool ((Stenetorp et al., 2012b; Stenetorp et al., 2012a)) which has been used in particular in the BioNLP domain. The approach is supported by a tool that we have built, which allows user to annotate documents with our different kinds of annotations and to query the set of documents and annotations. The tool thus enables user to rapidly find answers to a legal question in the domain.

General ideas are illustrated with an example where our approach may help a user to extract the most relevant legal excerpts for his problem. Consider the case of a professional newsman which regularly works for a journal as a freelance, so he is paid by the piece. Suddenly the journal ceases giving him work. He wants to know if he has some right to an indemnity. The base document to query is the labour code, which is 1800 pages long. This use case is cited as an example all along the paper, and is specifically considered in section 6.

The rest of the paper is divided as follows. In section 2, the approach is positioned with respect to other approaches for the access to legal documentation. Section 3) presents the types of annotations that have been defined in order to give an account of legally significant properties of the text. The OMTAT tool which has been built to visualize, explore or add annotations is described in section 4. Section 5 is specifically devoted to the query engine. The 6th and last section contains a short description of the use case.

*This work is part of the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083

2 Accessing Documentation in the legal Domain

2.1 Information Retrieval

Information Retrieval in the legal domain consists in retrieving law articles and case law decisions related to a given subject matter. The search can be by reference, for instance the "Loi Aubry" or "loi sur les 35 heures" (35hours working week law). It can also be a search of keywords in plain text, or in some specific sites of a previously defined structure. Advanced search options allow to search by document descriptors (previously annotated with the help of semantic ressources, like a thesaurus).

Major law-editors (Dalloz[1], Editions Francis Lefebvre [2], Lamy[3] , Lexis-Nexis[4]) offer to clients a large quantity of documents and an engine customized for their documents and their descriptors. Their services are accessed mostly by professionals, due to their commercial offers. Some smaller law editors make available, in the *Code du Travail* domain, mementos and/or fact-sheets that can be bought by non-lawyers economic players (named NLEP in the following): trade-unionists, human resources managers, managers of small business, etc. NLEP can then refer back to these summaries to answer practical questions. Our present approach proposes to make sources of law available for NLEP, through a semantic and structural search in documents which have been previously semantically and structurally annotated.

2.2 Semantic Approach

Information Retrieval in the legal domain has recently interested the academic community. (Berry et al., 2012) consider how different language models of the collection succeeded or failed to be used by domain specialists, (Mimouni, 2015) studies two approaches for querying a collection, viewed as a network of documents. The first is based on Formal Concept Analysis, the second on semantic web technologies, namely an ontology to annotate the collection and Sparql to query it.

In a semantic approach as the one adopted here, semantic resources, i.e. ontologies, thesauri or terminologies in the legal domain, play an important role. Resources of this kind exist but, due to business reasons, they are mostly not public.

Among public thesauri, EuroVoc is a multilingual thesaurus produced by the European Union[5]. It describes the terminology used by the different domains of activity of this Union and is available in twenty three languages. It is used, among others, by the European Parliament, the Publications Office of the European Union, some national and regional parliaments in Europe, as well as by national administrations and private users in different states, some members of the Europeaan Union and some not members.

Jurivoc[6] is a public, trilingual thesaurus used in the Swiss Confederation. It was produced by the Swiss Federal Court and the Insurance Federal Court.

These resources cope with domains other than the labour code and cannot be used as such for annotating the French *Code du Travail*. A terminology of French employment law is in construction by one of the authors who is a legal expert. Currently the built terminology is restricted to the use case. It allows to create a set of terminological annotations.

2.3 Standard based Initiatives

XML (eXtensible Markup Language) is a metalanguage allowing to define for a particular domain a set of semantic and structural marks. Several XML standards have been defined for the encoding of legal documents, like MetaLex and AkomaNtoso.

- MetaLex has been devised by the CEN as a Workshop Agreement which standardizes the way in which sources of law and references to sources of law are to be represented in XML. It involves an Open XML Interchange Format for Legal and Legislative Resources, so as to avoid locking a client to a provider (Boer et al., 2008).

- AkomaNtoso (Palmirani et al., 2005) defines a set of simple technology-neutral electronic representations in XML format of parliamentary, legislative and judiciary documents.

[1] http://www.editions-dalloz.fr/

[2] http://www.efl.fr/

[3] http://www.wkf.fr/accueil.html

[4] http://www.lexisnexis.fr/

[5] http://eurovoc.europa.eu/drupal/?q=fr

[6] http://www.bger.ch/fr/index/juridiction/jurisdiction-inherit-template/jurisdiction-jurivoc-home.htm

Note that XML standards are not exclusive of our approach, for semantic and structural markups can be converted into annotations. An advantage of our approach is the possibility to create relational annotations, that is annotations bearing on other annotations which are difficult to represent in XML.

3 Annotations

From a NLP point of view, an annotation links data to a text fragment or to other annotations. The data can be of many kinds, syntactic (POS, grammatical roles, etc.), semantic (lexical entry, terminological element, ontological entity, anaphora...), discursive (focus, concession, restriction, emphasis, etc.), and free comment. The text fragment can be all in one segment, or can involve several segments. In our approach as in most works focussed on technical contents, the annotations considered have constrained data (enumerated *a priori* in a closed list). The text fragments to which they are attached can be made of separate pieces, and they can be structured by other (lower level) annotations.

Different kinds of knowledge help legal texts users to search in the text, and so OMTAT uses different types of annotations to reflect the particular role of each.

Keywords (*named AnnotationK*) are exact denominations (except morphological variations) which have a specific meaning in the domain and can be by themselves a clue. For instance, *Partie* (Part), *Titre* (Title[7]), *Chapitre* (Chapter) *Article* (Article) are keywords : using a synonym of Part as *Fragment* (Fragment) is impossible, as the meaning of these words is only defined by the hierarchy in the table of contents. Note that the heading of a division is under the scope of the relevant keyword and is attached as an attribute of the annotation.

Terms (*named AnnotationT*) represent significant entities of the domain. Terms can be supported by words or multiwords, but their significance relies on the attached meaning, represented in the term-annotation label. Note that different words may receive the same meaning and that, most often, one of them is chosen to represent this meaning. In the *Code du travail*, *salaire*, *sanction* or *contrat de travail* are terms (the meaning of the first can also be supported by e.g. *traitement*, *appointements*, *paie*). As in many domains, legal terms are often gathered in specialized resources like terminologies or ontologies.

Relations (*named AnnotationR*) allow to link two terms by a specific relation. For instance, an *authority_for* relation links *entreprise de presse* to *journaliste* in *Est journaliste professionnel toute personne qui a pour activité principale, régulière et rétribuée, l'exercice de sa profession dans une ou plusieurs entreprises de presse et en tire l'essentiel de ses ressources* (Is a professional journalist any person whose main, regular and paid activity consists in exercising his profession in one or more newspaper companies and who obtains this way the main part of its resources).

References (*named AnnotationL*) annotations mark fragments which refer to other fragments of a legal text, most often with the help of a standard identifier, e.g. *La présomption de salariat prévue à* ***l'article L. 7121-3***, or: *Lorsque le travail du journaliste donne lieu à publication dans les conditions définies à* ***l'article L. 132-37 du code de la propriété intellectuelle***. When the reference is relevant for the question at hand, the text under consideration must be extended with the fragment referreed to by the annotation. Three main specific relations can link an annotation to a reference : *defined_in*, *decided_in*, *mentioned_in*

Events (*named AnnotationE*) link sets of arguments around a central predicate, the *trigger*. For instance, in *Le salaire perçu par un mannequin pour une prestation donnée* (The salary received by a model for a given performance), the trigger of the event is *perçu*, its theme is *salaire*, its agent is *un mannequin* and its cause is *une prestation donnée*.

Context (*named AnnotationC*) annotations cover a possibly large fragment having a functional

[7] In French law, Title is a level in the hierarchy of divisions, not to be mistaken for the text of a heading of any level, named its title

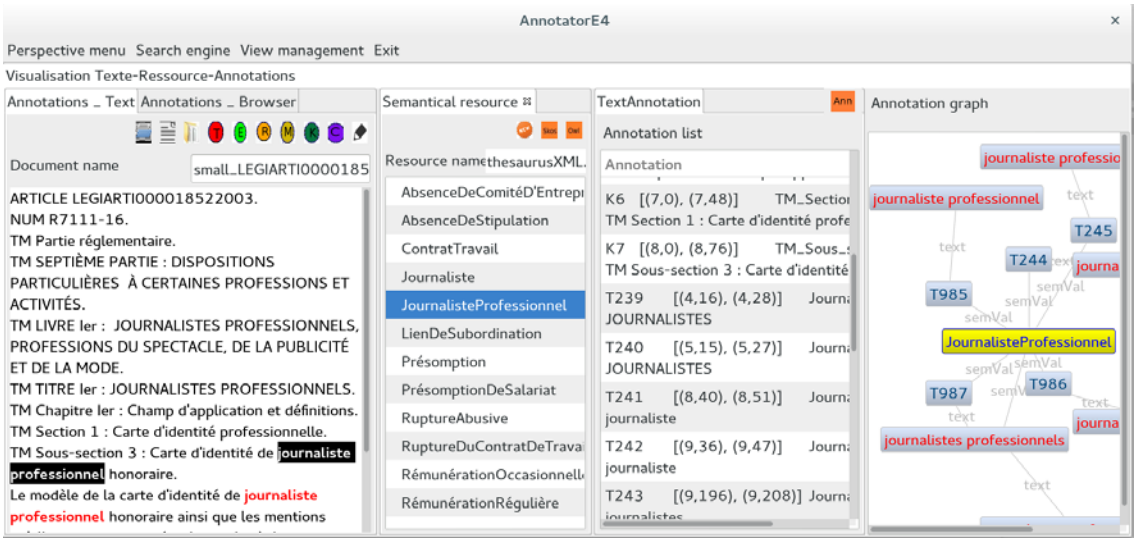


Figure 1: Main annotation view

role by which interpretation has some specificity. In the *Code du Travail*, context annotations correspond to entities in the table of contents, because the functional view is reflected in this table. For instance, article L7111-3 states *Si l'employeur est à l'initiative de la rupture, le salarié a droit à une indemnité qui ne peut être inférieure à la somme représentant un mois, par année ou fraction d'année de collaboration, des derniers appointements*[8]. It cannot be accurately understood without accounting for its position: in the legislative part, part VII (Rules specific to some professions), Livre 1 Title 1 (journalists), Chapter 2 (employment contract) Section 2 (breach of contract) - hence "the salaryman" need to be a professional journalist. Note that, in other texts as case law decisions for instance, judgments have functional parts which remain implicit, while they must nevertheless be recognized.

Terms, relations and events have been widely used in the BioNLP challenge, thanks to the outer encoding provided by the Brat tool[9] ((Stenetorp et al., 2012b; Stenetorp et al., 2012a)). Other annotations use the same kind of encoding and have been devised for the specific needs of legal annotation.

4 OMTAT Tool

In this section, we present the OMTAT (One More Text Annotation Tool) system, built as an E4 Eclipse application to implement annotations as described above. Many functionalities have been defined to create and/or visualize different annotations. The main window involves four views (see fig 1):

- Text view: it shows the text of the document (and its name). All the annotations are emphasized with different text colors according to their type.
- Semantic resource view: it shows the semantic resource (a thesaurus or an ontology). Any semantic resource in SKOS format or OWL format may be loaded. A SKOS semantic resource may be enriched when an annotationT is created and its semantic value does not exist in the resource. A OWL semantic resource cannot be modified. In the figure, the semantic resource is a thesaurus. It is built on the labour code and restricted to the use case.
- Text annotation view: it displays all the annotations defined on the document showed in the text view. A click on an annotation selects the corresponding text in the text view.

[8] Translation: If the breach is initiated by the employer, the salaryman has a right to an indemnity which cannot be less than the amount of money representing one month per year of fragment of a year of collaboration, of the last salary

[9] http://brat.nlplab.org/about.html In the Brat tradition, Terms are named *Types*

- Annotation graph view: it shows the selected annotations in the form of a graph which nodes correspond to values and edges correspond to labels.

Other global functionalities include:

- Text and sets of annotationK(s) may be extracted from XML files provided that markups are defined in a configuration file.
- A semantic resource provided with preferred labels may be cast on a text to create terminological annotations (AnnotationT).

From the annotation point of view, OMTAT has common points with well known tools, GATE [10] (Cunningham, 2002; Cunningham et al., 2011) and BRAT[11] (Stenetorp et al., 2012b), but has also significant differences. GATE considers one single category of annotations of the form (type, span, features) ; the type is a single word, features are key-value pairs. Gate can model structured annotations with the help of a "constituent" feature which records lists. GATE documentation states that "No special operations are provided in the current architecture for manipulating constituents" [12]. After version 7, the provided library allows modeling relations through a `members[]` array[13]. The 8.1 basic version does not offer a user interface to manually add relations as it does for plain annotations. A search tool locates the next occurrence of a string or regex and can restrict the search to the scope of existing annotations.

Brat has three categories of annotations: types, relations and events. *Types* are the basic blocks associating a (possibly discontinuous) portion of text with a type. A *Relation* involves two typed annotations and a labelled link. An *Event* is made of a main typed annotation (its *trigger*) linked to a variable number of arguments. The last two can be represented in Gate by relations, but Brat emphasizes their difference from a user point of view: the relation need no lexical support (e.g. there is no lexical link between a bacterium and the mouth which is its localization) while the trigger of an event is its lexical support and restricts possible arguments. Last, Brat is designed to be used as a centralized collaborative tool through a web server. Its user interface combines text and drawings to visualize or add all categories of annotations, while the annotation schema is centrally controlled. A search tool locates annotations of a given category according to the conjunction of conditions on their text and on their attributes.

In comparison, OMTAT is a single user tool under Java as Gate (but Gate has a collaborative extension). It has the three categories of Bratt, it can read its standoff format and defines some more categories to account for the structure of legal documents. It accepts to dynamically manage semantic resources (Gate also does) and includes a search engine which is described in the following section.

5 The Search Engine

An experimental search engine has been developed for the need of exploring the annotated corpus currently in memory. Its main role is to build so called w-tuples, tuples of elements (i.e. annotations, sentences and documents) constrained by a set of conditions. At the engine level, a query has a simple `Select ... From ... Where ...` form and returns s-tuples, tuples of attribute values. For example, figure 2 shows a plain graphic interface to the engine and a query returning the list of pairs $(S.text, Aart.text)$ where $S.text$ is the text of a sentence containing a Term annotation *Presumption of Salary*, and $Aart.text$ is the number of the article containing the sentence[14]. The function of each clause is shortly explained here:

- The `From` clause provides a set of open (in memory) documents in which the search will happen.
- The `Where` clause describes the form of w-tuples and the conditions that they must satisfy. For that purpose, every element in the tuple is named and typed (the type may be annotation, sentence or document; for the sake of conciseness, the first letter of the name involves the type). Conditions are then relations between attributes of the elements and possibly constant values. For example, $S.numsent < 5$ is a condition requiring that

[10] General Architecture and Text Engineering, http://gate.ac.uk

[11] Brat rapid annotation tool, http://brat.nlplab.org/

[12] Gate inline documentation, section 5.4.2

[13] Documentation, section 7.7

[14] Hovering the mouse over truncated sentences shows the full text. The OK button saves the result.

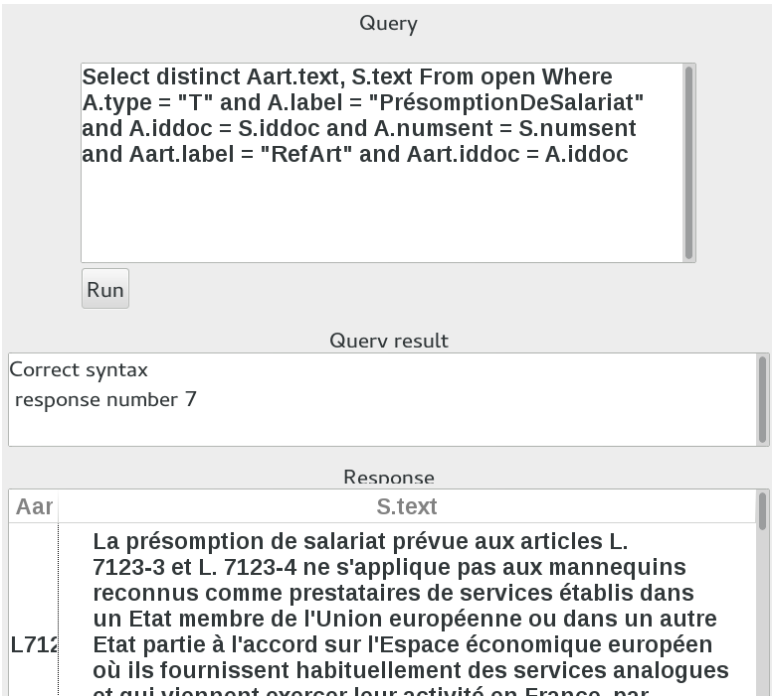


Figure 2: The Query view

the element S (a sentence) be at the beginning of the document (in the first five sentences), and $A\_obj.label = A\_agent.label$ is a condition requiring that the two elements A_obj and A_agent (both annotations) share the same label.

- The `Select` clause describes what s-tuple is returned for each built w-tuple. Only *element.attribute* names can be used here, and all the element names must be defined in the w-tuples, i.e. used in the `Where` clause. Any attribute of a defined name can be used in the Select. Last, a `Distinct` modifier allows to merge identical returned tuples.

For instance, the query

`Select distinct` $A1.numsent$
`From` $current$
`Where` $A1.numsent$ `=` $A2.numsent$
  `and` $A1.idannot$ `!=` $A2.idannot$
  `and` $A1.label$ `=` $A2.label$

returns the sentence number of those sentences in the current document in which the same label occurs twice.

The different kinds of annotations described in section 3 are implemented as classes, benefiting from inheritance. Attributes are provided by a specific inner mechanism of annotation classes allowing to define computed attributes which are not explicit in the Brat format (the size of the text, the number of arguments of an event, etc.). The kind of annotations is identified by the $type$ attribute. Some more attributes are common to all annotations, and others are specific to one or several types. In the present version, common attributes are the identifier of the annotation, its type, its label, its size, its start and end position, its sentence number and the identifier of its document. Other attributes may be for instance the subject and the object of a relation (type "R") or the division of a context annotation (type "C"), i.e. the kind of textual unit (chapter, paragraph) which defines its borders. Elementary conditions in the `Where` clause are only combined by conjunction; a limited form of disjunction is available through inequalities and string relations *starts with*, *ends with*.

Parsing provides a control of the query and of its conditions. Elementary conditions are sorted according to the names in their left and right hand part. w-tuples are then built recursively on the set of names involved. In the application of con-

ditions, the validity of many attributes can only be decided when the type is known; furthermore, some of them are optional in the type (*e.g.* the arguments of an event). So the validity of attributes used by conditions is dynamically determined while the w-tuples are built.

6 Use Case

6.1 Presentation

We have considered annotation of the French *Code du Travail*, which has 4644 articles. In a legal sense, a French Code groups and organizes the legal and regulatory texts produced along time for a given domain. Jurisprudential knowledge comes from other sources. A legal user of this text must discover as quickly and easily as possible which fragments are relevant to his problem, and may have to examine many excerpts.

Following a well established layout of French codes, we stored each article in a separate document which includes a reminder of its position in the table of content: the document starts with headings of all levels in the scope of which it is. At the moment, a small set of annotations have been marked in the text: *Contrat de travail* (employment contract), *Journaliste* (journalist), *Journaliste professionnel* (professional journalist), *Présomption* (presumption), *Rémunération* (remuneration).

Recall the example legal problem of section 1, for which our approach may help a NLEP to find useful excerpts of the labour code answering to his problem. A freelance journalist used to work for a newspaper, and at a moment the newspaper does not ask him anymore article. Being freelance, the journalist is paid for each provided product (article, photograph, drawing, etc..). In the case, he has never signed any written contract whatsoever. The stakes are if he has a right to an indemnity.

6.2 A first Analysis

Articles can be selected according to the presence of annotations *Journaliste* in the document, including titles. This yield 336 distinct answers, due in particular to one of the titles of level *Livre*, which enumerates various professions, so for instance articles about (theater, movies, ...)-players are also under this title. Restricting the annotation to occur in the body of the article reduces to 45 articles, 17 of which are in laws and can be focused on by requiring a K annotation with a *Partie Législative* label. Requiring a supplementary annotation on *Contrat de Travail* only leaves 5 articles.

Among these, article L 7112-1 states *Toute convention par laquelle une entreprise de presse s'assure, moyennant rémunération, le concours d'un journaliste professionnel est présumée être un contrat de travail. Cette présomption subsiste quels que soient le mode et le montant de la rémunération ainsi que la qualification donnée à la convention par les parties*[15]. Note also that the section is entitled *Présomption de salariat* (Presumption of wage relation), implying that wages are a form of payement specifically attached to employment contracts.

This means that, provided he is a professional, the freelance is presumed to be governed by an employment contract and hence to benefit of a right to an indemnity. The user must enjoy a minimal understanding of legal reasoning to supplement this basic information with the help of some more queries. Namely, he must know that, beside employment, other kinds of contracts may be relevant in order to "remunerate the support" of some individual, and that employment contracts themselves may have different categories. Searching for *Journaliste* and *Remuneration* annotations in the same article gives six results,

One of them is L. 7113-3: *Lorsque le travail du journaliste professionnel donne lieu à publication dans les conditions définies à l'article L. 132-37 du code de la propriété intellectuelle, la rémunération qu'il perçoit est un salaire*[16]. It means that it is possible to argue that the payment is of the kind specific to intellectual property, i.e. an authorship rights remuneration, and not an employment contract. The conditions of this possibility are to be found in the Intellectual Property Code. To maintain the legal complexity inside reasonable bounds, this path is not followed here.

[15] Any agreement by which a press company obtains, through remuneration, the support of a professional journalist, is presumed to be an employment contract. This presumption remains whatever can be the mode and the amount of the remuneration or how participants qualify the agreement.

[16] When the work of the professional journalist gives rise to a publication in such conditions as defined in article L. 132-37 of the Intellectual Property Code, the remuneration that he receives is a salary

6.3 More on the Contract

Every French worker knows that employment contracts belong to one of two categories : contracts with an indeterminate duration (CDI) or with a determinate one (CDD), and that they do not involve the same rights. More, article L. 7112-2, one of the five obtained from the first query, considers a breach of the CDI, while no mention is made of a CDD. Definitions can only be found considering articles which apply to the generic case. Searching for titles of the highest possible level annotated with *employment contract* yields Part 1 (individual employment relations) Livre II (Employment contract).

Searching in this *Livre* for articles annotated both with CDI and CDD provides two basic texts. In L. 1221-2 it is stated that the CDI is the default case: *Le contrat de travail à durée indéterminée est la forme normale et générale de la relation de travail. Toutefois, le contrat de travail peut comporter un terme fixé avec précision dès sa conclusion ou résultant de la réalisation de l'objet pour lequel il est conclu dans les cas et dans les conditions mentionnés au titre IV relatif au contrat de travail à durée déterminée*[17]. And in L. 1242-12, this is enforced by requirements on the form of the CDD: *Le contrat de travail à durée déterminée est établi par écrit et comporte la définition précise de son motif. A défaut, il est réputé conclu pour une durée indéterminée*[18].

7 Conclusion

This article has described the use of different textual annotations to help legal documents search take advantage of the document structure. An experimental use case demonstrates the utility of the approach. An implementation has been carried out which allows to explore, query and add annotations. Future work will allow us to deepen the study of links between semantic annotations and semantic document search.

[17] The CDI is the normal and standard form of employment relation. However, the employment contract may involve an end date precisely fixed at start or determined by the achievement of the object in view of which it is decided, in cases and conditions mentioned in Titre IV related to CDD

[18] The CDD is set up in writing and contains a precise definition of its motive. Failing that, it is deemed to be agreed for an indeterminate duration

Thanks

We thank Jorge Garcia Flores and Nadi Tomeh for their friendly review of our writings.

References

Michael W. Berry, R. Esau, and Bruce Keifer. 2012. The use of text mining techniques in electronic discovery for legal matters. In C. Jouis, I. Biskri, Jean-Gabriel Ganascia, and M. Roux, editors, *Next Generation Search Engines: Advanced Models for Information Retrieval*, pages 174–190. IGI Global.

A. Boer, A. Winkels, and F. Vitali, 2008. *Metalex xml and the legal knowledge interchange format.*, volume Computable Models of the Law - Lecture Notes in Computer Science -4884, pages 21–41. Springer.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*.

H. Cunningham. 2002. Gate - a general architecture for text engineering. In *Computers and the Humanities, Volume 36*, pages 223–254.

Nada Mimouni. 2015. *Interrogation d'un réseau sémantique de documents: l'intertextualité dans l'accès à l'information juridique*. Ph.D. thesis, Paris 13 - Sorbonne Paris Cité, janvier.

M. Palmirani, R. Brighi, and M. Massini. 2005. Automated extraction of normative references in legal texts. In NY USA ACM, editor, *Proccedings of the 9th international conference on Artificial intelligence and Law, ICAIL*, pages 105–106.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Sophia Ananiadou, and Akiko Aizawa. 2012a. Normalisation with the BRAT rapid annotation tool. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine*, Zürich, Switzerland, September.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012b. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.

Génération automatique de HashTags

Guillaume Tisserant
LIRMM & Awadac
tisserant@lirmm.fr

Mathieu Roche
TETIS & LIRMM
mroche@cirad.fr

Violaine Prince
LIRMM
prince@lirmm.fr

Résumé

Les hashtags sont des mots-clés que les utilisateurs de réseaux sociaux choisissent de mettre en avant dans leurs messages. Ils ont été popularisés sur le réseau social Twitter, qui a permis à ses utilisateurs de sélectionner des HashTags à suivre et d'afficher l'ensemble des messages contenant un HashTag suivi. Ils sont aujourd'hui utilisés sur les principaux réseaux sociaux, tels que Facebook, Google+, Diaspora*, et sont un facteur important de la diffusion de l'information sur Internet. Dans cet article, nous proposons une méthode fondée sur des informations statistiques, syntaxiques et sémantiques pour générer des HashTags.

1 Introduction

Avec l'usage grandissant d'Internet, une quantité d'information de plus en plus importante se trouve à notre disposition. La difficulté n'est donc plus seulement de multiplier les ressources auxquelles nous pouvons accéder mais de trier les informations auxquelles nous accédons. Pour permettre à ses utilisateurs de sélectionner le contenu qui l'intéresse, certains réseaux sociaux, comme *Twitter*, proposent aux utilisateurs de sélectionner des HashTags, qui correspondent à leurs centres d'intérêts et affichent dans leur flux les messages contenant des HashTags suivis. Les utilisateurs voient les HashTags contenus dans les messages de leur flux, et peuvent les ajouter à leur liste de HashTags suivis.

Toutefois, le choix de HashTags à suivre ou à utiliser pour indexer un de ses tweets peut être difficile à réaliser : un HashTag trop générique va indexer le message dans un flux très important de données, il ne sera donc pas mis en valeur alors qu'un HashTag trop spécifique risque de ne pas être suivi. L'exploitation des HashTags demande donc un effort plus important à l'utilisateur. La création d'un système automatique de recommandation de HashTags est donc une solution intéressante pour faciliter l'accès aux ressources par les utilisateurs.

2 Problématique

2.1 Qu'est ce qu'un HashTag?

Les HashTags sont des termes que les utilisateurs des réseaux sociaux, en particulier Twitter, choisissent de mettre en avant dans leurs messages en les faisant précéder du symbole #.

Un HashTag peut avoir plusieurs significations. Il peut servir à référencer un tweet par rapport à un sujet ou à participer à une discussion en cours sur le sujet associé au HashTag (Huang et al., 2010). Par exemple, les HashTags #2012, #Elysée, et #Elysée2012 ont servi à indiquer qu'un tweet évoquait la campagne présidentielle de 2012. Mais les HashTags servent aussi à marquer son appartenance à une communauté ou une idéologie (Conover et al., 2011). Twitter, pendant les dernières élections présidentielles en France, a vu se multiplier des noms de candidats (#Eva pour Eva Joly, #NS pour Nicolas Sarkozy), des partis (#UMP, #PS) ou des slogans (#Placeaupeuple) utilisés comme HashTags.

Récemment, de nombreux travaux ont montré l'intérêt d'exploiter les HashTags dans le cadre de tâches de fouille de texte. (Conover et al., 2011), par exemple, utilisent les HashTags comme descripteurs pour la classification et mettent en avant le fait qu'ils sont plus pertinents que les autres termes. (Ozdikis et al., 2012) utilisent les HashTags pour faire du clustering. Ils montrent que les HashTags permettent un meilleur partitionnement des tweets. Ils montrent aussi que l'enrichissement sémantique pour des tâches de partitionnement est de meilleure qualité en se fondant sur les HashTags plutôt que sur les termes. Cela montre que les HashTags sont des données à la fois exploitables par des êtres humains et par des programmes automatiques.

2.2 Comment générer des HashTags?

La recommandation de HashTags est un domaine encore peu exploré (Kywe et al., 2012). La majorité des approches s'appuient sur des méthodes statistiques, comme (Zangerle et al., 2011) qui utilisent la pondération *TF-IDF* ou (Godin et al., 2013) qui exploitent le modèle *Latent Dirichlet Allocation*. Ces approches se fondent sur l'idée qu'un HashTag contient une information qui a pour but d'indexer un tweet par rapport à un sujet. Mais un HashTag possède aussi une sémantique importante évoquée dans le tweet. Générer des HashTags depuis les tweets demande donc de détecter des termes qui soient à la fois sémantiquement intéressants et marqueurs d'une opinion ou d'une appartenance à un groupe. Nous allons, dans cet article, proposer une méthode à la fois statistique et sémantique, permettant de détecter les termes les plus discriminants pour l'indexation et les plus intéressants sémantiquement.

3 Analyse

Pour développer notre méthode de génération de HashTags, nous avons commencé par analyser les termes utilisés en tant que HashTags dans un corpus de tweets, et nous les avons comparé à des termes issus d'une analyse statistique d'un corpus de tweets, et des termes issus d'une ressource sémantique en rapport avec notre corpus. Dans cette section, nous présenterons ces différentes ressources, ainsi que notre corpus.

3.1 Les ressources utilisées

Pour comprendre quels termes pouvaient être des HashTags potentiellement intéressants, nous sommes partis d'un corpus de tweets politiques, et nous avons comparé les termes utilisés en tant que HashTags avec des termes statistiquement discriminants, et avec une liste de termes politiquement clivants, mettant en avant des opinions politiques. La méthode statistique pour favoriser les termes discriminants provient d'une méthode appelée *GenDesc* (Tisserant et al., 2014) et la ressource sémantique

utilisée pour sélectionner les termes politiquement clivants provient du GWAP (Game With A Purpose) *PolitIt* (Tisserant, 2015).

3.1.1 Le corpus

Pour tester nos méthodes et faire nos mesures statistiques, nous avons utilisé un sous-ensemble du corpus POLOP (Political Opinion Mining) (Bouillot et al., 2012). POLOP rassemble des tweets écrits en français par des élus de différents partis politiques pendant la campagne présidentielle de 2012. Nous avons travaillé sur un sous-ensemble du corpus de 2500 tweets équitablement répartis entre 5 partis politiques. Ces mouvements sont l'*UMP*, le *MoDem*, le *PS*, le *Front de Gauche*, et *EELV*. Les tweets sélectionnés font en moyenne 81 caractères. Le plus petit tweet fait 31 caractères, le plus long en fait 140.

3.1.2 *GenDesc*

GenDesc (Tisserant et al., 2014) est une méthode que nous avons développée pour répondre à des problématiques de classification de textes. L'objectif de notre méthode est de détecter les termes les moins discriminants et de les remplacer par des descripteurs plus génériques de façon à obtenir une meilleure représentation en vue de son utilisation par un algorithme de classification automatique. Nous utilisons une formule statistique que nous appelons D pour privilégier les termes les plus discriminants :

$$D(x) = \frac{\text{occClasse(x)}}{\text{occCorpus(x)}}$$

où occClasse(x) est le nombre d'occurrences de x dans la classe qui le contient le plus et occCorpus(x) représente le nombre d'occurrences de x dans l'intégralité du corpus.

La mesure D s'est révélée pertinente pour des tâches de classification (Tisserant et al., 2014). Nous allons montrer comment cette mesure, en détectant les termes les plus discriminants des tweets, peut être utilisée pour une tâche de génération de HashTags. Les *classes* considérées seront les différents mouvement politiques.

3.1.3 *PolitIt*

PolitIt est un GWAP qui traite de la "polarité politique" des termes. Dans ce jeu, une interface propose des termes aux joueurs qu'ils doivent associer à un des six courants politiques proposés : *Extrême-gauche*, *Ecologie*, *Gauche modérée*, *Droite modérée*, *Droite*, *Extrême droite*. Lorsqu'ils considèrent qu'aucun courant ne correspond mieux que les autres, ils peuvent passer à un autre terme sans donner de réponse. Les données obtenues grâce aux parties jouées permettent de rattacher des termes aux centres d'intérêts des courants politiques. Par exemple, l'entité nommée *Adam Smith* est associée par les joueurs à la *droite*. Cela ne veut pas forcément dire que le descripteur représente un concept défendu par le courant politique rattaché. Par exemple, le terme *nucléaire* est rattaché au courant politique *Ecologie*. Cela s'explique par le fait que, bien qu'ils s'y opposent, le nucléaire est un sujet de préoccupation important pour les mouvements écologistes. Ces données peuvent donc être adaptées pour détecter les termes clivants dans notre corpus de tweets politiques.

3.2 Les termes sélectionnés par chaque mesure

La Table 1 montre un échantillon des termes appartenant à une des catégories. Les termes de *PolitIt* présentés sont ceux ayant le plus grand nombre de réponses attachant le terme au même courant politique. Les HashTags mis en avant sont les plus utilisés dans le corpus. Les ter-

mes en provenance de *GenDesc* sont ceux ayant la valeur de la mesure D la plus élevée. Une version plus complète de ce tableau est donnée dans (Tisserant, 2015).

PolitIt	HashTag	*GenDesc*
centriste	air	puteaux
dieu	éducation	metz
assurance	marseille	mélenchon
elf	jdd	besançon
bourse	karachi	front
rpr	interview	laurent
rtt	optimisme	edf
ss	hollande	nazaire
terre	crise	nucléaires
altermondialiste	sénat	démanteler

Table 1: Mots n'apparaissant que dans un seul des trois ensembles

Nous voyons qu'une partie des termes retournés par *GenDesc* semblent les plus difficiles à rattacher à un mouvement politique. Le terme *démanteler* ou le prénom *Laurent*, par exemple, sont impossibles à analyser sans leur contexte. Toutefois, en prenant en compte le contexte, certains de ces termes deviennent politiquement orientés. *Metz*, par exemple, fait référence à un meeting qui a eu lieu pendant la campagne législative de 2012, au moment où le corpus de tweets a été constitué.

Une partie des HashTags nécessite aussi un contexte pour les rattacher à un mouvement politique. Toutefois, même privés de leur contexte, nous pouvons considérer qu'ils représentent des concepts politiques. Par exemple, si nous prenons les termes *éducation* ou *sénat*, nous pouvons considérer qu'ils appartiennent au champ lexical de la politique, même s'ils ne peuvent être rattachés à un mouvement politique spécifique.

Les termes de *PolitIt* peuvent en grande partie être rattachés à un courant politique, même hors de tout contexte. Certains termes comme *RPR* ou *altermondialiste* font même directement référence à des courants politiques. Toutefois, il ne faut pas forcément en déduire que ces termes vont être utilisés par les courants auxquels ils sont reliés. Par exemple, le terme *SS* est rattaché à l'*extrême droite* dans *PolitIt*, mais il est principalement utilisé sur Twitter par des sympathisants d'*extrême gauche* pour parler du *Front National* de manière péjorative.

4 Contribution

Nous allons dans cette Section présenter deux méthodes de génération de HashTags. Chaque méthode sera évaluée quantitativement et qualitativement sur un corpus de tweets politiques.

Nous proposons dans la Section 4.1 une méthode pour sélectionner, à partir de tweets, des termes candidats pour être des HashTags. Puis, dans la Section 4.2, nous proposerons une méthode permettant de générer des HashTags composés de plusieurs mots.

4.1 Génération de HashTags simples

Nous avons vu précédemment que les termes provenant à la fois de *GenDesc* et *PolitIt* sont des HashTags potentiels intéressants. Nous allons nous appuyer sur ce constat pour proposer une première méthode de génération de HashTags.

4.1.1 Méthodologie

L'idée est d'exploiter les termes détectés par *GenDesc* et ceux provenant de *PolitIt* pour générer des HashTags. Les termes provenant de *GenDesc* sont discriminants pour les tweets. Cela indique qu'ils peuvent donner des HashTags intéressants pour marquer politiquement un tweet. Les termes provenant de *PolitIt*

sont des termes ayant une sémantique politique forte. À ce titre, ils ont de fortes chances d'être un marqueur d'attachement à un mouvement politique.

Pour vérifier notre hypothèse, nous avons mené l'expérimentation suivante : Nous avons sélectionné 25 termes pour chaque catégorie (*GenDesc*, *PolitIt* et *GenDesc* ∩ *PolitIt*). Les termes de *GenDesc* étant ceux avec la mesure D la plus importante, et les termes de *PolitIt* sont ceux ayant été rattachés le plus grand nombre de fois au même courant politique. Pour l'intersection de *GenDesc* et *PolitIt*, nous avons pris les termes ayant la mesure D la plus importante qui sont rattachés à plus de 50 % à un même courant politique dans *PolitIt*.

4.1.2 Résultats

Pour vérifier si les HashTags générés pouvaient s'avérer pertinents, nous avons observé, grâce au site hashtags.org[1], s'ils sont aujourd'hui utilisés comme HashTag. Les mesures ont été effectuées en Juillet 2014, soit plus de deux ans après la construction du corpus. Les HashTags sont considérés comme utilisés régulièrement s'ils ont été utilisés plus de cent fois en moyenne par jour. Les résultats de l'expérimentation sont donnés en Table 2.

GenDesc	*PolitIt*	*GenDesc* ∩ *PolitIt*
52 %	76 %	92 %

Table 2: Tag utilisés fréquemment

4.1.3 Analyse

Nous remarquons qu'à peine plus de 50 % des termes avec une valeur de la mesure D élevée sont utilisés comme des HashTags, alors que ceux provenant de *PolitIt* sont à 76 % utilisés comme HashTags. Cela permet de montrer que l'information sémantique issue de GWAP est plus pertinente que l'information statistique pour la génération de HashTags. Toutefois, nous voyons que l'information statistique peut être pertinente pour la génération de HashTags. En effet, en prenant l'intersection de *GenDesc* et *PolitIt*, nous obtenons un meilleur résultat (92 %) qu'en utilisant les termes de *PolitIt* (76 %).

Le fait que la combinaison des deux méthodes soit plus efficace que l'utilisation des méthodes isolées vient du fait que *GenDesc* et *PolitIt* apportent des informations différentes et complémentaires :

- *GenDesc* nous permet de savoir qu'un terme est marqueur d'une classe politique,
- son apparition dans *PolitIt* montre qu'il appartient au champ lexical de la politique.

4.2 Génération de HashTags composés

Nous avons vu dans la section précédente que nous pouvions générer des HashTags pertinents composés d'un unique terme en combinant des informations statistiques et sémantiques. Mais une partie importante des HashTags sont en réalité composés de plusieurs mots.

4.2.1 Méthodologie

La problématique de génération de HashTags composés de plusieurs mots est plus complexe que celle de génération de HashTags simples. En effet, il faut pouvoir proposer des combinaisons de termes représentant des hashtags potentiels, et ensuite sélectionner ceux dont la combinaison offre une sémantique intéressante permettant d'identifier rapidement la thématique du tweet.

Nous avons décidé de nous appuyer sur l'utilisation de patrons syntaxiques pour effectuer une sélection de termes candidats. Puis,

[1] www.hashtags.org

pour sélectionner les termes les plus pertinents pour être utilisés en tant que HashTags, nous avons utilisé les informations provenant de *PolitIt* et *GenDesc*.

Étape 1 : Patrons syntaxiques

La première étape de la génération consiste à choisir un ensemble de syntagmes candidats. Nous avons choisi d'utiliser des patrons syntaxiques pour sélectionner des syntagmes candidats. Cette approche est proche de celle adoptée par certaines méthodes d'extraction de terminologie (Aussenac-Gilles et al., 2000). Nous avons recouru à trois patrons syntaxiques classiques (Daille, 1994), présentés dans la Table 3.

Patron syntaxique	Exemple
NOM - ADJECTIF	listes électorales transition énergétique
ADJECTIF - NOM	haute surveillance affreux dictateur
NOM - PREPOSITION -NOM	syndicalisme de lutte gaz de schiste

Table 3: Patrons syntaxiques et exemples de termes associés présents dans le corpus.

Pour détecter les patrons syntaxiques dans les tweets, nous avons choisi d'utiliser l'étiqueteur grammatical SYGFRAN (Chauché, 1984).

Étape 2 : Filtre statistique endogène

Nous avons ensuite appliqué un filtre statistique endogène sur nos candidats, pour ne conserver que les syntagmes dont au moins un était considéré comme pertinent par *GenDesc*. Ce filtre permet de supprimer les HashTags n'étant pas considérés comme discriminants par notre mesure statistique. Ainsi, des syntagmes comme "*journal de campagne*" ou "*texte à trous*" vont être supprimés de la liste des HashTags candidats.

Étape 3 : Filtre sémantique

Nous avons ensuite appliqué un filtre sémantique, pour ne garder que les couples contenant des termes appartenant à *PolitIt*. Ce filtre nous permet de ne conserver que des syntagmes représentant des concepts politiques. Ainsi, certains syntagmes candidats comme "*fdg créé*" ou "*projet irresponsable*" vont être écartés de la liste des HashTags candidats.

Étape 4 : Filtre statistique exogène

Un certain nombre de HashTags générés à partir des patrons syntaxiques ne représentent pas de concepts. Le nombre important de ce type de HashTags s'explique, en partie, par la mauvaise construction grammaticale des tweets, qui a tendance à induire en erreur l'analyseur syntaxique. Nous avons donc utilisé un filtre statistique exogène pour détecter la pertinence de l'association de termes. L'idée est de nous appuyer sur un corpus différent qui est à la fois indépendant et de taille supérieure pour y mesurer la fréquence d'apparition des syntagmes sélectionnés. Ce filtre a pour but de supprimer des syntagmes ne représentant pas forcément un concept, comme "*consommation collaborative*" ou "*petitjournal politesse*". Nous avons choisi d'utiliser internet comme corpus pour ce filtre. Nous avons mesuré la fréquence d'apparition des syntagmes grâce au moteur de recherche Bing, en considérant le nombre de résultats retournés (Turney, 2001). Ce filtre nous a permis de supprimer des groupes de termes apparaissant rarement ensemble sur Internet, et qui ne représentent pas forcément une sémantique intéressante.

4.2.2 Résultats

Après application du processus, nous avons relevé qu'un grand nombre de HashTags générés étaient en rapport direct avec des évènements survenus à la période où le corpus

a été constitué. Nous avons choisi d'évaluer chaque HashTag en prenant en compte deux types d'informations sémantiques :

- la représentation d'un **concept** du champ lexical politique.
- le marquage d'une **orientation** politique.

Des exemples de HashTags illustrant ces notions sont donnés dans la Table 4.

Concept politique	*#voteutile* *#pouvoirdachat*
Orientation politique	*#buffetsurcanalplus* *#gaucheàbastia*
Concept politique et **Orientation** politique	*#drapeaurouge* *#alliancecentriste* *#agriculturepaysanne*
HashTag apolitiques	*#nouvellechanson* *#texteàtrous*

Table 4: Exemple de HashTags générés et classés en fonction de leur sémantique politique

Nous avons évalué notre méthode de génération de HashTags composés. Pour cela, nous avons annoté manuellement 40 HashTags pour chaque type de filtre utilisé (*GenDesc*, *PolitIt*, *GenDesc* ∩ *PolitIt*, *GenDesc* ∩ *PolitIt* ∩ Web). Les résultats sont donnés dans la Table 5.

4.2.3 Analyse

La Table 5 nous montre que le filtre fondé sur *GenDesc* est efficace pour écarter les HashTags générés n'ayant pas d'*orientation politique*. Au contraire, le filtre fondé sur *PolitIt*, se montre plus efficace pour supprimer les HashTags ne représentant pas un concept politique. L'utilisation des deux filtres combinés permet d'obtenir un pourcentage de HashTags représentant un concept politique supérieur à

	Sémantique Politique	Orientation Politique	Sémantique Politique ∩ Orientation Politique
Aucun	27.5 %	5 %	5 %
GenDesc	42.5 %	50 %	27.5 %
PolitIt	55 %	22.5 %	17.5 %
GenDesc ∩ *PolitIt*	62.5 %	50 %	32.5 %
GenDesc ∩ *PolitIt* ∩ Web	80 %	52.5 %	47.5 %

Table 5: Pourcentage de HashTags générés représentant un concept politique ou une orientation politique en fonction des filtres utilisés. La ligne *Aucun* correspond à l'ensemble des HashTags sélectionnés grâce aux *patrons syntaxiques*.

n'importe lequel des deux filtres utilisés seul. Le recours au filtre utilisant *Bing* combiné aux deux autres filtres permet d'améliorer encore la qualité des HashTags générés. **La combinaison des trois filtres correspondant à l'application du processus dans sa globalité permet de générer des HashTags dont 80 % sont porteurs d'une sémantique politique et 47.5 % sont à la fois porteurs d'une sémantique politique et marqueur d'une orientation politique.**

Le fait que 80 % des HashTags générés avec l'utilisation des trois filtres représentent un concept politique indique que l'algorithme ne propose que 20 % de HashTags réellement non pertinents. Par ailleurs, plus de la moitié des HashTags générés sont porteurs d'une *orientation politique*. La combinaison de filtres que nous proposons nous permet donc de générer des HashTags pertinents, à la fois porteurs de sens et d'une orientation politique.

5 Conclusion

Nous avons décrit dans cet article deux méthodes de génération de HashTags. Nous avons vu que les données statistiques comme les données sémantiques permettaient de développer des méthodes de génération de HashTags. Nous avons développé une méthode mêlant informations sémantiques, informations syntaxiques, et approches statistiques s'appuyant sur des données endogènes et exogènes. Nous avons montré que la combinaison de ces méthodes permet d'obtenir de meilleurs résultats que chacune des méthodes utilisée séparément.

Toutefois, nos travaux sur la génération de HashTags contiennent plusieurs limites. Nous nous sommes placés dans un contexte où nous avions connaissance des thématiques des tweets. Nous pensons étendre notre méthode à un cadre non supervisé, sans connaissances *a priori* des thématiques abordées dans le corpus ni des différentes opinions exprimées.

Une autre limite importante de nos travaux vient du faible nombre de patrons syntaxiques utilisés pour la sélection de HashTags candidats. Or de nombreux HashTags ont des structures complexes, non représentées par ces patrons. Nos futurs travaux s'appuieront sur les syntagmes verbaux et l'association de termes ou de n-grammes de caractères extraits à partir des tweets.

References

Aussenac-Gilles, N., Biébow, B., Szulman, S., et al. (2000). Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In *Actes Ingénierie des Connaissances (IC)*, pages 93–104.

Bouillot, F., Poncelet, P., Roche, M., Ienco, D., Bigdeli, E., and Matwin, S. (2012). French presidential elections: what are the most efficient measures for tweets? In *Proc. of the Workshop on Politics, elections and data*, pages 23–30. ACM.

Chauché, J. (1984). Un outil multidimensionnel de l'analyse du discours. In *Proc. of Int. Conf. on Computational Linguistics*, pages 11–15.

Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the political alignment of twitter users. In *Proc. of Conference on Social Computing (SocialCom)*.

Daille, B. (1994). *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. PhD thesis.

Godin, F., Slavkovikj, V., De Neve, W., Schrauwen, B., and Van de Walle, R. (2013). Using topic models for twitter hashtag recommendation. In *Proc. of Int. conference on World Wide Web companion*, pages 593–596.

Huang, J., Thornton, K. M., and Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proc. of Conference on Hypertext and hypermedia*, pages 173–178. ACM.

Kywe, S. M., Hoang, T.-A., Lim, E.-P., and Zhu, F. (2012). On recommending hashtags in twitter networks. In *Proc. of Int. Conference on Social Informatics*, SocInfo, pages 337–350.

Ozdikis, O., Senkul, P., and Oguztuzun, H. (2012). Semantic expansion of hashtags for enhanced event detection in twitter. In *Proc. of International Workshop on Online Social Systems*.

Tisserant, G. (2015). *Généralisation de données textuelles adaptée à la classification automatique*. PhD thesis, Univ. Montpellier.

Tisserant, G., Prince, V., and Roche, M. (2014). Gendesc: Vers une nouvelle représentation des données textuelles. *Num. sp. "Fouille de Données Complexes", RNTI*, pages 127–146.

Turney, P. (2001). Mining the web for synonyms: Pmi-ir versus LSA on TOEFL. In *Proc. of European Conference on Machine Learning*, pages 491–502.

Zangerle, E., Gassler, W., and Specht, G. (2011). Recommending#-tags in twitter. In *Proc. of Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR*, volume 730, pages 67–78.

Novel Metaphor and Scientific Discourse Come to Terms: A Case Study of Metaphorical Prototerms in Biology

José Manuel Ureña Gómez-Moreno
University of Castile-La Mancha / Avd. Camilo José Cela, sn/ 13071, Ciudad Real, Spain
Josemanuel.urena@uclm.es

Abstract

Novel metaphorical expressions have been understudied in traditional approaches to terminology because they normally behave as sporadic units incapable of structuring whole discourse events. To show that this is not always the case, this paper presents a case study of novel BIOECONOMICS metaphors in an academic marine biology research article (Landa 1998). They were analysed following two paradigms: (i) the Career of Metaphor Theory (Bowdle and Gentner 2005), a solid framework for the description of novel metaphor in usage; and (ii) the text-linguistics approach to term description (Collet 2004), which suggests a set of criteria for term definition that challenges the prescriptive tenets of monolithic terminology models. The analysis of unexpected metaphors and similes identified in the text suggests that these units should be regarded as proto-terms experienced as deliberate rhetorical *and* conceptual devices. On a pragmatic level, the metaphors are shown to be part and parcel of the writer's discursive strategy to communicate specialised knowledge to her peers and further science. On a conceptual level, the metaphors are found to be essential building blocks and structuring elements of the mental model of the article.

1 Introduction

The Career of Metaphor Theory (CaMT)[1] (e.g. Bowdle Gentner, 2005) is one of the most representative models of metaphor description within the cognitive linguistics strand (cf. e.g. Steen 2007 for a detailed account of the rest of the models). Bowdle and Gentner examine metaphor in usage, looking at both spontaneous and conventionalised instances of metaphoric use in context. This proposal thus provides potential ground for a discourse-led metaphor analysis, becoming a valid framework for this study.

Bowdle and Gentner claim that novel metaphors are processed by comparison, i.e. alignments between target and base concepts; in contrast, conventionalised metaphors are processed by categorisation, where comprehension of the metaphor "requires that one use the base concept to elicit a metaphoric category that it typifies" (Bowdle and Gentner, 2005: 194). CaMT further posits that whether metaphors are processed directly (i.e. as stable metaphoric categories) or indirectly (as comparisons) will depend both on their degree of conventionality and on their linguistic form (Bowdle and Gentner, 2005: 193). This study focuses on the behaviour of novel metaphors in a marine biology research article. Many novel metaphorical expressions in this article are thus suggested not to be processed directly as well-entrenched categories with stable linguistic forms, but indirectly as innovative *comparisons* formalised as unconventional linguistic pairings. Being indirectly comprehended, these comparisons involve a complex sequential process whereby the intended metaphoric meaning is derived by the expert reader only when the pre-existing literal (or conventionalised figurative) meaning of the base term cannot be sensibly applied in the biology discourse.

This assumption is reinforced by the evident rhetorical function of each of the linguistic forms instantiating the novel metaphors brought by the writer into the marine biology article, not inviting but rather making the reader constantly map the source domain onto the target domain for

[1] The most reasonable acronym to use here would be CMT. However, this could be confusing for metaphor scholars in that Conceptual Metaphor Theory is often abbreviated as CMT.

specialised knowledge comprehension. The analysis of these linguistic forms is relevant to CaMT insofar as this model has traditionally left aside the role of the rhetorical form in which a metaphor is expressed (Steen 2007: 78).

Another reason to use CaMT in this study is that there is authoritative research within or in consonance with this theory that claims for the existence of deliberate metaphor (e.g. Steen, 2009; Krennmayr, Bowdle, Mulder and Steen, 2014). This research provides evidence that sometimes language users pay *attention* to their use of metaphor for making cross-domain comparisons (Steen, 2009: 180). As Steen (ibid.) goes on to note, this normally takes place in the deliberate metaphorical design of texts and discourse units. As will be shown, this is precisely the scenario that is set up in the marine biology research paper analysed in this study.

Applying CaMT to the analysis of figurative expressions in expert-to-expert scientific communication should also help demonstrate that the stability preserved by canonical metaphorical terms — i.e. widely acknowledged instantiations of fixed linguistic regularities carrying specialised meaning — may be positively altered by the introduction of novel metaphors capable of conceptually articulating a domain-specific text/discourse event. The novel metaphors examined in this study are thus evidence of the ignored eclectic nature of specialised discourse, which can also produce highly creative metaphorical expressions that critically assist in introducing innovative knowledge and illustrating scientific findings for theory construction.

In CaMT, variation of linguistic form also involves similes. According to the *principle of grammatical concordance*, similes, which are grammatically identical to literal expressions of comparison, should invite explicit (albeit metaphoric) comparisons between target and base domains (Krennmayr et al., 2014: 70). Krennmayr et al. (2014) also suggest that the signalling effect of similes helps integrate a metaphoric frame into people's mental representation of a text. The research article analysed in this study contains a number of similes that are shown to form a part of the writer's argumentation strategy to present and describe specialised knowledge to the specialist readership. These similes are thus useful because of their explicit interpretative guidance, anticipating the analogy to the expert reader by effectively establishing and explicitly signalling metaphoric comparisons between two concepts. As with novel metaphors, certain similes used in the marine biology domain are also expected to aid integration of a metaphoric schema into experts' mental text representation and comprehension. Bowdle and Gentner (2005: 211) note, novel similes and metaphors involve novel base terms that refer only to *domain-specific* concepts [my emphasis].

Collet's (2004) approach to term description combines text-linguistic and Language for Specific Purposes assumptions. This proposal draws from earlier theory on context-oriented terminology (e.g. Bourigault and Slodzian, 1999), and departs from prescriptive paradigms — especially, Wüster's (1979) General Theory of Terminology, which argues for monolithic, decontextualised specialised meaning description. Collet's model is interesting because it suggests a new definition of term, based on a set of requirements that a lexical unit needs to meet to be considered a terminological unit. This analytic method is instrumental to the characterisation of the metaphorical units examined in the present study.

The first requirement involves placing the focus of analysis beyond the level of the sentence, considering the text the best-suitable instrument for term definition and description (Collet 2004: 103). Secondly, a subject-oriented text is regarded as the product of a communicative act or event where the lexical items used by an expert take on particular and specific meaning to produce and communicate specialised knowledge. Thirdly, for the sake of text *coherence*, a writer adjusts the meaning content of a term that he uses to his understanding of the realities that it refers to (Collect, 2004: 109). This way, terms help the writer achieve texture. Fourthly, to achieve *cohesion*, terms tend to vary their linear structures in specialised language texts, exhibiting a range of different lexical-syntactic configurations that make up a *paradigm*. A paradigm is a closed set composed of the full-length of the term and all of its alternate shorter forms (Collet 2004: 108).

Albeit novel and unconventional to the knowledge field of biology, the figurative expressions extracted from the research article examined in this paper (see section 4) are shown to be *semantically charged linear structures* (Collect, *ibid.*), which designate abstract or concrete realities studied in the special-subject text where these metaphorical expressions occur. Thus, we can speak of specific entity-word pairings/correlates that are exclusively created and

activated through metaphor in a concrete *text* belonging to a particular specialised knowledge domain. As will be shown, a good number of newly created metaphors retrieved from the research study examined are realised in a variety of alternate linguistic forms that make up paradigms or full linear structures within the texts where they occur. Like well-entrenched terms, these arrays of alternate metaphorical forms build *coreferential chains* (Collet, 2004), which typically contribute to text cohesion in domain-specific writing. This phenomenon has also been documented by other authors advocating for a text-linguistic approach to terminology, such as Rogers (2007), who speaks of *lexical chains*.

For all reasons given above, the novel biology metaphors analysed in this research can be regarded as *proto*-terms, i.e. lexical items that exhibit the same conceptual and lexical-syntactic features as terms but still need to be systematically used across specialised research articles and books.

The usefulness of similes from a text-linguistic approach in terminology studies resides in the intrinsic conceptual and linguistic characteristics of these figures of thought. As earlier explained, a text-linguistic analysis of specialised meaning goes beyond the scope of the sentence to focus on longer contextualised linguistic units, which effectively contribute to text construction as cohesion- and coherence-producing agents (Collet, 2004). Linking this claim to CaMT's consideration of similes as instruments aiding integration of metaphoric schemas in text representation and understanding, it can be argued that the use of novel metaphor base structures longer than single words and simple phrases easily reflect the conceptual processes of scientific communication. Concretely, similes participate in the description of expert knowledge where newly created comparisons are established. In doing so, similes show their conceptual potential as rhetorical devices, capturing the mind's eye, and creating highly imagistic representations of *entire* research articles.

2 Case study

2.1 Data

The analysis of novel bioeconomics metaphors conducted in this paper reports on empirical data extracted from economist Landa's (1998) research study, which was published in the academic journal *Environmental Biology of Fishes* (see full details in References). The analysis illustrates how a scholar manages to deliberately exploit a set of innovative metaphors with a view to conceptually and linguistically scaffolding her train of thought throughout a specific scientific discourse event. Fine textual analysis of a single text through bulk data retrieval has already been successfully performed by previous terminology studies (cf. Pecman 2014 for terminological variation and cognition).

Despite concentrating on novel metaphors found in one single text, the present research also draws upon a compilation of marine biology articles published in high-impact academic journals in order to test the authentic novelty of creative metaphor candidates against such articles. The dataset consists of 1,938,472 tokens/words. Table 1 includes the name of the journals as well as numerical information about them. The corpus articles were searched with the search option of *Wordsmith Tools®* — a lexical analysis software programme — for novel metaphor candidates. If hits of these candidates were obtained, they were then not classified as novel metaphors.

Journals	Number of Articles	Number of Tokens
Marine Biology	32	286,736
Environmental Biology of Fishes	31	266,065
Phycologia	30	244,758
Hydrobiologia	30	235,477
Journal of Experimental Marine Biology and Ecology	22	172,441
Journal of Fish Biology	18	126,570
Fish Physiology and Biochemistry	17	108,960
Ecotoxicology	17	106,929
Coral Reefs	16	102,830
Symbiosis	15	119,346
Biosemiotics	13	108,041
NATO Advanced Study Institutes Series	8	60,319
	Total: 249	Total: 1,938,472

Table 1: Academic journal articles in the corpus.

2.2 Data processing

Identification of novel metaphor and simile candidates

As shown by contexts (1) and (2), newly created metaphors and similes normally appear between inverted commas or in italics in the biology text corpus. This was attested by browsing the corpus through the search option of *Wordsmith Tools®*.

(1) The choroid *rete mirabile* is a large **horseshoe-shaped, gland-like** structure located around the optic nerve in the choroid layer of the eye of many species of fishes (*Nato Advanced Study Institutes Series* 1, 1975).

(2) In recent years the fruit fly, *Drosophila melanogaster* has become the "come-back kid" in biology, though some might question wheher research on this model animal ever peaked. The initial interest in the fruit fly goes to the days of Thomas Hunt Morgan and his infamous "fly room", the **ground zero** of the genetics movement. Today the focus on the "**black box**" (*Biosemiotics* 2009, 2:181-191)

Searching a corpus for inverted commas is normally a strategy of great use not only in general language studies of metaphor, but also in the analysis of newspaper articles and popular science publications (Goatly, 2002: 73). Even though their efficiency in specialised discourse had hardly been tested, inverted commas and italics were found to frequently act as visual direct and indirect metaphor markers to the specialist readers, signalling the unexpected surprise effect that spontaneous metaphors cause in them for not being conventionalised units in the field. The conclusion drawn is that scholars writing their research articles are generally aware that novel metaphorical units should be marked somehow to indicate that they are uncommon expressions, alien to the biology discourse. Based on this, the first strategy devised to identify novel metaphor candidates was to search the entire corpus for inverted commas, and next, focus the search on Landa's article. The markers "" and '' turned out to be extremely productive cues for potentially novel metaphors. Words in italics were also examined, also pointing to a number of linguistic candidates.

Testing and analysing novel metaphor candidates

Linguistically speaking, a metaphorical expression requires the identification of some kind of semantic tension or incongruity between its basic sense and that sense activated in a particular communicative situation. A valid strategy to find instances of metaphoric usage from a discourse-led perspective (cf. e.g. Cameron, 2007: 118) relies on two criteria: (i) the presence of a lexical item (the vehicle or base) that has a meaning that can be said to contrast with its meaning in the discourse context; (ii) the potential for extra meaning to be produced as a result of bringing together the vehicle's standard or de-contextualised meaning and its meaning activated in a specific discourse event. Based on these criteria, most of the lexical items that appear between inverted commas or in italics in Landa's article were found to have a metaphorical meaning.

In many passages of her article Landa creates totally innovative metaphors in the form of quasi-terminological items or proto-terms, exhibiting a linguistic arrangement that differs from that of well-entrenched terminological metaphors in bioeconomics — a well-entrenched field of inquiry that includes a consirable number of well-established terminological metaphors (see below). For instance, she sets a comparison between *vote-with-the-feet* (base concept) and *vote-with-the-fins* (target concept), the latter being a new co-occurrence of lexical items. Even with single-word metaphorical expressions from the source domain, such as *club*, Landa comes up with alternate linguistic forms to figuratively refer to marine biology entities and phenomena. For instance, she suggests the novel metaphor phrases *informal club* and *multi-product club* to describe fish schools with particular behavioural patterns.

Those metaphorical expressions in the analysed article that preserve the linguistic forms displayed in the base domain are also considered to be novel because they still entail newly created comparisons between two (not three) concepts in alignment (not categorisation). At the conceptual level, this can be explained by recourse to *extended mappings* between the base and target domains, a notion suggested by Bowdle and Gentner (2005: 212) in CaMT (see continuous arrows in Figure 1 and explanation below). As these authors put it, to the extent that concepts are often understood at least partly in terms of relations to other concepts within a particular knowledge domain, metaphoric mappings can be

expected to extend beyond the named target and base concepts to more global conceptual systems. As the text corpus demonstrates, economics is a field of expertise that has historically and systematically been exploited by biologists, giving rise to a large number of highly recurrent, well-established terminological metaphors, such as *capitalise on, economise on*, and *energy cost*, in the biology domain. For this reason, we can speak of a subfield known as *bioeconomics*. Conceptually speaking, and following CaMT (Bowdle and Gentner 2005: 209), the conventionalised mappings operating between pairs of concepts from the base (ECONOMICS) to the target (BIOLOGY) domain have ultimately been overridden by single mappings, which exclusively direct the target concepts from superordinate concepts that make up upper-level metaphoric categories. For this reason, these metaphors are argued to be processed directly. Figure 1 illustrates this re-arrangement in a concept-structure schema.

As Landa (1998: 355) explains, the theory of clubs and the theory of public goods form a part of the public choice theory within ECONOMICS, containing conventional metaphorical terms, such as *free rider* (somebody who gives up on an established economic paradigm to live by their own principles), *Pareto-optimal* (referring to a financial situation where one person is made better off and no one is made worse off), and *club good* (benefit obtained when belonging to a particular economic force). Figure 1 shows how these ECONOMICS concepts are projected onto the BIOECONOMICS domain, prompting *novel* cross-domain mappings that involve horizontal alignments of *two* (not three) concepts (see contexts in *Analysis and Discussion of Empirical Data*, which provide textual evidence of these systematic comparisons). For this reason, these metaphors are argued to be processed *in*directly. As a result of these metaphoric mappings, *free rider*, *Pareto-optimal*, and *club good* are made to designate domain-specific entity-word correlates carrying particular specialised meanings in a concrete communicative act within the marine biology discourse.

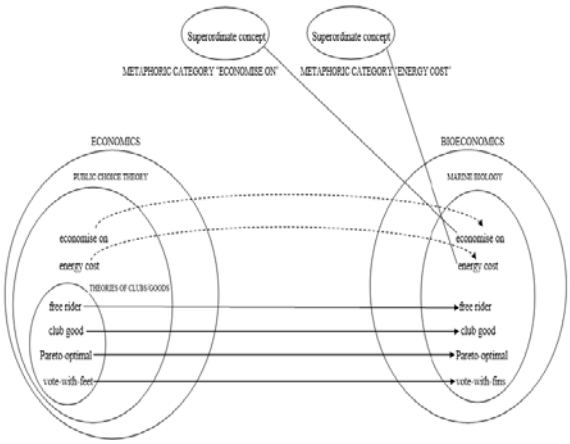


Figure 1: Overridden conventional (discontinuous arrows) and novel (continuous arrows) cross-domain mappings between ECONOMICS and BIOECONOMICS in Landa's article. Conventional mappings give way to metaphoric categories.

These newly created metaphors arise from extended (novel) mappings between the economics base domain PUBLIC CHOICE THEORY (concretely, the subdomains THEORIES OF CLUBS AND PUBLIC GOODS), which includes conventionalised metaphors, and the target subdomain MARINE BIOLOGY. Integrating the rest of innovative comparisons (the most representative ones are discussed in section 4) into this target subdomain resulted in an entire network of interrelated novel metaphors that conceptually and linguistically vertebrate Landa's discourse.

3 Results and Discussion

Identifying metaphor candidates, testing them, tracking metaphors across corpus texts and grouping them are the four core processes in every method of metaphor-led discourse analysis. Having gone through all four stages, the next step in this study was to analyse and describe the novel metaphors and similes recruited as contextualised linguistic units in Landa's article. This was done with a view to providing empirical evidence of the following claims:

(i) the novel marine biology metaphors examined are deliberately and systematically exploited by Landa with an evident rhetorical purpose, which is to bring the expert readers' attention to the contents of the article, overtly encouraging them to constantly map the ECONOMICS base domain onto the

BIOECONOMICS target domain for integrated specialised knowledge comprehension;

(ii) in connection with the last idea in (i), the role of the novel metaphors and similes is also to conceptually and linguistically articulate the author's paper to communicate scientific knowledge;

(iii) the novel metaphors conform to Collet's (2004) criteria, which allow these metaphors to be considered as proto-terms or terminological units in the making.

As explained earlier, in her article Landa deliberately makes sustained comparisons between theories of clubs and public goods in economics, which are conventionalised conceptual metaphor themselves, and fish schooling behaviour. Such comparisons instantiated as novel metaphors and similes in the text. The first piece of empirical evidence showing that they really are recurrent creative metaphors that systematically occur throughout the entire article is their high frequency of appearance. Table 2 includes the number of linguistic occurrences of the novel BIOECONOMICS metaphors identified in the article[2]. These metaphors are presented as lexemes and as the diverse linguistic forms (tokens) that they take on in the text.

Lexemes	Tokens (number of occurrences)
Free rider	free rider(s) (8); free riding (2); free-riding (2); free ride (4); free rides (1); quasi-free rider(s) (5); quasi-free riding (1)
Club	club (42); exclusive club (4); self-enforcing exclusive club (3); informal club (2); multi-product club (3); size club (2); mixed club (1); non-discriminatory club (2)
Club good	club good(s) (24)
Club member	club member(s) (23)
Selfish fish	selfish fish (28)
Invisible fin	invisible fin (1); invisible fin process (1)
To vote with fins	vote-with-their-fins (3); voting-with-its-fins (1)
To lift a finger	lifting a fin(ger) (1)
Pisces economicus	Pisces economicus (1)
Pareto-optimal	Pareto-optimal (2); Pareto-optimality (3)
Inspector	inspectors (2); inspecting behaviour (1)
Caste	caste of guard fish (1); caste such as guard fish (1)
Congestion/crowding	congestion/crowding (1)

Table 2. Lexemes and tokens of novel BIOECONOMICS metaphors in Landa's article.

Context (3) is the first example containing textual evidence of the sustained comparisons between ECONOMICS and MARINE BIOLOGY in Landa's article. Conventional metaphors, which were also found in many other corpus texts, are shown in small capitals. The novel and highly unconventional metaphors are shown in bold.

(3) The prevalence of fish schools seem to point to the schooling fish as a 'SOCIAL FISH'. But is SCHOOLER a 'social fish' or is it really a '**selfish fish**', or what Boulier & Goldfarb (1991) call ***Pisces Economicus*****,** the counterpart to ***Homo economicus***, the selfish, calculating or rational economic man of economic literature?

Context (3) includes the novel marine biology metaphor *selfish fish*[3] and the highly

[2] This table does not include novel similes because they involve relatively long stretches of text, not just single or multiword lexical units, as is the case for novel metaphors. Examples of similes are discussed below in this section.

[3] The meaning of *selfish* is understood here from an anthropocentric perspective, thus involving conscious (vs. mechanistic/instinctive) intersubjective aspects, such as desire, beliefs, attentional foci, and intentions. In this sense, *selfish* does have a figurative meaning in Landa's text since only humans clearly have all of these cognitive and psychological capacities and states (cf. Zlatev, Racine, Sinha and Itkonen 2008). Landa herself makes the distinction clear between humans and fish in terms of cognitive capacities by

unconventional expression *Pisces economicus*, which arise by comparison with two terminological units that have domain-specific meanings in economics: *selfish economic man* and *Homo economicus* (cf. Boulier and Goldfarb, 1991). Landa draws on these two expressions to bring up creative metaphors in biology that enable her to raise questions about the actual behaviour of schooling fish (a well-entrenched terminological metaphor, as shown by the text corpus). In doing so, Landa is constantly establishing inter-textual references between her arguments in marine biology and economic theories (cf. also Buchanan and Tullock's 1962 club theory below). This is thus an evident example of the additional development of the BIOECONOMICS metaphorical frame (by means of extended mappings) through inter-textuality. By combining novel and conventional BIOECONOMICS metaphors, Landa substantially enriches her specialised discourse event, providing an entire expert community with specific explanations about specialised biology concepts. The innovative nature of the novel metaphors that she introduces clearly serves as a powerful rhetorical tool to illustrate and define concepts, and eventually, construct and further science in an attractive manner.

The rhetoric of novel figurative expressions to conceptually structure a whole scientific discourse event also manifests in the form of even more explicit comparisons, such as the simile described in context (4).

(4) Visual cues and odors of conspecifics provide LOW COST signals for fish of the same species to identify each other, just as ethnicity serves as low cost non-price signal in Landa's (1981) theory of the EHMG.

Landa (1998: 359) aligns the concept *ethnically homogeneous middle-man group (EHMG)*, proposed by herself in economics, with *fish school* with the help of the simile signal *just as*. *EHMG* refers to the idea that middle-men prefer choosing trading partners who are members of their own kinship or ethnic group (e.g. Chinese or Jewish). The clear interplay between the conventional (terminological) metaphor *low cost* and the simile *EHMG-fish conspecifics* is used by Landa to explain how (effectively) recognition strategies in fish work. From a CaMT perspective, this interplay is intended to stimulate the imagery of the article's expert readers, who immediately incorporate the novel BIOECONOMICS metaphors and similes in their mental text representation. The effect is an easier understanding of specific marine biology concepts.

Landa uses the economic theory of clubs to set a comparison between a group of people who come together/join forces in order to reap economic benefits (i.e. a club) and fish schools. The novel metaphor *club* arises from this comparison to explain the highly coordinated behaviour of schooling fish to obtain hydrodynamic benefits (see context 5). The theory of clubs is further exploited in context (5), where Landa defines another type of schooling fish, *(quasi-)free rider*[4], and the benefit sought by schooling fish, the *club good*. Again, innovative metaphors (in bold) work together with fully-fledged metaphorical terms (small capitals) to produce specialised knowledge in a discourse event.

(5) A fish SCHOOL provides a dramatic example of collective action in nature; it is a **club** which confers benefits on its members […] A **selfish** SCHOOLING fish can reduce its own ENERGETIC COSTS EXPENDED in swimming by positioning itself correctly with respect to those immediately proceeding it. The *follower* fish is literally **free riding** on the hydrodynamic benefits (**the club good**) provided by **club members**. But the **selfish** SCHOOLING fish cannot completely **free ride**: the best it can do for itself is to be a '**quasi-free rider**'. This is because in order for each individual **selfish fish** to benefit from the LOWER COSTSthat come from CONSUMPTION of the **club good** the individual fish must continually adjust the direction of swimming to stay with the group […] The move from lone fish to SCHOOLING member, either as a leader or as a follower, is '**Pareto-optimal**', a term economists use to describe a situation in which one person is made better off, no one is made worse off.

As context (5) shows, the metaphors *free riding*, *club good*, *club members*, and *Pareto-optimal*, which are fully-fledged terminological units from the economics field (cf. Buchanan and Tullock's

comparing fish with the "*calculating* and *rational* economic man" [my emphasis] in context (3).

[4] In economics, the terminological metaphor *free rider* refers to someone who benefits from resources, goods, or services without paying for the cost of the benefit.

1962 public choice theory), acquire specific and precise novel meaning content to designate realities other than people when inserted in a particular scientific biology text in a specific communicative situation. This fact somehow goes along the lines of Rogers' (2007) (con)text-dependent, heterodox view of terminological meaning and term description, which departs from isolated specialised meaning encapsulated in headwords and term entries in specialist dictionaries and glossaries. According to Rogers (2007: 15), the semantic relation *in text* is one of reference by terms as word forms to what they stand for on particular occasions of their utterance. In other words, reference is an utterance-dependent notion in specialised discourse as well, she concludes. Based on this non-prescriptive consideration of terminological meaning, the novel biology metaphors above do designate specific referents, and thus, can be regarded as proto-terms. They would only need to be conventionalised, i.e. systematically used by experts in their scientific research articles, to gain the status of fully-fledged terms in the marine biology field.

Even though these linguistic units cannot be considered terms in their own right (conventionalised and well-established linguistic items with stable specialised meaning across the full spectrum of publications in a specialised knowledge field), they perform a highly restrained referential function with respect to a highly-constrained specialist domain (biology) in a *particular* scientific discourse event. This is an aspect that is characteristic of terms (Rogers, 2007: 15). Consequently, the novel metaphors in Landa's article have become *semantically charged linear structures* (Collet, 2004: 105), whose restrained referential function brings about specialised meaning content. This meaning content adds to the coherence of this communication act, entitling the novel metaphors to help the writer explain specialised knowledge throughout the entire research article, and eventually, conceptually structure her discourse.

Still another reason to treat the novel metaphors analysed as proto-terms is the linguistic variability that affects some of them. Context (5), and in general, Landa's full text include a range of linguistic forms of different novel metaphors (see Table 2 above for the full array of forms). For instance, the metaphor *free rider*, including the nominal compounds *free rider* and *quasi-free rider*, the deverbal compound *free-riding*, and the verb forms *free ride* and *free riding*. All these linguistic variants clearly form a cohesive network that ensures, the flow of specialised information, thus critically contributing to text cohesion. This concatenation of linguistic alternates is a counterpart example of what Collet (2004: 99) calls *coreferential chain* for the spectrum of lexical-syntactic variants of a terminological unit that co-occur in a particular specialised text.

Conclusions

Based on empirical data extracted from a marine biology research article (Landa, 1998), this paper shows that novel metaphors and similes, two figures of thought traditionally understudied in specialised language research, can constitute a critical design feature of scholars' meaning-making capacity in scientific discourse. Textual evidence is given that Landa systematically and deliberately exploits a variety of novel metaphors as extended comparisons from the metaphoric BIOECONOMICS domain to describe domain-specific knowledge. From a pragmatic perspective, these metaphors are experienced by the expert readership as unexpected and innovative linguistic units. Specifically, Landa uses them as rhetorical devices in order to catch the specialist reader's attention and stimulate his/her imagery for specialised concept understanding.

The Career of Metaphor Theory (Bowdle and Gentner, 2005) was used as the theoretical and analytic model to identify and describe the novel metaphoric nature of the expressions examined. Contrary to traditional studies following monolithic terminology theory, this research demonstrates that Collet's (2004) text-linguistic approach to term description is valid to account for the conceptual and linguistic features of novel marine biology metaphors. Because they conform to Collet's term definition criteria, novel metaphors carrying specialised meaning in Landa's article should be regarded as proto-terms, awaiting full acknowledgment and wide use in academic publications to become fully-fledged terminological units.

References

Andrew Goatly. 2002. Text-linguistic comments on metaphor identification. Language and Literature, 11(1):70-90.

Brian Bowdle and Dedre Gentner. 2005. The career of metaphor. Psychological Review, 112(1):193-216.

Bryan Boulier and Robert Goldfarb. 1991. *Pisces economicus*: the fish as economic man. Economics and Philosophy, 7(1):83-86.

Eugene Wüster. 1979. Einführung in die allgemeine Terminologielehre und in die terminologische Lexikographie. UNESCO ALSED LSP Network.

Gerard Steen. 2007. Finding Metaphor in Grammar and Usage. John Benjamins, Amsterdam /Philadelphia.

James Buchanan and Gordon Tullock. 1962. The Calculus of Consent. Universitv of Michigan Press.

Janet Landa. 1998. Bioeconomics of schooling fishes: selfish fish, quasi-free riders, and other fishy tales. Environmental Biology of Fishes, 53:353-364.

Jordan Zlatev, Timothy Racine, Chris Sihna and Esa Itkonen. 2008. The Shared Mind: Perspectives on Intersubjectivity. John Benjamins, Amsterdam/Philadelphia.

Lynne Cameron. 2007. Confrontation or complementarity? Metaphor in language use and cognitive metaphor theory. Annual Review of Cognitive Linguistics, 5:107-136.

Margaret Rogers. 2007. Terminological equivalence in technical translation: A problematic concept? St Jerome and technical translation. Synaps, 20:13-25.

Mojca Pecman. 2014. Variation as a cognitive device. Terminology, 20(1):1-24.

Tanja Collet. 2004. What's a term? An attempt to define the term within the theoretical framework of text linguistics. Linguistica Antverpiensia New Series - Themes in Translation Studies, 3:99-111.

Tina Krennmayr, Brian Bowdle, Gerben Mulder and Gerard Steen. 2014. Economic competition is like auto racing. Metaphor and the Social World, 4(1):65-89.

Short Papers

Extraction of Definitional Contexts from Restricted Domains by Measuring Synthetic Judgements and Word Relevance

César Aguilar
Pontificia Universidad Católica de Chile
Santiago de Chile
caguuilara@uc.cl

Olga Acosta
Pontificia Universidad Católica de Chile
Santiago de Chile
olgalimx@gmail.com

Abstract

In this article we present an ongoing work for extracting conceptual information from specialized-domain texts. Concepts are forms of dividing the world in classes and they are the fundamental pieces for constructing ontologies. In this sense, ontology learning is the (semi-) automatic support for constructing an ontology. Input data are required for the ontology learning and this data are the basic source from which to learn the relevant concepts for a domain, their definitions as well the relations holding between them. With this necessity in mind, we propose here a methodology that takes into account the level of synthetic judgements and word relevance in a sentence in order to filter out and rank sentences. Sentences with high relevance and low level of synthetic judgements should have at least a predicative verb characteristic of analytical definitions for being good candidates.

1 Introduction

Concepts are one of the most fundamental pieces of the cognition: humans daily use concepts for interacting with others and the world. According to Smith (1988), concepts mirror the way that we divide the world into classes, and much of what we learn, communicate, and reason involves relations among these classes. Additionally, Rosch (1978) argues that concepts promote the cognitive economy because the human beings attempt to gain as much information as possible about its environment while minimizing cognitive effort and resources.

Currently, due to the accelerated growth of digital information on the Web and other media as well the urgent necessity of obtaining relevant information in a fast and efficient way from these huge text sources, automated methods or approaches have been developed. For instance, in Maedche and Staab (2004) define *ontology learning* as a number of complementary disciplines that feed on different types of unstructured and semi-structured data in order to support a semi-automatic ontology engineering process. In line with this, Cimiano (2006) describes various sub-processes for constructing an ontology from texts where the concept extraction is an important phase. So, the ontology learning needs input data from which to learn the relevant concepts for a given domain.

According to these ideas, in this paper we sketch a methodology for recognizing candidates to analytical definitional contexts, according to the work developed by Sierra *et al*. (2008). We organize our work as follows: in section 2 we present general information about analytical definitions and the automated extraction of conceptual information. In section 3 we describe the function of adjectives as modifiers of a noun as well the distinction among descriptive and relational adjectives and the relation of descriptive adjectives with synthetic judgements in an attributive form. In section 4 we summarize the methodology proposed. In section 5 we show some preliminary results. Finally, in section 6 we present the future work.

2 Conceptual information

We consider as *conceptual information* the information expressed by specialized definitions, particularly in analytical definitions constituted by *Genus Term* and *Differentia*, following the criteria formulated by Smith (2004). In fact, this author considers that information expressed by

these kinds of definitions is relevant to create ontologies based in lexical relations, specifically hyponymy/hypernymy and meronymy/holonymy relations. Smith argues that these relations, from a philosophical point of view, are basic and universal.

2.1 Analytical Definitions

An analytical definition is a formula for describing a concept, denoted by a linguistic tag, in terms of a superordinate concept (*Genus Term*), and a differentia distinguishing the concept defined from others with the same Genus Term.
For example, the next definition provides a description of the concept *lightning conductor* using one of the most common verbs (i.e., to be) for introducing a definition. In this case, the genus is the concept *device* while the differentia describes the function of the *lightning conductor*:

> [Lightning conductor $^{\text{Term}}$] is a [device $^{\text{Genus Term}}$] [that allows to protect the electrical systems against surges of atmospheric origin $^{\text{Differentia}}$].

2.2 Definitional contexts

Sierra *et al*., (2008) proposed a based-pattern method for extracting terms and definitions in Spanish. This relevant information is expressed in textual fragments called definitional contexts (or DCs) and are constituted by: a term, a definition, and linguistic or metalinguistic forms, such as verb phrases, typographical markers and/or pragmatic patterns, for example:

> The **primary energy**, in general terms, is defined as an energetic resource that has not been affected for any transformation, with the exception of its extraction.

We can see here a DC sequence formed by the term *primary energy*, the definition *that resource that…* and the verb pattern *is defined as*, as well other characteristic units such as the pragmatic pattern *in general terms* and the typographical marker (bold font) that in this case emphasizes the presence of the term.
For achieving this objective, the authors employ verb patterns operating as connectors between terms and definitions. Such patterns syntactically are predicative phrases (or PrP), configured around a verb that operates as a head of this PrP (e.g., to be, to characterize, to conceive, to consider, to describe, to define, to understand, to know, to refer, to denominate, to call, to name).

3 Adjectives

Based on Demonte (1999), adjectives are syntactic units modifying the noun's meaning and associating it with one or various attributes. There are two kinds of adjectives which assign properties to nouns: descriptive and relational adjectives. On the one hand, descriptive adjectives refer to constitutive features of the modified noun. These features are exhibited or characterized by means of a single physical property: color, form, character, predisposition, sound, and so on: *la silla verde* (e.g., *the green chair*). On the other hand, relational adjectives assign a set of properties, i.e., all the characteristics jointly defining names as *sea*: *puerto marítimo* (e.g., *maritime port*). In terminology, relational adjectives represent an important element for building specialized terms, e.g.: *inguinal hernia*, *venereal disease*, *psychological disorder* and others are considered terms in medicine. In contrast, *rare hernia*, *serious disease* and *critical disorder* seem more descriptive judgments and closely related with a specific context.

3.1 Syntactical Identification of Non-Relevant Adjectives

In line with what was just mentioned, if we consider the internal structure of adjectives, two kinds of adjectives can be identified: permanent and episodic adjectives (Demonte, 1999). The first kinds of adjectives represent stable situations, permanent properties characterizing individuals. These adjectives are located outside of any spatial or temporal restriction (i.e., *psicópata- psychopath*). On the other hand, episodic adjectives refer to transient situations or properties implying change and with time-space limitations. Almost all descriptive adjectives derived of participles belong to this latter class as well all adjectival participles (i.e., *harto-jaded*, *limpio-clean*). Spanish is one of the few languages that in syntax represent this difference in the meaning of adjectives. In many languages this difference is only recognizable through interpretation. In Spanish, individual properties can be predicated with the verb *ser*, and episodic properties with the verb *estar*.
Another linguistic heuristics for identifying descriptive adjectives is that only these kinds of adjectives accept degree adverbs, and they can be part of comparative constructions, for example, *muy alto* (Eng.: *very high*). Finally, only descriptive adjectives can precede a noun because

—in Spanish— relational adjectives are always postposed, i.e.: *la antigua casa* (Eng.: the old house).

3.2 Synthetic Judgements and Descriptive Adjectives

According to Kant (2013), *analytic* sentences are those whose truth seems to be knowable by knowing the meanings of the constituent words alone (e.g., *gynecologists are doctors*), unlike the more usual *synthetic* ones (e.g., *gynecologists are rich*), whose truth is knowable by both knowing the meaning of the words and something about the world.

We believe that synthetic judgements in an attributive position (e.g., *rich gynecologists*) are common in non-relevant sentences in specialized domains. This kind of judgements can be recognized from the descriptive adjectives obtained by linguistic heuristics mentioned in section 3.1.

4 Methodology

We present here our methodology for extracting conceptual information from a medical domain corpus. The input data consist of a corpus with POS tagged with FreeLing (Carreras *et al*., 2004).

4.1 Sentence Segmentation

The heuristics assumed here in order to segment our corpus by sentences take into account that a sentence must be separated by a point, to have at least a main verb, and the number of words must be greater than 10 words because the most short DC would have a single word term, the most long predicative verb-is defined as, a possible article preceding genus, genus term and, in this case, some arbitrary limit of words for the *differentia*).

4.2 Filtering out Sentences by Predicative Verbs

The set of sentences obtained by the above step are filtered out by considering predicative verbs mentioned in section 2.2, that is, if there is at least a predicative verb; then it is a good candidate to DC. For the case of *to be*, if it is the first word of the sentence, then it is discarded.

4.3 Chunking

We have used the library of Natural Language NLTK (Bird, Klein and Loper, 2009) in the Python language, for implementing a chunker in order to extract descriptive adjectives with heuristics described in section 3.1.

In this work, we propose a phase of quantification of *synthetic judgments* in candidate sentences as a further filter of non-relevant sentences. We assumed here that synthetic judgments are descriptive adjectives in an attributive position (e.g., *rare syndrome*). So, the higher amount of synthetic judgments in a sentence, the more likely sentence is non-relevant. We considered the set of descriptive adjectives obtained by heuristics as a mechanism for this quantification of syntheticity.

Acosta, Aguilar and Sierra (2013) point out relational adjectives have a higher probability of being part of terms. The heuristics considered in this experiment are:

<RG><AQ>
<VAE><AQ>
<D.*|P.*|F.*|S.*><AQ><NC>

Where RG, AQ and VAE as tagged with FreeLing, correspond to adverbs, adjectives and the verb *estar*, respectively. The tags *<D.*/P.*/F.*/S.*>* correspond to determinants, pronouns, punctuation signs and prepositions. The expression *<D.*/P.*/F.*/S.*>* is a restriction to reduce noise, since elements wrongly tagged by FreeLing as adjectives are extracted without this restriction.

4.4 Weighting Words

We evaluated relevance of simple words by means of a corpus comparison approach by applying the relative frequency ratio (Manning and Schütze, 1999) between two different corpora as in (1). Given that the syntactical pattern of most common terms in Spanish is <NC><AQ> (Vilvaldi, 2004), we take into account only nouns and adjectives in both corpora:

$$weigth(w_i) = log_2\left(\frac{f_{w_{i,D}}}{N_{w_{i,D}}} \Big/ \frac{f_{w_{i,R}}}{N_{w_{i,R}}}\right) \quad (1)$$

Where $f_{w_{i,D}}$, $N_{w_{i,D}}$ correspond to the absolute occurrence frequency of w_i and the size of the domain corpus, respectively. Similarly, $f_{w_{i,R}}$, $N_{w_{i,R}}$ correspond to absolute occurrence frequency of w_i and the size of the reference corpus. The measure in (1) is only calculated for w_i's, where $\frac{f_{w_{i,D}}}{N_{w_{i,D}}} > \frac{f_{w_{i,R}}}{N_{w_{i,R}}}$. Otherwise, w_i can be used as part of a list of non-relevant words for purposes of quantifying non-relevance in sentences. On the other

hand, words only occurring in domain are weighted as in (2). We assume that the reference corpus is large enough for filter out non-relevant words, hence words only occurring in the domain corpus will have a higher probability of being relevant so that the word's frequency can reflect its importance:

$$weight(w_{i,D}) = log_2(1 + f_{w_{i,D}}) \quad (2)$$

4.5 Relevance of Sentences

The ranking of sentences is done by adding up the individual ranks of words present in the sentence. Formally, if *s* (that is, a sentence) has a length of n words, $w_1 w_2 \ldots w_n$, where $n>10$, then the ranking of the candidate *s* is the sum of the weights of all the individual words $w_i \in W$, where *W* are all of the relevant words weighted as mentioned in section 4.4. In contrast, if $w_i \notin W$, then its weight is zero.

5 Preliminary Results

Considering descriptive adjectives automatically extracted by heuristics for quantifying syntheticity, the first results show to be a good filter in order to remove non-relevant fragments by setting thresholds related with the number of descriptive adjectives in sentences. At the same time, the ranking of words achieves to sort sentences according to its relevance for the domain. Additionally, given that only sentences with predicative verbs are considered, a subset of the better ranked sentences are analytical DCs.

If we take into account words where relative frequency in reference is greater or equal than in domain (given its higher occurrence in reference than in domain, we assume they are non-relevant words) as part of this list for removing non-relevant sentences by setting thresholds (here, nouns and adjectives are included) improve significantly the results.

6 Future results

In a future phase of this experiment, we will implement a syntactic phase in order to remove more non-relevant sentences. For instance, sentences with *to be* verb are the most common sentences and which produce so much *noise* in results. Given this, we consider that a syntactic phase capable to assure the occurrence of specific syntactic structures will be an important advance in order to perform a better filtering.

On the other hand, we will continue with the recollection of more information for increasing the sections of science and technology in our reference corpus, in order to improve the word weighing and the calculation of relevance sentences.

Acknowledgments

This paper has been supported by the National Commission for Scientific and Technological Research (CONICYT) of Chile, Project Numbers: 3140332 and 11130565.

References

Olga Acosta, Gerardo Sierra and César Aguilar. 2011. Extraction of Definitional Contexts using Lexical Relations. International Journal of Computer Applications, 34(6): 46-53.

Steven Bird, Ewan Klein and Edward Loper. 2009. Natural Language Processing whit Python. O'Reilly, Sebastropol, Cal.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In Proceedings of the 4th International Conference on Language Resources and Evaluation LREC 2004, ed. by Maria Teresa Lino *et al*., pp. 239-242. ELRA Publications, Lisbon, Portugal.

Philipp Cimiano. 2006. Ontology Learning and Population from Text. Springer, Berlin.

Violeta Demonte. El adjetivo. Clases y usos. La posición del adjetivo en el sintagma nominal. In *Gramática descriptiva de la lengua española*, ed. by Ignacio Bosque and Violeta Demonte. Vol. 1, Ch. 3, pp. 129-215. Espasa-Calpe, Madrid.

Immanuel Kant. 2013. Crítica de la razón pura, edited and traslated to Spanish by Pedro Ribas. Taurus, Madrid.

Alexander Maedche and Steffen Staab. 2004. Ontology Learning. In Handbook on Ontologies, ed. by Steffen Staab and Rudi Studer, pp. 173-190. Springer, Berlin.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Mass.

Rosch. 1978. Principles of categorization. In Cognition and Categorization, ed. by Elinor Rosh and Barbara Lloyd, pp. 27-48. Lawrence Erlbaum Associates, Hillsdale, NJ.

Gerardo Sierra, Rodrigo Alarcón, César Aguilar and Carme Bach. 2008. Definitional verbal patterns for

semantic relation extraction. Terminology, 14(1): 74-98.

Barry Smith. 2004. Beyond concepts: ontology as reality representation. In Formal Ontologies in Information Systems, ed. by Achille Varzi and Laure Vieu, pp. 73-84., IOS Press, Amsterdam.

Edward Smith. 1988. Concepts and Thought. In Psychology of human thought, ed. by Robert J. Sternberg, pp. 19-49. Cambridge University Press, Cambridge, UK.

Jorge Vivaldi. 2004. Extracción de candidatos a términos mediante la combinación de estrategias heterogéneas. Ph. D. Dissertation. IULA-UPF, Barcelona.

How Terms Meet in Small-World Lexical Networks: The Case of Chemistry Terminology

Francesca Ingrosso
SRSMC UMR 7565,
CNRS-Université de Lorraine
Boulevard des Aiguillettes, BP 70239
54506 Vandœuvre-lès-Nancy Cedex, France
francesca.ingrosso@univ-lorraine.fr

Alain Polguère
ATILF UMR 7118,
CNRS-Université de Lorraine
44 av. de la Libération, BP 30687
54063 Nancy Cedex, France
alain.polguere@univ-lorraine.fr

Abstract

We present a new type of terminological model based on formal network structures called *lexical systems*. Those are non-hierarchical lexical graphs where the bulk of lexical relations is formally encoded by means of Meaning-Text lexical functions. This paper describes how this approach to lexical structuring can be applied to the modeling of terminologies, more specifically, to the French and English terminology of chemistry. The first section explains the importance of terminology in chemistry and introduces the aim of our project. Section 2 is a brief presentation of formal characteristics of lexical systems. Section 3 illustrates the type of terminological descriptions we are implementing with the specific case of the chemical term *catalysis*.

1 Structuring the Lexicon of Chemistry

1.1 Key Role of Terminology in Chemistry

Terminology plays a key role in chemistry research. For instance, chemical terms, by their very morphological structure, are closely related to the behavior and properties of substances they designate. As noted by R. Hoffmann[1] and P. Lazlo (1991), the knowledge of the name of a chemical compound, that strictly reflects the compound structure, gives the chemist the "control" over the molecule. Additionally, the terminology of chemistry is extremely vast and fluctuant. The importance of using a proper terminology in chemistry has lead to the creation, in 1919, of the IUPAC: International Union for Pure and Applied Chemistry. IUPAC elaborates rules for the nomenclature of molecules, in order to avoid definitional ambiguities and ensure harmonization of terminological proposals when new molecules are discovered. It has made available on-line for chemists the so-called *Gold Book* (McNaught and Wilkinson, 1997): an extensive dictionary-like description of English chemistry terms.

In spite of such efforts, the terminology of chemistry is loosely normalized. There is also a lack of multilingual perspective as most scientific papers are written in English, which can lead to serious problems in the context of the teaching of this discipline in schools and universities.

1.2 Terminological Networking

In the on-line IUPAC *Gold Book*, term descriptions are eminently relational, as illustrated by the entry for the nominal term *bond* below.[2]

(1) **bond**
There is a chemical bond between two atoms or groups of atoms in the case that the forces acting between them are such as to lead to the formation of an aggregate with sufficient stability to make it convenient for the chemist to consider it as an independent 'molecular species'.
See also: agostic, coordination, hydrogen bond, multi-centre bond

The definition (*There is a chemical bond ...*) in this description is doing its job of establishing connections between *bond* and related terms such as *force*, *aggregate*, etc. It is however an unstructured and non-formalized model for such connections. A greater applicative potential could

[1] Nobel Prize in Chemistry 1981.

[2] http://goldbook.iupac.org/B00697.html

be achieved by explicitly encoding the linguistic structure of a term definition. If such structure mirrors the logical organization of concepts in the corresponding scientific domain, the lexical definition can be an efficient tool for the understanding of scientific texts, for scientific writing and for teaching chemistry.

Additionally, as illustrated by terminological pointers at the end of (1) – *agostic*, *coordination*, *hydrogen bond*, *multi-centre bond* – the mastering of a chemistry term and of the corresponding notion depends on the ability to position this term within the network of other terms that gravitate in its semantic space.

Polysemy is another acute problem in the terminology of chemistry, that is generally ignored in existing resources. Polysemy manifests itself in two ways.

A) It can occur within the terminology itself, when a single form is used to denote different terminological notions – for instance, *to catalyze* as '[for a substance] to cause a certain type of chemical reaction' in (2) *vs.* '[for a chemist] to make this reaction take place' in (3):[3]

(2) *These fiber catalysts can efficiently* ***catalyze*** *the Knoevenagel condensation of benzaldehyde and ethyl cyanoacetate in water (yields: 95-98%).*

(3) *These Ta2O5-T samples were characterized by TG / DTA, XPS, nitrogen adsorption, XRD, and UV-Raman, and were employed to* ***catalyze*** *the gas-phase dehydration of glycerol (GL) to produce acrolein (AC) at around 315 degrees C.*

B) Polysemy can also spread over both chemistry terminology (2)-(3) and general language (4).[4]

(4) *Cities are always building new stadiums with the justification that they'll* ***catalyze*** *the local economy.*

All these observations show that it is necessary to organize the terminology of chemistry according to rigorous theoretical and descriptive principles.

The project we are presenting has a very practical aim: the design and construction of a terminological database in chemistry, for both the English and French languages. It also has theoretical implications as it explores a new approach to the structuring of terminologies based on non-hierarchical graph structures (see lexical systems, section 2 below), where each term is an element in a global lexical network in which it is related to the rest of the domain terminology, as well as to general language lexicon, by means of Meaning-Text lexical functions (Mel'čuk, 1996). Lexical functions have already proved to be an efficient tool to model relations between terms (L'Homme, 2002). In our project, however, the recourse to lexical functions is embedded within a formal proposal for the graph structuring of lexical knowledge – lexical systems – that we believe is particularly suited to account for the interaction between terminologies of different domains – e.g., chemistry terms used in physics – as well as between "purely" terminological units and units that belong to the general language – e.g. *water* as a type of molecule and *water* as a substance.

2 Terminologies as Lexical Systems

The terminological models we are elaborating are grafted on two general language lexical resources: the *English* and *French Lexical Networks* (Gader et al., 2014; Lux-Pogodalla and Polguère, 2011), respectively en-LN and fr-LN.

The design of the en- and fr-LNs is based on a new type of lexical model called *lexical system* (Polguère, 2014). From a formal point of view, a lexical system is a graph whose vertices are lexical units of the lexicon under description and whose edges are lexical relations of essentially two types:

1. semantic relations – (*chemical*) *bond* is linked to *to bond*, *interaction*, *compound* ...;
2. combinatorial relations – (*chemical*) *bond* combines with *covalent*, *ionic* ...

Both types of relations are modeled by means of lexical functions (section 1.2 above): paradigmatic lexical functions in the first case and syntagmatic lexical functions in the second case. Though lexical functions provide the bulk of graph structuring in lexical systems, other types of relations are also implemented. For instance,

[3] Chemistry examples are borrowed from *Web of Science* (`http://webofscience.com/`).

[4] *New York Times*, COCA corpus (`http://corpus.byu.edu/coca/`).

semantic embedding is implemented via links weaved within lexical definitions: cf. the definition of CATALYSIS **I.1**, section 3.3 below, that formally links this term to two semantically embedded terms: REACTION **1** and GIBBS ENERGY.

Such graphs belong to the family of so-called *small-world networks* (Watts and Strogatz, 1998) and their topological properties allow for the automatic identification of semantic spaces through clusterization. Figure 1 illustrates the semantic space of BOND_{N} **I.2** – the chemistry sense of the noun BOND_{N} in the current version of the en-LN[5].

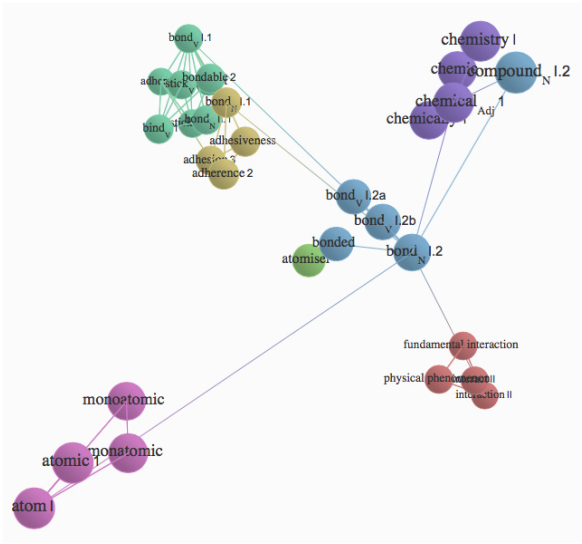


Figure 1: Semantic space of BOND_{N} **I.2**.

Beside being computer-tractable structures, lexical systems are equivalent to "virtual dictionaries", as all properties of lexical units are encapsulated in graph vertices – lexical definitions, grammatical information, citations from corpora (i.e. contexts), etc. Thanks to a specially designed lexical graph editor, it is possible to build a lexical model and, thus, a terminology, by methodically weaving lexical systems (Polguère, 2014). In the specific case of the work presented here, we are describing the terminology of chemistry by weaving the terminological network of this discipline directly on top of the general language en- and fr-LNs. This will allow not only for the proper connection of terminologies in both language, but also for the "interpretation" of chemical terms relative to the non-specialized lexical stock, with which these terms naturally interact in standard research activity as well as in scientific texts.

3 Example: The Case of CATALYSIS I.1

We now illustrate our approach with the en-LN description of the noun CATALYSIS. It possesses the same polysemic structuring as the corresponding verb CATALYZE – see (2), (3) and (4), section 1.2. Therefore, two chemistry senses have to be distinguished within the nominal vocable:

- CATALYSIS **I.1** [*The catalysis occurred via the formation of a chloromethylated triflate complex, and electrophilic addition to an aromatic hydrocarbon.*]
- CATALYSIS **I.2** [*Were they doing catalysis, and if so, how did they recover the catalyst?*]

We will focus on the term CATALYSIS **I.1**, which is the nominal counterpart of the basic sense of the verb exemplified in (2).

3.1 From Lexical Graph to Article-View

When weaving lexical systems with the tailor-made graph editor named *Dicet* (Gader et al., 2012), lexicographers are provided with a textual rendering of lexical information: the *article-view* of the headword. We present the article-view of CATALYSIS **I.1** in Figure 2 below.

It is essential to note that the article-view is only the textual display of fundamentally relational information encoded in the lexical network. For instance, what appears as:

S_1 : **spec** catalyst **I**

in Figure 2 is generated (i) from an **S_1** lexical function link (typical name for the 1st actant of the headword) holding between CATALYSIS **I.1** and CATALYST **I** and (ii) from a grammatical characteristic link connecting this latter unit to the linguistic usage note "spec", that characterizes CATALYST **I** as being a term.[6]

In order to truly apprehend the formal nature of the lexical model in which terminologies are embedded, it is therefore necessary to distance oneself from textual article-views and focus on the

[5] Graph visualizations are based on the Tmuse algorithm (Chudy et al., 2013) and are generated with tools provided by Kodex.Lab (`http://kodexlab.com`).

[6] Even citations – cf. the `[EX]` (ample) zone in Figure 2 – are implemented as connections between individual citations and lexical units they contain.

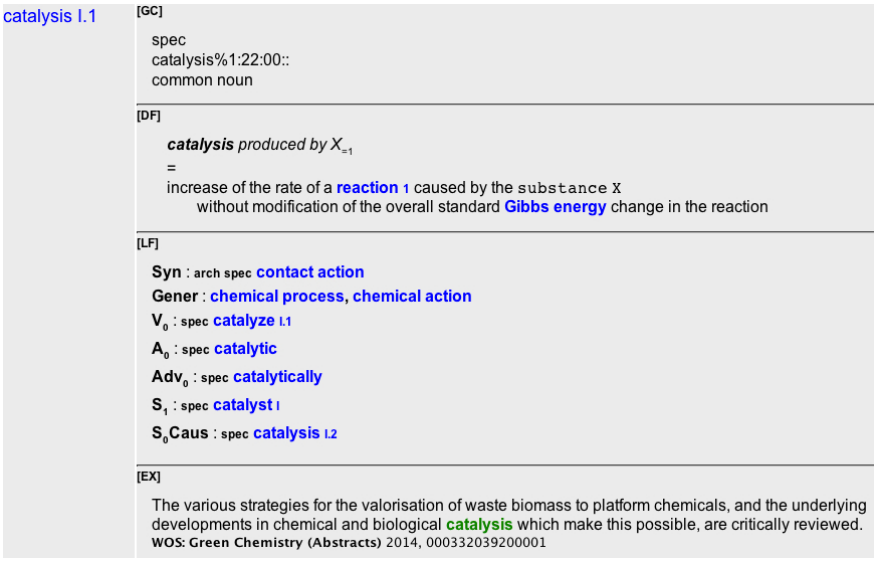


Figure 2: Current en-LN article-view for CATALYSIS **I.1**.

core structuring element of our model: the multidimensional system of lexical function relations that connects lexical units.[7]

3.2 Web of Lexical Function Relations

The structurally most relevant information in Figure 2 appears in the lexical function zone [LF]. It corresponds to the set of paradigmatic links that originate from the CATALYSIS **I.1** headword and connect it the rest of the lexical system. (At present, no syntagmatic link has been encoded for this specific term.) It is this information, together with incoming lexical function links, that position the term in the global structure of the en-LN and defines its semantic space.

In our terminology, the semantic space of a lexical unit such as CATALYSIS **I.1** is much more than just the subgraph constituted of all outgoing and incoming lexical function links. It is the topologically significant cluster of semantically-related nodes that gravitate around CATALYSIS **I.1**, as illustrated in Figure 3.

This semantic space features not only lexical units that are **directly** connected to CATALYSIS **I.1** – e.g. CONTACT ACTION or CATALYST **I** –, but also indirectly connected terms – e.g. GREEN CHEMISTRY or CHEMICAL CHANGE – that entertain significant semantic proximity with this headword.

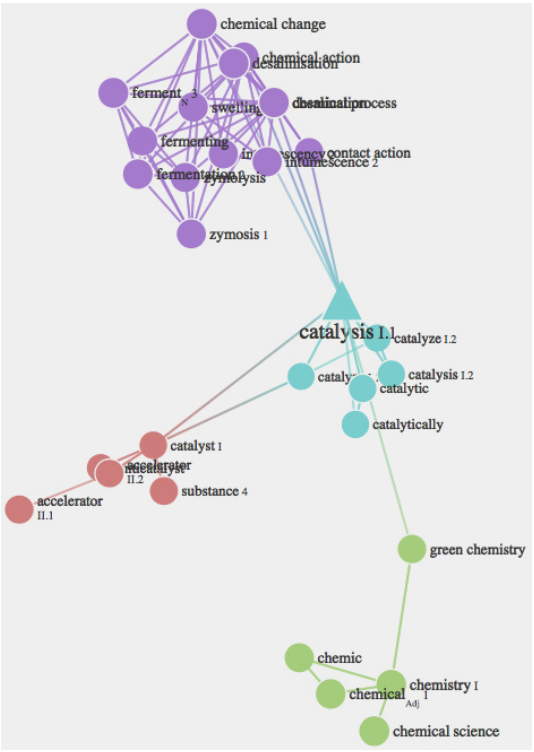


Figure 3: Semantic space of CATALYSIS **I.1**.

At present, semantic space clustering is based

[7] The distinction between article-view and lexical graph perspectives on the en-LN bears some similarity with *written* vs. *graph information modes* in Pram Nielsen (2013).

of Proxemy analysis (Gaume, 2008). It is optimized by taking into consideration the *semantic weight* of each individual lexical function. For instance, a paradigmatic lexical function such as $\mathbf{S_1}$ possesses the maximal semantic weight "2" in the en-LN model of lexical functions, while the $\mathbf{Oper_1}$ lexical function denoting support verb collocates possesses the minimal semantic weight "0".[8]

We believe that lexical systems – small-world graphs of lexical units connected by paradigmatic and syntagmatic relations – are powerful alternatives to more traditional taxonomic models for at least two reasons: (i) they favor semantic space connectivity over a more restricted class-based organization and (ii) they unite both semantic and combinatorial connections within the same formal apparatus (lexical functions).

3.3 Definitional Embedding of Notions

To conclude, we wish to say a few words about definitions and their role in the structuring of terminological knowledge. As indicated in section 2, lexical definitions also participate in the weaving of lexical systems, though to a lesser extent, by implementing semantic embedding. In the specific case of CATALYSIS **I.1**, two lexical units appear in the article-view as clickable targets of definitional embedding links: REACTION **1** and the terminological idiom GIBBS ENERGY. This is made possible by the formal encoding the definition: an XML-like tagging of the definitional text, that we will not detail here for lack of space.

Acknowledgments

We are grateful to TIA 2015 anonymous reviewers for their comments on a preliminary version of our paper. This research is supported by the *PEPS Mirabelle 2015* program (CNRS and Université de Lorraine) for interdisciplinary research.

References

Yannick Chudy, Yann Desalle, Benoît Gaillard, Bruno Gaume, Pierre Magistry and Emmanuel Navarro. 2013. Tmuse: Lexical Network Exploration. In: *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, Asian Federation of NLP, Nagoya, 41–44.

Nabil Gader, Veronika Lux-Pogodalla and Alain Polguère. 2012. Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor. In: *Proceedings of the Third Workshop on Cognitive Aspects of the Lexicon (CogALex III)*, The COLING 2012 Organizing Committee, Mumbai, 109–125.

Nabil Gader, Sandrine Ollinger and Alain Polguère. 2014. One Lexicon, Two Structures: So What Gives? In Heili Orav, Christiane Fellbaum and Piek Vossen (eds.): *Proceedings of the Seventh Global Wordnet Conference (GWC2014)*, Tartu, 163–171.

Bruno Gaume. 2008. Mapping the Forms of Meaning in Small Worlds. *Journal of Intelligent Systems*, 23:848–862.

Roald Hoffmann and Pierre Lazlo. 1991. Representation in Chemistry. *Angewandte Chemie International Edition in English*, 30:1–16.

Marie-Claude L'Homme. 2002. Fonctions lexicales pour représenter les relations sémantiques entre termes. *Traitement automatique de la langue (T.A.L.)*, 43(1):19–41.

Veronika Lux-Pogodalla and Alain Polguère. 2011. Construction of a French Lexical Network: Methodological Issues. In: *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop*, Ljubljana, 54–61.

Alan D. McNaught and Andrew Wilkinson. 1997. *IUPAC. Compendium of Chemical Terminology (the "Gold Book")*, 2nd edition, Blackwell Scientific Publications, Oxford. On-line corrected version: `http://goldbook.iupac.org` (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins.

Igor Mel'čuk. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In Leo Wanner (ed.): *Lexical Functions in Lexicography and Natural Language Processing*, Language Companion Series 31, John Benjamins, Amsterdam/Philadelphia, 37–102.

Alain Polguère. 2014. From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27(4):396–418.

Louise Pram Nielsen. 2013. User experimentation with terminological ontologies. In: *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence TIA 2013*, Paris, 185–188.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' network. *Nature*, 393:440–442.

[8] There is no significant semantic link, in most cases, between a noun and its support verbs. The relationship between *flu* and its support verb *to have*, between *decision* and *to make* ... is first and foremost combinatorial, not semantic.

Constitution d'une base bilingue de marqueurs de relations conceptuelles pour l'élaboration de ressources termino-ontologiques

Luce Lefeuvre
CLLE-ERSS, UMR 5263
CNRS & Université Toulouse Jean-Jaurès
Toulouse, France
Luce.lefeuvre@univ-tlse2.fr

Anne Condamines
CLLE-ERSS, UMR 5263
CNRS & Université Toulouse Jean-Jaurès
Toulouse, France
Anne.condamines@univ-tlse2.fr

Résumé

Les marqueurs de relations conceptuelles sont un moyen efficace de détecter des contextes utiles à l'élaboration de ressources termino-ontologiques. De nombreux travaux existent, mais aucun recensement n'a été effectué. Nous souhaitons construire une base de marqueurs de relation pour l'hyperonymie, la méronymie et la cause, en français et en anglais. La prise en compte de la variation dans l'analyse de ces marqueurs nous permettra de caractériser leur fonctionnement.

1 Introduction

Notre étude se situe dans le cadre du projet ANR CRISTAL (Contextes RIches en connaissanceS pour la TrAduction terminoLogique) dont l'un des objectifs consiste à affiner la notion de Contextes Riches en Connaissances (Meyer, 2001) en prenant en compte différents paramètres de variation tels que la langue (français vs anglais), le domaine (oncologie vs volcanologie), le genre (scientifique vs vulgarisé) et l'utilisateur (traducteur vs terminologue). Nous adoptons ici le point de vue du terminologue. Nous nous intéressons aux relations que peuvent entretenir au moins deux termes, en considérant que ces relations sont un type de connaissance qu'il est possible de découvrir dans un corpus spécialisé. Le projet s'inscrit ainsi dans la thématique de la construction de ressources termino-ontologiques.

L'un des moyens d'accéder à ces connaissances consiste à utiliser des marqueurs de relations conceptuelles. Non ignorés de l'ingénierie des connaissances, de la lexicographie ou de la terminologie, ces éléments linguistiques n'ont pas fait l'objet d'un recensement systématique, ni d'une analyse à grande échelle.

Nous mentionnons en section 2 les travaux dans la lignée desquels nous nous situons. La section 3 décrit la méthodologie que nous avons adoptée. Nous présentons quelques résultats en section 4, et discutons des perspectives de travail en section 5.

2 Travaux antérieurs

La notion de marqueur de relation a souvent été abordée pour élaborer des réseaux de termes, que ce soit en ingénierie des connaissances, en terminologie, ou en traitement automatique des langues. Constitués d'éléments lexico-syntaxiques, typographiques ou dispositionnels (Auger et Barrière, 2008), ils peuvent être utilisés pour expliciter la relation qui unit deux termes. Cette connaissance peut être représentée par un triplet de la forme « Terme 1 - Marqueur - Terme 2 », dans lequel le marqueur précise la relation existant entre les deux termes. Par exemple, la relation d'hyperonymie (générique - spécifique) peut être indiquée par le marqueur « X est un Y + caractéristiques différentielles » (« *Le cancer est une maladie caractérisée par la prolifération incontrôlée de cellules* ») ; et la relation de méronymie (ou partie - tout) peut être indiquée par le marqueur « X être {formé/constitué} de DET Y » (« *le volcan primitif est en majorité constitué de coulées d'andésites* »). Les marqueurs étudiés concernent principalement trois relations : l'hyperonymie, la méronymie, et la cause. Considérées comme structurantes, et supposées universelles, elles apportent des éléments de connaissance sur les termes d'un domaine.

De nombreux travaux s'attachent ainsi à décrire les marqueurs de ces relations (Alarcon-Martinez, 2009 ; Hearst, 1992 ; Garcia, 1998 ; Cruse, 2002 ; Séguéla, 2001 ; Condamines et Rebeyrolle, 2000). Ces études descriptives doivent permettre d'exploiter les marqueurs de relation à l'aide d'outils dédiés, afin de détecter le plus automatiquement possible des triplets structurant les ressources termino-ontologiques.

D'autres travaux plus récents s'intéressent à la variation de ces marqueurs selon le genre textuel, le domaine, ou la langue (Condamines, 2002 ; Marshman, 2006 ; Marshman et L'Homme, 2006 ; Pearson, 1998). Ces travaux montrent que la productivité et la répartition des marqueurs varie parfois fortement d'un domaine ou d'un genre à l'autre. Ils soulignent la nécessité de prendre en compte la variation dans la description des marqueurs de relation, afin d'en étudier la « portabilité » (Marshman et L'Homme, 2006).

Bien que la littérature sur ce sujet soit abondante, il n'existe pas de base de données recensant l'ensemble des marqueurs des relations d'hyperonymie, de méronymie et de cause, ni d'analyse systématique à grande échelle de ces marqueurs. Notre contribution sera de constituer cette base de données et d'analyser chaque candidat-marqueur afin d'en donner une description linguistique fine.

3 Méthodologie

Notre travail s'est déroulé selon deux étapes :

1) Élaboration de la liste des candidats-marqueurs en français et en anglais pour les relations d'hyperonymie, de méronymie et de cause
2) Analyse des occurrences des candidats-marqueurs français en corpus.

Nous détaillons dans la suite chacune de ces étapes.

3.1 Constitution de la base de marqueurs

La base de marqueurs de relation a été construite en deux phases :

1) Recensement des marqueurs de relation pour le français. À partir des travaux existants et dans la lignée des travaux mentionnés en section 2, nous avons fait une liste la plus exhaustive possible des marqueurs français pour trois relations : hyperonymie, méronymie, cause.

2) Élaboration de la liste des marqueurs de relation pour l'anglais (Fabre, 2014). Une première liste de marqueurs a été dressée à partir d'une étude bibliographique. Cette liste a ensuite été enrichie par la traduction de certains marqueurs de relation français. Une première validation a été effectuée en vérifiant dans le COCA corpus[1] les contextes d'apparition des nouveaux candidats-marqueurs anglais obtenus. La relecture de cette liste par une linguiste anglophone a ensuite permis de valider la liste finale.

Le tableau suivant recense le nombre de candidats-marqueurs obtenus pour chaque relation et pour chaque langue[2].

Marqueurs de relation conceptuelle	FRANÇAIS	ANGLAIS
Hyperonymie	33	35
Méronymie	95	99
Cause	192	247

Tableau 1. Nombre de candidats-marqueurs par relation et par langue.

3.2 Évaluation en corpus

La seconde étape de notre travail a concerné l'analyse à grande échelle des candidats-marqueurs en français, en prenant en compte les différents paramètres de variation que nous avons listés plus haut. Notre corpus traite ainsi de deux domaines : la volcanologie, qui appartient aux Sciences de la Terre, et l'oncologie, qui appartient aux Sciences de la Vie. Pour chacun de ces domaines, nous avons pu constituer un corpus scientifique très spécialisé et un corpus vulgarisé, en français et en anglais. Les corpus scientifiques sont constitués de textes issus de revues spécialisées, écrits par des experts à destination d'experts du domaine ou de domaines connexes. Les corpus vulgarisés sont constitués de textes issus de revues ou de sites internet de vulgarisation ; ils sont écrits par des experts du domaine et sont à destination du grand public. Les textes français ont été écrits par des auteurs francophones, et les textes anglais par des au-

[1] Davies, M. (2008-). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Disponible en ligne : http://corpus.byu.edu/coca/.

[2] La liste des marqueurs français et anglais sera disponible sur le site du laboratoire CLLE-ERSS : http://w3.erss.univ-tlse2.fr/

teurs anglophones. Le tableau 2 ci-dessous synthétise ces informations.

	Oncologie	**Volcanologie**
Corpus scientifique	200 000 mots / langue	400 000 mots / langue
	2002 – 2008	1980 - 2012
Corpus vulgarisé	200 000 mots / langue	400 000 mots / langue
	2001 - 2008	1980 - 2002

Tableau 2. Constitution du corpus d'étude.

Nous avons extrait de ce corpus les contextes comportant les candidats-marqueurs recensés. Pour chaque candidat-marqueur de chaque relation, nous avons annoté le contexte comme suit :

- « Oui » : la relation est présente

« *Un dynamisme explosif, extrusif et / ou intrusif **a généré** des cônes stromboliens, des necks basaltiques* » (volcanologie, scientifique).
Le candidat-marqueur « Det X générer Det Y » lie les termes « *dynamisme explosif, extrusif et / ou intrusif* » d'une part et « *cônes stromboliens* » et « *necks basaltiques* » d'autre part par la relation de cause.

- « Non » : le candidat-marqueur n'indique pas la relation attendue

« *Mais notre but **est un** autre volcan **très** actif et dangereux* » (volcanologie, vulgarisation)
Le candidat-marqueur testé « Y être DET X très Adj » n'indique pas la relation d'hyperonymie attendue entre « *but* » et « *volcan* ».

- « Plutôt oui » : le candidat-marqueur exprime la relation conjointement avec un autre élément.

« *Trop de repos ou un manque d'activité peuvent **diminuer** l'oxygénation des tissus musculaires* » (oncologie, vulgarisation)
La nominalisation « oxygénation » associée au candidat-marqueur « diminuer » nous permet d'interpréter la relation comme causale. Deux éléments du triplet sont ainsi présents.

- « Plutôt non »: la relation est difficile à interpréter ; ou alors les éléments en relation ne nous intéressent pas dans l'optique de construire des ressources termino-ontologiques (ce ne sont pas des termes du domaine par exemple).

« *Cette découverte **motive** son élection à l'Académie des sciences* » Relation de cause (volcanologie, vulgarisation)
Il ne nous semble pas pertinent d'intégrer les éléments en relation à une ressource terminologique liée au domaine de la volcanologie.

- « Indéterminé »: nous ne pouvons évaluer la relation (par manque d'indices linguistiques ou par manque de connaissances sur le domaine).

« *Hormones hypophysaires : Ce sont des hormones **sécrétées par** l'hypophyse, glande cérébrale située juste sous le cerveau* » (oncologie, vulgarisation)
Les candidats-termes « *hormones* » et « *hypophyse* » peuvent être reliés par une relation de cause ou une relation de fonction. Aucun indice linguistique ne nous permet de statuer pour l'une ou l'autre des relations.

Environ 10000 contextes ont été annotés selon ces critères.

4 Résultats

Comptabilisant ensemble les « oui » et « plutôt oui », nous avons effectué deux types de calculs : la fréquence d'apparition des candidats marqueurs dans les corpus, et la productivité de chaque candidat-marqueur. Cette productivité correspond au pourcentage des énoncés contenant un candidat marqueur pouvant être interprétés comme contenant la relation attendue.

Nous avons ainsi pu mettre au jour quelques phénomènes de variation liés au domaine ou au genre textuel que nous présentons ici.

4.1 Influence du genre textuel

Les candidats-marqueurs de la relation de méronymie sont organisés selon différentes catégories, qui peuvent préciser par exemple : le type de liaison que les parties d'un ensemble entretiennent (fusion, jonction, inclusion), le type même des parties, si ces parties sont organisées ou non (organisation, non organisation), si elles proviennent de la décomposition d'objets, si elles correspondent à l'expression d'un lieu. Plusieurs candidats-marqueurs n'apparaissant pas du tout dans les corpus, nous avons choisi d'observer la façon dont les occurrences des candidats-marqueurs sont réparties à travers les catégories plutôt que de les comparer de façon isolée.

Catégories de regroupement	**Occ. VULGARISATION**	**Occ. SCIENT**	TOTAL
Inclusion	71	109	180
Non-organisation	37	3	40
Organisation	12	10	22
Types de parties	28	28	56
Lieu	38	40	78
Parties de même genre	29	20	49
TOTAL	215	210	425

Tableau 3. Répartition des occurrences totales de certains candidats-marqueurs de la relation de méronymie par catégorie.

Le tableau 3 ci-dessus présente la répartition des occurrences des candidats-marqueurs selon certaines catégories. On remarque que dans les catégories « Inclusion » et « Non-organisation », les occurrences ne sont pas réparties de manière équilibrée. Les candidats-marqueurs exprimant l'inclusion d'une partie dans une autre sont plus fréquents dans le corpus scientifique. Les candidats-marqueurs indiquant que les parties ne sont pas organisées entre elles sont plus fréquents dans le corpus vulgarisé. Un Chi-test[3] ($p \leq 0{,}001$) a confirmé la différence des deux corpus par rapport aux catégories des candidats-marqueurs.

La catégorie « Inclusion » comporte des candidats-marqueurs comme « X {comprendre/ abriter/comporter/compter/inclure/intégrer} DET Y », ou « Y (être) {classé/classifié/catalogué/ rangé/placé/inclus/étiqueté/catégorisé/groupé} dans DET X ». Leur fréquence plus importante en corpus scientifique peut être due à deux facteurs. Le premier concerne la notion d'inclusion elle-même, qui peut être difficile à appréhender, et que l'on retrouve souvent dans les domaines des mathématiques, de la logique, de la biologie, de la minéralogie. L'autre facteur concerne les éléments en relation dans ces structures. Dans la plupart des contextes contenant ces candidats-marqueurs, les éléments en relation sont des candidats-termes : « acte chirurgical » et « curage axillaire », « complexe volcanique » et « cratère » par exemple. Si l'on ne connait pas la signification de ces termes, un effort de compréhension est nécessaire pour saisir le lien de méronymie qu'il peut exister. On peut ainsi émettre l'hypothèse que l'apparition de ces candidats-marqueurs est liée à une volonté des auteurs, experts de leur domaine, de s'adresser à leurs pairs, sans avoir à détailler à la fois la relation d'inclusion et la spécificité des termes en relation.

La catégorie « Non-organisation » comporte quant à elle des candidats-marqueurs comme « X {être/résulter/de/issu de} DET {tas/amas/ ramassis/masse/accumulation/entassement} de (DET) Y » ou « {tas/amas/ramassis/masse/ accumulation/entassement} de (DET) Y {dans/en/pour former /pour constituer/donner} (DET) X ». La présence d'éléments du lexique comme « tas » ou « accumulation » rend ces structures facilement compréhensibles. Elles ne fournissent pas d'information précise sur les liens que peuvent entretenir les parties. Assez générales et peu spécialisées, elles peuvent être comprises par tous les lecteurs ; quand bien même les éléments en relation seraient des candidats-termes comme « lave » et « dôme » ou « cellules » et « ganglions lymphatiques » par exemple. Ce manque de précision peut expliquer la très forte fréquence d'apparition de ces candidats-marqueurs en corpus vulgarisé. Les auteurs ne peuvent en effet pas détailler toutes les connaissances d'un domaine spécialisé.

Finalement, il semblerait que le genre textuel ait une influence à plusieurs niveaux : au niveau des catégories de la relation de méronymie, au niveau des candidats-marqueurs eux-mêmes, au niveau des éléments en relation.

4.2 Influence du domaine

Le fonctionnement des candidats-marqueurs de cause semble varier de manière significative en fonction du domaine (figure 1).

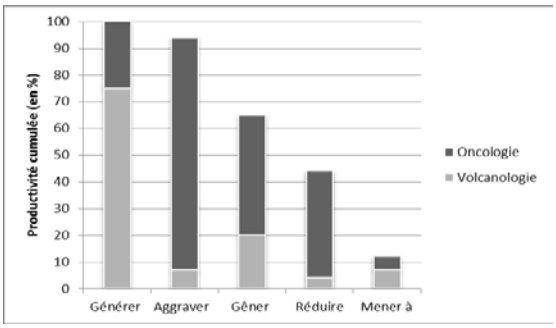


Figure 1. Répartition de quelques candidats-marqueurs de cause selon le domaine.

Dans le domaine de l'oncologie, les candidats-marqueurs de cause les plus représentés (*aggra-*

[3] Je remercie sincèrement Basilio Calderone, membre de CLLE-ERSS pour son aide.

ver, *gêner*, *réduire*, *diminuer*) appartiennent aux catégories /influencer/ et /gêner/, que l'on peut paraphraser par « X cause une influence/une gêne sur Y ». Dans le domaine de la volcanologie, les candidats-marqueurs les plus représentés (*générer*, *mener à*, mais aussi *déclencher*, *créer*, *engendrer*) sont liés à la catégorie /créer/, qui indique qu'un phénomène ou une situation X est la cause de l'existence d'un phénomène ou d'une situation Y. Les objectifs distincts des deux domaines peuvent expliquer ces différences. L'oncologie, et la médecine plus généralement, a pour but de soigner, d'étudier le développement des maladies, de décrire des symptômes, des effets secondaires liés aux traitements. En objet des candidats-marqueurs présents, on retrouve des éléments du lexique comme "séquelles", "dépression", "lymphœdème", "cancer", qui sont liés aux symptômes, aux diagnostics, aux traitements du cancer. La volcanologie a pour objectif d'étudier l'origine ainsi que les mécanismes du volcanisme. Elle s'intéresse à la création des volcans, mais également à ce qu'ils produisent, ce qui va de concert avec la catégorie /créer/ de la relation de cause. On retrouve ainsi en objet des candidats-marqueurs de cause présents des éléments lexicaux qui désignent les produits des volcans : "cendres", "lahars", ou qui concernent la typologie des volcans : "structures", "cônes". Dans les deux cas, il semble bien que ce soit le domaine qui ait une influence sur l'apparition des candidats-marqueurs de cause.

5 Perspectives

Les premiers résultats nous ont permis de valider nos hypothèses sur l'influence du genre et/ou du domaine sur le fonctionnement des marqueurs de relation. Nous souhaitons pour la suite mener des analyses plus fines, afin de mettre en évidence des fonctionnements propres à chaque sous-corpus en lien avec la nature de sa variation. Cela nous permettra de mettre au point des catégories de fonctionnement des marqueurs de relation en fonction du domaine et du genre. Nous pourrons ainsi dresser une typologie des marqueurs de relation, indiquant les cas dans lesquels les marqueurs sont productifs : dans tous les corpus, dans le domaine de la volcanologie, dans le genre vulgarisé, etc.

Le second aspect que nous souhaitons développer concerne l'amélioration de la productivité des marqueurs. Pour cela, nous souhaitons utiliser différentes ressources externes pour contraindre le co-texte. Ces ressources, de type lexical, nous permettront à la fois de sélectionner et de filtrer les contextes extraits. L'utilisation de la liste des candidats-termes ainsi que celle des nominalisations déverbales nous permettront par exemple de sélectionner des triplets complets. Le lexique transdisciplinaire scientifique pourra nous permettre de filtrer certains contextes n'apportant pas de connaissances spécifiques sur le domaine.

Enfin, il serait intéressant de projeter des couples de termes dont on connaît la relation afin de pouvoir découvrir des marqueurs spécifiques au domaine.

Références

Alarcon Martinez, R. (2009). *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*. Thèse de doctorat (non publiée) de l'Université Pompeu Fabra (discipline Sciences du Langage), Barcelone.

Auger, A., & Barrière, C. (2008). Pattern based approaches to semantic relation extraction: a state-of-the-art. *Terminology*, 14(1), 1-19.

Condamines, A. (2002). Corpus analysis and conceptual relation patterns. *Terminology*, 8(1), 141-162.

Condamines, A., & Rebeyrolle, J. (2000). Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In J. Charlet, M. Zacklad, G. Kassel, D. Bourigault, (eds.), *Ingénierie des Connaissances, évolutions récentes et nouveaux défis* (pp. 225-242). Paris: Eyrolles.

Cruse, A. (2002). Hyponymy and its Varieties. In R. Green, C.A. Bean, & S.-H Myaeng (eds.), *The semantics of relationships* (pp. 3-22). Dordrecht/Boston/London, Kluwer Academic Publishers.

Fabre, L. (2014). *Élaboration d'une liste de marqueurs de relations conceptuelles en anglais*. Rapport de stage de Master 2 (discipline Linguistique Anglaise) au sein du laboratoire CLLE-ERSS, Université Toulouse – Jean Jaurès, Toulouse.

Garcia, D. (1998). *Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système Coatis*. Thèse de doctorat de l'Université Paris IV - Sorbonne (discipline Informatique), Paris.

Hearst, M. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes.

Marshman, E. (2006). *Lexical Knowledge Patterns for the Semi-automatic Extraction of Cause-effect and Association Relations from Medical Texts: A Comparative Analysis of English and French*. Thèse de doctorat de l'Université de Montréal (discipline Traduction), Montréal.

Marshman, E., & L'Homme, M.-C. (2006). Portabilité des marqueurs de la relation causale : étude sur deux corpus spécialisés. In F. Maniez, P. Dury, N. Arlin & C. Rougemont (eds.), *Corpus et dictionnaires de langues de spécialité*. *Actes des Journées du CRTT 2006* (pp. 87-110), Nantes.

Meyer, I. (2001). Extracting Knowledge-rich Contexts for Terminography: A Conceptual and methodological Framework. In D. Bourigault, M.C. L'Homme & C. Jacquemin (eds.), *Recent Advances in Computational Terminology* (pp. 279-302). Amsterdam/Philadelphia: John Benjamins.

Pearson, J. (1996). The Expression of Definition in Specialized Texts: A Corpus-based Analysis. In M. Gellerstam et al. (eds.), *Proceedings of the Seventh Euralex International Congress* (pp. 817–824), Göteborg.

Séguéla, P. (2001). *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Thèse de doctorat de l'Université Paul Sabatier (discipline Informatique), Toulouse.

Enhancing Terminological Knowledge With Upper Level Ontologies

Selja Seppälä
University at Buffalo
Buffalo, NY, USA
seljamar@buffalo.edu

Amanda Hicks
University of Florida
Gainesville, FL, USA
aehicks@ufl.edu

Abstract

In this communication, we advocate the use of upper level ontologies such as the Basic Formal Ontology (BFO) to enhance terminological resources and research. First, we present common issues in ontologized terminological work. Then, we review two projects that illustrate the potential advantages of integrating rigorous formal upper level ontologies. Finally, we discuss possible challenges and conclude with a summary of the benefits that such ontologies can bring to both terminological theory and practice.

1 Introduction

Terminologies encode lexical and background knowledge that experts have about their domain of expertise. These resources can be associated with a more explicit ontology-like representation of the entities in the relevant domain. Such representations may include, for example, non-lexicalized concepts. This extends mere terminologies to more sophisticated knowledge representations. Being language independent, ontologized terminologies have the advantage of integrating multilingual terminologies. When augmented with axioms, they can be used in reasoning systems.

Terminological works, where they refer to ontologies at all, generally use Gruber's definition of an ontology as "an explicit specification of a conceptualization." (Gruber, 1995). Ontologies built on the basis of this definition thus depend on peoples' concepts. As a result, the Gruber approach may lead to several distinct ontological representations of the same domain, whether expressed in the same natural language or in different ones. This definition may also lead to a multiplication of ontological terms expressing categories and relations to represent the same or distinct conceptual systems.

However, a multiplication of ontological metalanguages (categories and relations) tends to create knowledge silos (Smith and Ceusters, 2010). In particular, when these metalanguages are domain-specific. Even within a single domain, using distinct metalanguages can limit interoperability of systems using ontological representations of terminologies. Furthermore, from the terminological research viewpoint, a multiplication of categories and relations hinders the advancement of our understanding of conceptual systems, of the internal structure of terms and definitions, etc. To avoid these limitations, we propose that terminologists developing terminological resources and carrying out research would greatly benefit from using an upper level ontology, such as the Basic Formal Ontology (BFO), to integrate resources and research.

In this communication, we present and discuss existing works integrating upper level ontologies, and underline the main advantages of augmenting terminological knowledge with categories and relations from an upper level ontology such as BFO.

2 Limitations of Ontological Terminologies

As shown in Seppälä (2012), common issues in ontologized terminologies are:

- Lack of rigorously defined categories and relations. The interpretation of the metalanguage is left to our intuitive understanding of

the terms used for expressing the used categories and relations.

- *is_a overloading* (Guarino, 1998): the *is_a* relation used for structuring the domain ontology does not distinguish the genuine *is_a* subsumption relation from the *instance_of* relation, and sometimes even from the *part_of* relation.

- Multiplication of domain-specific, sometimes ad hoc, categories and relations.

- When upper level categories are used, limitation to a few top-most categories, which are completed with domain-specific ones (Faber, 2002; Kageura, 2002).

The above limitations result in practical and research-related consequences for terminological works, which can be summarized as follows:

- Confusing and incompatible representations of the same domain.

- Non-interoperable terminologies, which hinders the possibility of sharing and reusing terminological resources.

- Non-generalizable observations of terminological phenomena, which hinders research towards a proper understanding of content-related principles governing term formation, definition composition, and conceptual system organization. This eventually hinders the development of widely (re)usable terminological tools, for example, for creating new terms and writing definitions.

- Non-comparable results of terminological research for lack of a common well-defined domain- and language-independent metalanguage, which hinders the development of a mature integrated science.

These shortcomings can be addressed by adopting well-defined domain- and language-independent upper level categories and relations (ontological metalanguage) of the sort accounted for in formal upper level ontologies.

3 Enhancing Terminologies with Upper Level Ontologies

A formal upper level ontology can be defined as "a representation of the categories of objects and of the relationships within and amongst categories that are to be found in any domain of reality whatsoever." (Spear, 2006)

To illustrate the potential advantages for terminology of using formal upper level ontologies, we describe two projects that integrate such ontologies. There are a few upper level ontologies that can be used by mid-level or domain-specific ontologies to define and relate their categories in a non-ambiguous manner, using logical axioms if needed. The projects described hereafter use, respectively, the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Masolo et al., 2001) and the Basic Formal Ontology (BFO) (Arp et al., 2015).

3.1 The KYOTO Project

The KYOTO project aims at representing domain-specific terms in a computer-tractable axiomatized formalism to allow machines to reason over texts in natural language (Vossen et al., 2010). The system developed in this project comprises a platform for multilingual text mining and information extraction that was tested on documents from the environmental domain. The semantics of the terms are defined through the KYOTO ontology which is based on DOLCE. WordNets and specialized vocabularies of different languages are linked to ontology classes on the basis of a mapping of the English WordNet to the KYOTO ontology. "This basic ontology and the mapping to WordNet are used to model the shared and language-neutral concepts and relations in the domain." (Vossen et al., 2010, 4) The system can thus "detect similar data across documents in different languages, even if expressed differently." (Vossen et al., 2010, 2)

In Vossen et al. (2013), the authors extracted statements from texts about the Chesapeake Bay using Kybots, scripts based on ontological and linguistic patterns in annotated text. The results of baseline fact extraction were compared with Kybot extraction and Cterm extraction, both of which utilize the KYOTO ontology. The result was that the baseline and Kybot profiles had high recall, 100% and 91% respectively. The baseline had low precision (18%), whereas the precision of the

Kybot profiles was better, though not optimal, at 31%. In short, leveraging ontological information in domain-specific fact extraction NLP resulted in high recall and improved precision.

3.2 The BFO-Based Ontological Analysis Framework

The second project consists in analyzing the contents of definitions using the categories and relations of BFO (Seppälä, 2012; Seppälä, 2015b). The author puts forward an ontological analysis framework that is domain- and language-independent and that can be used in any kind of terminological conceptual analysis task. The categories and their characteristics are also used as models that serve to predict the contents of definitions. These may be used as templates in tools to help in definition writing.

The results of the pilot study reported in Seppälä (2012; 2015b) show that these BFO-Templates account for about 75% of the contents of definitions of terms from 15 distinct domains. The rest of the definition contents can be described using the BFO categories and relations.

The well-defined BFO vocabulary can thus be used as a metalanguage to describe definition contents, term formation, and the organization of conceptual systems in a way that research findings can be compared and integrated. In practice, BFO-based ontologized terminologies would have the advantage of being interoperable, as it is already the case for the mid-level and domain-specific ontologies (and the corresponding terminologies) that extend BFO, such as the Ontology for Biomedical Investigations (OBI) and the Ontology for Biobanking (OBIB)[1].

4 Possible Obstacles to Use of Upper Level Ontologies

Using upper level ontologies may sometimes prove challenging. Possible issues may be:

- Upper level ontologies evolve and their categories are, at times, still under development. In those cases, it may not be straightforward under which category to place a term.
- Specifications of the upper level ontology may be sparse and lacking, and sometimes too formal (OWL, first order logic) to be easily understood by terminologists.
- An adequate use requires familiarity with the upper level ontology chosen.

A solution to such issues would be to use existing mappings of WordNet to upper level ontologies as aids for integration. A future mapping of WordNet to BFO should facilitate the integration of BFO in terminological projects (Seppälä, 2015a).

5 Conclusion

We saw that ontologized terminologies present a number of shortcomings that can be addressed by integrating a formal upper level ontology. We illustrated the advantages of such an enhancement by reviewing two projects that use such ontologies. To summarize, the main benefits of using a language- and domain-independent upper level ontology are, on the practical side, the possibility to integrate multilingual and multi-domain terminological resources with one another and with information system tools. The latter can thus use the inferences drawn on the basis of the upper level ontology to reason over and manipulate multilingual natural language texts. Using a well-defined formal upper level ontology as a basis for terminological work would make sharing and reuse of terminologies easier: identifying and sharing common terms, constructing new definitions using the same building blocks (information types and logical axioms), etc. Such a framework avoids semantic conflicts and need for mapping.

On the research side, using a well-defined ontological metalanguage allows: carrying out rigorous and comparable conceptual analysis work in terminology; making language- and domain-independent generalizations about term formation, definition content structure, and terminological systems' organization, which can help develop empirically based content standards and writing aid tools; creating comparable research results that contribute to developing a mature integrated terminological science.

[1] For a full list, see `http://ifomis.uni-saarland.de/bfo/users`. For an illustration of interoperability and its advantages, see the presentation on *The OBIB Ontology for Biobanking*, by Chris Stoeckert, Jie Zheng, and Mathias Brochhausen `http://ncorwiki.buffalo.edu/index.php/CTS_Ontology_Workshop_2015`.

Moreover, a metalanguage using the categories and relations of an upper level ontology for describing terminological data (for example, terms', definitions', and conceptual systems' structure) can fruitfully complement any terminological resource whether or not already ontologized. Cimiano et al. (2011) propose, for example, a model to formally link lexicons (with relevant linguistic descriptions) to ontologies.

Using more specifically a BFO-based metalanguage would further enhance our understanding of the relationship between the lexical, linguistic, conceptual, and ontological levels of terminologies. Indeed, BFO is a realist ontology that represents the things that exist in the world and the relations between them, independently of our conceptualizations thereof. A BFO-based metalanguage may thus provide an additional level of understanding to existing descriptive frameworks.

We therefore encourage terminologists to fully embrace the best ontological practices to enhance their research and resources.

Acknowledgments

This work was supported in part by the Swiss National Science Foundation (SNSF) and by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the SNSF, the National Institutes of Health, or the NCTE. Thanks also to Aurélie Picton and Barry Smith.

References

Robert Arp, Barry Smith, and Andrew D. Spear. 2015. *Building Ontologies with Basic Formal Ontology*. MIT Press, Cambridge, MA.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

Pamela Faber. 2002. Terminographic definition and concept representation. In Johann Haller Belinda Maia and Margherita Ulyrich, editors, *Training the Language Services Provider for the New Millennium*, pages 1–14 [343–354]. Universidade do Porto, Porto.

Thomas R. Gruber. 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human Computer Studies*, 43(5):907–928.

Nicola Guarino. 1998. Some ontological principles for designing upper level lexical resources. In *Proceedings of First International Conference on Language Resources and Evaluation. ELRA-European Language Resources Association, Granada, Spain*, pages 527–534. Citeseer.

Kyo Kageura. 2002. *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. Terminology and Lexicography Research and Practice 5. John Benjamins, Amsterdam.

Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. 2001. Wonderweb deliverable D18 ontology library (final). *ICT Project*.

Selja Seppälä. 2012. *Contraintes sur la sélection des informations dans les définitions terminographiques: vers des modèles relationnels génériques pertinents*. Ph.D. thesis, Département de traitement informatique multilingue (TIM), Faculté de traduction et d'interprétation, Université de Genève.

Selja Seppälä. 2015a. Mapping WordNet to the Basic Formal Ontology using the KYOTO ontology. In *Proceedings of ICBO 2015*.

Selja Seppälä. 2015b. An ontological framework for modeling the contents of definitions. *Terminology*, 21(1):23–50.

Barry Smith and Werner Ceusters. 2010. Ontological Realism: A Methodology for Coordinated Evolution of Scientific Ontologies. *Applied Ontology*, 5:139–188.

Andrew D. Spear, 2006. *Ontology for the Twenty First Century: An Introduction with Recommendations*. Institute for Formal Ontology and Medical Information Science, Saarbrücken, Germany.

Piek Vossen, German Rigau, Eneko Agirre, Aitor Soroa, Monica Monachini, and Roberto Bartolini. 2010. KYOTO: an open platform for mining facts. In *Proceedings of the 6th Workshop on Ontologies and Lexical Resources*, pages 1–10.

Piek Vossen, Eneko Agirre, German Rigau, and Aitor Soroa. 2013. Kyoto: A knowledge-rich approach to the interoperable mining of events from text. In *New Trends of Research in Ontologies and Lexical Resources*, pages 65–90. Springer.

A Methodology for Identifying Terms and Patterns Specific to Requirements as a Textual Genre Using Automated Tools

Maxime Warnier
CLLE-ERSS (UMR 5263)
Université Toulouse – Jean Jaurès & CNRS
Centre National d'Études Spatiales
maxime.warnier@univ-tlse2.fr

Anne Condamines
CLLE-ERSS (UMR 5263)
Université Toulouse – Jean Jaurès & CNRS
anne.condamines@univ-tlse2.fr

Abstract

As a step in a project whose final goal is to propose a Controlled Natural Language for requirements writing at CNES (Centre National d'Études Spatiales), we intend to build the grammar of the textual genre of the requirements. One of the main issues faced when analyzing our corpus is the (sometimes subtle) difference between the terms and syntactic structures pertaining to the genre and those linked to the domain (in our case, the development of space systems) – a difference that is generally not taken into account by automated tools. In this paper, we present a methodology aimed at detecting candidate terms and textual patterns specific to the genre by combining results obtained from a terminology extractor and a data mining tool with a validated resource in use for indexing documents at CNES. The results are then illustrated by a selection of examples from our corpus.

1 Introduction

This study is part of a wider project aiming at improving the writing of requirements[1] at CNES (Centre National d'Études Spatiales), the French Space Agency.

Indeed, the requirements (as well as the specifications, that is, the documents in which they are included) are mostly written in a natural language – in this case, in French –, and as a consequence they may sometimes contain well-known related problems, such as ambiguity and vagueness (Pace & Rosner, 2010). A Controlled Natural Language (CNL) is a possible solution to avoid or at least substantially limit these problems by setting constraints on the lexicon, the syntax or the semantics (Kuhn, 2014).

However, in order for this CNL to be actually applied, we believe that it should not be unnecessarily restrictive and, in particular, not too far removed from the way engineers are already used to write the documents – otherwise, they will probably merely ignore it. In other words, we wish to propose a CNL inspired by already existing data, following a corpus-driven and corpus-based methodology that we describe more in details in (Condamines & Warnier, 2014).

This methodology relies on the existence of a *textual genre*, which Bhatia (1993) defines as "a recognizable communicative event characterized by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs", as it is clearly the case for requirements writing (since it is a recurring task performed by employees working in similar companies), and in particular of a *sublanguage*, defined by Somers (1998) as "an identifiable genre or text-type in a given subject field, with a relatively or even absolutely closed set of syntactic structures and vocabulary". We were already able to provide some evidence in favor of this hypothesis (if not for all requirements, at least for requirements written in French at CNES) and we are now trying to build the grammar (that is to say the set of rules followed – consciously or not – by the speakers of this community to produce acceptable utterances) of this particular genre by semi-automatically analyzing specifications of two former projects.

In the present study, we will focus on the results obtained by a terminological extraction. More specifically, we will propose a method to sort them (as we are interested only in the terms pertaining to the genre, not in those pertaining to

[1] According to one of the definitions given by IEEE (1990), a *requirement* is: "a condition or capability that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed documents".

the domain) and subsequently to use them as a filter to retrieve textual patterns belonging to the grammar of the genre. An example of similar work, based on collocations and n-grams, is given by the transdisciplinary scientific lexicon (Tutin, 2007).

2 Genre vs. domain

Although this grammar should ideally be independent of the field (aerospace industry, aeronautics, software engineering, etc.), in practice, the distinction is not so simple as regards specifications[2]. While some features are indeed inherent in the nature of the documents (because they describe something that does not exist yet, but will have to exist and to conform with the requirements, the use of the future tense and injunctions, for instance, are common), others, however, are closely related to the field to which belongs the future "object" being described. It may reasonably be assumed that the lexical level – since it directly refers to the object in question – is most significantly affected by the domain, but we cannot reject the hypothesis that syntactic structures too may differ from one field to another.

For that reason, if we want to define a terminology of requirements, we must keep in mind that the candidate terms proposed by the terminology extractors may actually belong either to the genre or to the domain. Unfortunately, although the possibility to filter terms by domain has already been highlighted as a user need (Blancafort et al., 2011), traditional extractors do not provide any means to distinguish *a priori* between genre and domain, because they are designed mostly for more didactic corpus, where the field matters much more than the genre (e.g. in order to establish the terminology in use in a company or in a knowledge domain). Furthermore, similar problems are to be expected when using other kinds of automated tools (such as data mining software), as they will also mix the two different types of words and terms.

Specifications are thus unusual, specialized corpora and they bring new challenges to terminology extraction in general. In particular, considering the fact that the candidate terms linked to the domain are probably more numerous than those linked to the genre, we want to find a way to exploit the results without a need for manually revising all of them. In the next section, we present the small experiment we conducted on our corpus of specifications as a possible way to reach this goal, but also to reuse these results to filter textual patterns identified by a text mining tool.

3 Methodology

3.1 Corpora

All the operations described hereafter were performed on two corpora of requirements in French extracted from several specifications provided by the CNES. (All tables and figures were removed from the requirements, because their automatic analysis would have been more difficult.) The first corpus concerns the project called "Pleiades" [3] (two very-high-resolution satellites for Earth observation) and is composed of nearly 120,000 words; the second corpus, related to the smaller project "Microscope"[4] (a microsatellite, whose main objective is to verify a physical principle), contains nearly 44,000 words. Although the requirements were written under similar circumstances and represent the same levels of specifications for the two projects, it is worth noting that Pleiades and Microscope have totally different scales and purposes. Consequently, the fields to which they relate are at least partially distinct.

3.2 Candidate terms

First of all, candidate terms for both corpora were extracted using the terminology extractor developed for the Talismane toolkit (Urieli, 2013); based on a syntactic analysis, it extracts only contiguous noun phrases. The first list we obtained (Pleiades) contained 1,551 candidates, while the second one (Microscope) contained 716 candidates (minimum frequency = 5).

Since they included candidate terms for the genre and for the domain (see section 2), and since we are interested only in the former, all the entries present in a list of terms used at CNES for indexing documents in their knowledge base were removed. This list of domain terms (used here as a "stop list") has been augmented for many years thanks to internal documents of various types and carefully validated by domain

[2] The distinction between *genre* and *domain* itself is actually far from trivial (Lee, 2001).

[3] https://pleiades.cnes.fr/en/PLEIADES/index.htm

[4] http://missions-scientifiques.cnes.fr/MICROSCOPE/

experts. We therefore assume that the terms that it contains are representative of the fields covered by the different projects conducted at CNES over the past years; furthermore, it is safe to think that it should not contain terms belonging to the genre of requirements, because they would not be helpful for indexation (since they are too general). After this step, only 1,355 entries remained for Pleiades (a difference of almost 200 entries) and 598 for Microscope (more than 100 candidates were thus discarded).

In order to remove even more candidate terms supposedly linked to the field, we decided to keep only entries present in both lists (Pleiades and Microscope). This resulted in a much shorter list of just 300 candidate terms (meaning 1,055 were exclusive to Pleiades and 298 to Microscope). This step makes sense because the specifications of Pleiades and Microscope are comparable at many levels, but also because, as already mentioned, the two projects are sufficiently distinct. Hence, whereas the first selection was useful to eliminate candidates related to the field at a more general level (e.g. "satellite" or "simulation"), here some of the candidates were not kept because they are more dependent to one of the two projects, and thus more specialized (e.g. "magnétomètre" 'magnetometer' or "masse interne" 'internal mass'). (However, because the corpus of specifications from Pleiades is almost three times larger than the other corpus, it is also probable that some terms, such as "priorité" 'priority', could have appeared in the Microscope corpus as well.)

Lastly, we proceeded to a manual revision of the remaining candidate terms to eliminate some entries that were obviously noise. The final list contains 267 candidate terms (to be compared with the original list, which would have contained over 1,850 different candidates, or almost 2,000 if the extraction had been performed on the two corpora as a whole). Interestingly, the terms seem to concern both functional requirements (e.g. "fonctionnalité" 'functionality') and non-functional requirements (e.g. "disponibilité" 'availability').

3.3 Textual patterns

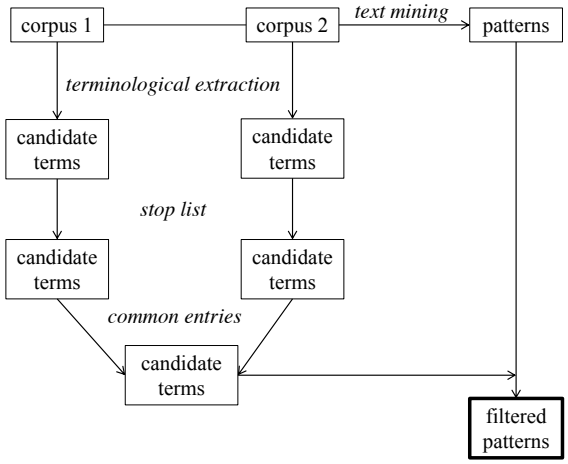
Of course, a grammar of genre should not be limited to the lexicon, as it would be the case with the results of the terminological extraction. We would like to identify recurring syntactic structures or, at least, frequent textual patterns[5] with the help of text mining tools.

For this purpose, we used SDMC (Quiniou et al., 2012) to retrieve patterns of *lemmas* (i.e. canonical forms of the words) frequent in the two corpora, such as "comme décrire dans le tableau" 'as describe in the table', appearing seventeen times in total. These patterns have variable lengths. Here again, the main problem is the huge number of results: almost 14,000 patterns were proposed, making a manual revision extremely time-consuming.

In order to reduce this number to a more reasonable proportion, we have decided to keep only patterns containing at least one of the remaining candidate terms (for the sake of simplicity, the noun phrases were reduced to their heads); indeed, we assume that the structures based on terms belonging to the genre are themselves more likely to be typical of this same genre. This restriction limited the number of patterns to approximately 6,000, among which "être connaître avec un [précision][6] meilleur que (number)" 'be know with a [precision] better than (number)', "être conforme au [format]" 'be consistent with the [format]" and "devoir respecter le [contrainte]" 'must respect the [constraint]'.

The list can be further reduced by focusing on patterns containing a verb. In this way, we consider an intermediary level between the lexicon and the discourse.

To conclude this section, the main steps of the process we described are represented by Figure 1.



[5] Patterns of this kind are the basis of the so-called "boilerplates" (Hull et al., 2005), which are basically fixed structures filled with variable elements at determined positions.

[6] The candidate terms are between square brackets.

Figure 1. Main steps of the proposed methodology.

4 Results

In this section, we briefly discuss some of the results we obtained after applying the process described previously.

4.1 Regarding terms

Some terms belonging to the space domain remain: initialisms ("ASH", "DGAPC"), terms too general to be useful for indexation ("mission", "centre de contrôle" 'control center'), terms of the field ("tuyère" 'nozzle', "calibration").

Others, by contrast, belong more to the genre. They may describe a need ("*besoin* de test+programmation+restitution" '*need* for a test+programmation+restitution') or the characteristics of the objet that is described ("*taille* du buffer temporaire+du paquet TM" '*size* of the temporary buffer+TM packet", "*durée* de désaturation+la manœuvre" '*duration* of desaturation+the manoeuvre'); they can specify expected functions ("*fonction* de gestion+filtrage" '*function* of management+filtering'); or they can be related to the management of the project: possible problems ("*défaillance*" '*failure*', "*défaut*" '*defect*'), necessary documentation ("*rapport* d'avancement+d'expertise" 'progress+expertise *report*'), validation ("*acceptation*" '*acceptance*', "*confirmation*", "*autorisation*" '*authorization*').

Some terms can belong either to the field or to the genre, depending on their modifier: "*date* de début du produit" 'starting *date* of the product' (genre) vs. "*dates* de début et de fin de vidage TM" 'starting and ending *dates* of the emptying of the TM' (field, because of the domain terms "vidage TM").

4.2 Regarding structures

The most frequent verbs in the patterns are: "être" 'to be', "devoir" 'must', "permettre" 'to allow', "mettre" 'to put', "prendre (en compte)" 'to take (into account)', "fournir" 'to provide', "pouvoir" 'to be able', "définir" 'to define', "passer (en mode+dans l'état)" 'to enter (a mode+a state)', "contenir" 'to contain', "donner" 'to give', "utiliser" 'to use', "gérer" 'to manage', "sélectionner" 'to select', "rejeter" 'to reject', "traiter" 'to process', "correspondre" 'to correspond', "générer" 'to generate', "décrire" 'to describe', "tenir" 'to hold', "exécuter" 'to execute', "vérifier" 'to verify', "calculer" 'to calculate'.

Some structures based on these verbs are typical of the corpus:

[Det N permettre de (V+deverbal noun)]: "le DUPC permettra de modifier localement les paramètres du calcul".

[Det N fournir Det N1 (à Det N2)]: "cette interface fournit les positions navigateur de l'instrument".

[Det N utiliser Det N2 (pour V)]: "le système GIDE utilisera le protocole FTP pour effectuer les transferts".

[Det N fournir (à Det N2) Det N3]: "le système de navigation fournira au système informatique central une réference de temps".

[Sur réception de cette TC, le LVC exécute la procédure de mise ON+OFF de Det N (, par l'envoi de commandes (sur+vers+à Det N3))]: "sur réception de cette TC, le LVC exécute la procédure de mise ON de la carte IOT sélectionnée, par l'envoi de commandes discrètes sur l'OBMU" (only in Pleiades).

[Det deverbal noun doit s'exécuter (conditions)]: "la consolidation du scenario de travail au CECT doit s'exécuter en moins de 15 secondes" (only in Microscope).

[Det N (avoir la capacité de+être (capable de+autorisé à)) traiter Det N2]: "le CCC doit avoir la capacité de récupérer et traiter 291 Mo de TM par jour".

These regular structures are therefore part of the grammar of the genre of requirements (at CNES).

5 Conclusion

As emphasized in section 2, specifications of space systems represent a particular type of corpus, because the terms of the domain and the terms of the genre are closely linked – making it difficult to automatically distinguish them. In section 3, we described the methodology we applied to keep only the terms belonging to the textual genre, using an existing resource (built for other needs) and a comparison between two corpora. This also allowed us to identify some structures (textual patterns) belonging to the grammar of the genre, which are used for writing functional requirements (describing expected functions) as well as for non-functional requirements (describing qualities or constraints applied to the system). The grammar could be refined thanks to existing guides to writing specifica-

tions that specify the various sections of the documents and the different types of requirements, which are likely to be expressed in different ways.

Nevertheless, it also appears that it is not always possible to draw a line clearly separating terms of the field and terms of the genre, since some terms may belong to both categories. In any case, the interpretation of the results remains dependent on the objective(s) being pursued.

Finally, we used this experiment as a proof-of-concept; before we can generalize it, we would have to ask for validation by experts (experienced writers). It would also be very interesting to compare our corpus to specifications written in another domain.

References

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. London: Longman.

Blancafort, H., Heid, U., Gornostay, T., Méchoulam, C., Daille, B., & Sharoff, S. (2011). User-centred Views on Terminology Extraction Tools: Usage Scenarios and Integration into MT and CAT Tools. In *Conference "Translation Careers and Technologies: Convergence Points for the Future (TRALOGY)*. Paris, France: INIST.

Condamines, A., & Warnier, M. (2014). Linguistic Analysis of Requirements of a Space Project and Their Conformity with the Recommendations Proposed by a Controlled Natural Language. In B. Davis, K. Kaljurand, & T. Kuhn (Eds.), *Controlled Natural Language* (pp. 33–43). Springer International Publishing.

Hull, E., Jackson, K., & Dick, J. (2005). *Requirements engineering*. London: Springer.

IEEE Standard Glossary of Software Engineering Terminology. (1990). *IEEE Std 610.12-1990*, 1–84. http://doi.org/10.1109/IEEESTD.1990.101064

Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, *40*(1), 121–170.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle. Retrieved from http://ro.uow.edu.au/artspapers/598/

Pace, G. J., & Rosner, M. (2010). A Controlled Language for the Specification of Contracts. In N. Fuchs (Ed.), *CNL 2009 Workshop* (pp. 226–245). Marettimo: Springer.

Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. (2012). What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics? In *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'12)* (pp. 166–177). New Delhi, India.

Somers, H. (1998). An Attempt to Use Weighted Cusums to Identify Sublanguages. In D.M.W. Powers (Ed.), *NeMLaP3/CoNLL 98 : New Methods in Language Processing and Computational Natural Language Learning* (pp. 131–139). ACL.

Tutin, A. (2007). Modélisation linguistique et annotation des collocations: une application au lexique transdisciplinaire des écrits scientifiques. *Formaliser Les Langues Avec L'ordinateur: Actes Des Sixièmes, Sofia 2003, et Septièmes, Tours 2004, Journées Intex-Nooj*, *3*, 189.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Université de Toulouse 2 - Le Mirail, Toulouse.

Posters and Demonstrations

Descriptors for the detection of the chemical risk

Natalia Grabar
UMR8163 STL
CNRS, Université Lille 3
Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

Thierry Hamon
LIMSI-CNRS, Orsay
Université Paris 13
Sorbonne Paris Cité, France
hamon@limsi.fr

Abstract

We propose an experience on the automatic detection of sentences conveying the notion of chemical risk. Our objective is to study which resources are useful for the automatic detection of such sentences. Lexical, semantic and opinion-oriented content of the sentences is studied. Our results indicate that not only lexical and semantic content must be taken into account, but also markers related to the modality, opinion and polarity.

1 Introduction

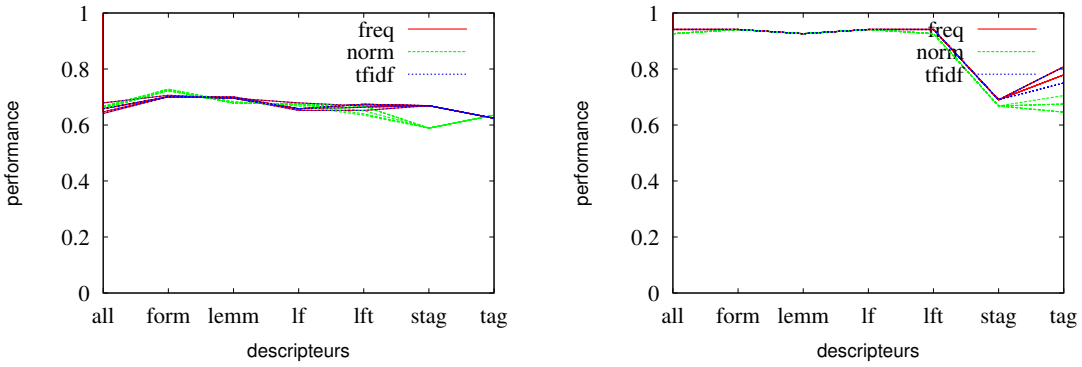
Chemical risk is relative to situations in which chemical products are dangerous for human or animal health and consumption, and for environment. The automatization of the process can help the experts to control and manage large amounts of scientific literature, that have to be analyzed to support the decision making process (van der Sluijs et al., 2008). The sentences that must be recognized are for instance: *The Panel concluded that the current NOAEL for BPA (5 mg/kg b.w./day) would be sufficiently low to exclude any concern for this effect*, or *Despite this lack of evidence, the possibility of poultry and egg consumption as an exposure route to HPAIV remains a concern to food safety experts*. Such sentences are to be assigned in categories related to the chemical risk: the first sentence is related to the significance of the results, while the second is related to the quality of the scientific hypothesis. If such sentences are detected in scientific publications or reports, it means that these publications or reports contain information not fully reliable and can possibly indicate the insufficiency of the corresponding studies and the presence of the risk.

The chemical risk is poorly studied, although the notion of the risk is addressed by other works: building of the dedicated resources (Makki et al., 2008), exploring of known industrial incidents (Tulechki and Tanguy, 2012), computing the exposition to the risk (Marre et al., 2010). Our objective is to study which resources are useful for the automatic detection of the sentences which convey the notion of the chemical risk.

2 Material and Methods

In addition to the lexical and semantic content of the text, we use several kinds of resources in order to favour one aspect or another. These resources contain markers oriented on modality, opinion and polarity expressed by the authors on the proposed experiements: (1) uncertainty (*possible, should, may, usually*) indicates that there are doubts on the results presented, their interpretation, etc.; (2) negation (*no, neither, lack, absent, missing*) indicates that the results have not been observed, that the study does not respect the expected norms, etc.; (3) limitations (*only, shortcoming, insufficient*) indicates that there are some limits of the work, such as unsufficient sample size, small number of tests or doses explored, etc.; (4) approximation (*approximately, commonly, estimated*) indicates other kinds of insufficiency related to imprecise values of substances, samples, dosage, etc.

The work is done with the corpus on chemical risk reporting on several chemical experiments with bisphenol A (EFSA Panel, 2010). It contains over 80,000 occurrences. The reference data are obtained through a manual categorization of the corpus sentences: 425 sentences are assigned to 55 classes of the chemical risk.



(a) Significance of the results

(b) Natural variability

Figure 1: F-measure obtained during the categorization of sentences into classes of the chemical risk.

We tackle the problem through the supervized categorization with the `Weka` platform (Witten and Frank, 2005). Sentences correspond to the units, while 7 classes (most frequent) of the chemical risk are the categories to which the sentences have to be assigned. The resources and the linguistic annotation of corpus (Schmid, 1994) provide several descriptors. These are used to build several sets of descriptors. They represent the semantic and linguistic content of the sentences: *forms* (the forms such as they occur in the corpus), *lemmas* (lemmatized forms), *lf* (combination of forms and lemmas), *tag* (POS tags, such as nouns, verbs, adjectives), *lft* (combination of forms, lemmas and POS-tags), *stag* (semantic tags of words, such as uncertainty, negation, limitations), *all* (combination of all the descriptors available). The descriptors are weighted with various methods (*freq* raw frequency, *norm* normalization by the length of the sentences, and *tfidf* tf-idf normalization).

3 Results

Figure 1 presents some results obtained for two categories: *Significance of the results* and *Natural variability of the results*. We can observe some difference according to the descriptors: the exploitation of forms, semantic tags (with *Significance of the results*) and various combinations of descriptors provide results that are often better for these two categories and for other categories. We assume that these two kinds of descriptors (lexical and semantic content of corpus and the descriptors related to modality, polarity and opinion (Vinodhini and Chandrasekaran, 2012)) provide complementary views on the content and should be combined. These results also indicate that chemical risk is not fully conceptual category but is also related to subjective and contextual values.

References

EFSA Panel. 2010. Scientific opinion on Bisphenol A: evaluation of a study investigating its neurodevelopmental toxicity, review of recent scientific literature on its toxicity and advice on the danish risk assessment of Bisphenol A. *EFSA journal*, 8(9):1–110.

J Makki, AM Alquier, and V Prince. 2008. Ontology population via NLP techniques in risk management. In *Proceedings of ICSWE*.

A Marre, S Biver, M Baies, C Defreneix, and C Aventin. 2010. Gestion des risques en radiothérapie. *Radiothérapie*, 724:55–61.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49.

N Tulechki and L Tanguy. 2012. Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques. In *TALN*, pages 439–446.

Jeroen P van der Sluijs, Arthur C Petersen, Peter H M Janssen, James S Risbey, and Jerome R Ravetz. 2008. Exploring the quality of evidence for complex and contested policy decisions. *Environ. Res. Lett.*, 3(2).

G Vinodhini and RM Chandrasekaran. 2012. Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6):282–292.

I.H. Witten and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

Terminological research in Ukraine

Natalia Grabar
UMR8163 STL
CNRS, U Lille 3
Villeneuve d'Ascq
France
natalia.grabar@univ-lille3.fr

Nataliia Shyshkina, Halyna Zorko
V.M. Glushkov Institute of Cybernetics
National Academy of Sciences of Ukraine
Kyiv, Ukraine
{nata_shy,zorko_gv}@voliacable.com

Thierry Hamon
LIMSI-CNRS, Orsay
U Paris 13
Sorbonne Paris Cité
France
hamon@limsi.fr

Abstract

Our purpose is to present research done in Ukraine on terminology-related questions. We address three aspects: theoretical questions studied, automatic terminology-related tools and existing terminological resources.

1 Terminological research in Ukraine

Terminological research in Ukraine has long tradition:

- the journal *Лексикографічний бюлетень* (*Lexicographical bulletin*) has been founded back in 1951;
- the activity of *Інститут української наукової мови ВУАН* (*Institute of Ukrainian Scientific Language of Academy of Sciences*) and *Технічний комітет стандартизації науково-технічної термінології* (*Technical Committee for Scientific and Technical Terminology Standardization*) consists in normalization and standardization of Ukrainian language and terminology;
- a dedicated conference *Українська термінологія і сучасність* (*Ukrainian Terminology and Contemporaneity*) has been held for the 10th year in 2015.

On the whole, the research in terminology is similar to the research work done in other countries, although the geographical and political situation in Ukraine is specific: poor development during and after the soviet period, overwhelming neighborhood of the Russian society, first years of the opening of Ukrainian researchers to the internationally recognized journals and conferences. Accordingly, the main body of the scientific work is currently published in Ukrainian. Another fact is the domination of theoretical work, although computational methods are emerging. We present various realizations in the terminology area across three lines: theoretical questions, automatic terminology-related tools and existing terminological resources and their utilization.

1.1 Theoretical questions

Terminology research in Ukraine is an active area, although the main terminological work shows mainly theoretical and linguistic orientation, such as:

- history of national lexicology, computational lexicography and terminology (Самойлова, 2004; Ivashchenko, 2013);
- history of national lexicography and terminography dictionaries starting from XVI century and up to the modern times (Балог, 2004);
- computer problems related to the standardization of terminology (Рожанківський and Кузан, 2000);
- semantic processes in the modern Ukrainian terminology (Тищенко, 2004).

1.2 Tools

There are very few automatic tools for the terminology work. We can cite Part-of-Speech taggers: UGtag tagger (Kotsyba et al., 2009) which does not perform syntactic disambiguation and the TNT model for Ukrainian (Babych, 1997).

1.3 Terminological resources

Several specialized areas have been addressed in order to propose their terminological description:

- physics (Кочерга and Мейнарович, 2010);
- law (Тименко, 2004);
- computer sciences (Коссак, 2000; Dmytruk, 2009);
- religion (Пуряєва, 2004);
- literature (Shatalina, 2005);
- Crimean Tatar language (Alieva, 2005; Memetova, 2007).

These are mainly descriptive studies which aim is to confine the area and to identify the main notions and terms.

In accordance with the recent international orientations, new research directions of the area aim at:

- building the electronic corpora (Kelikh et al., 2009; Демська, 2011),
- using electronic corpora for the building of terminologies and dictionaries (Монахова, 2009; Бугаков, 2006; Глибовец and Решетнев, 2014),
- transforming traditional dictionaries in electronic format (Левченко and Кульчицький, 2013).

Up to now, very few works are oriented on the use of terminologies. The currently addressed tasks are related to the software localization (Shyshkina et al., 2010), and to the indexing of speech therapy terminology (Лалаєва et al., 2004).

The majority of these resources are not freely available, although some of them can be queried online: the national corpus of the Ukrainian language[1] (Дарчук, 2010) and online dictionaries[2].

2 Conclusion

The terminological research is an active area in Ukraine. It is dominated by theoretical work and is mainly published in Ukrainian. Nevertheless, as the topics researched are similar to those studied in other countries, we believe that the Ukrainian research community will be soon well positioned and recognized at the international level.

[1] *http://www.mova.info/corpus.aspx?l1=209*

[2] *http://lcorp.ulif.org.ua/dictua/*

Acknowledgments

This work is funded by the LIMSI-CNRS AI project *Outiller l'Ukranien*.

References

VN Alieva. 2005. Onomatopoeic words in the Crimean Tatar language. *Uchenye zapiski*, 18(57):8--11.

B Babych. 1997. Representation and interpretation of ambiguous deep syntactic structures. *Ukrainian Linguistics*, 21:89--100. in Ukrainian.

V Dmytruk. 2009. Typological features of word-formation in computing, the internet and programming in the first decade of the XXI century. In *УДК*, pages 1--11.

VL Ivashchenko. 2013. Historiography of terminology: metalanguage and structural units. In *UDC*, pages 1--22.

E Kelikh, S Buk, P Grzybek, and A Rovenchak. 2009. Project description: designing and constructing a typologically balanced ukrainian text database. In *Методи аналізу тексту*, pages 125--132.

N Kotsyba, A Mykulyak, and IV Shevchenko. 2009. Ugtag: morphological analyzer and tagger for the ukrainian language. In *Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009)*.

ES Memetova. 2007. Lexicophraseological expressive means of the Crimean Tatar language. *Uchenye zapiski*, 18(57):37--39.

OF Shatalina. 2005. Literature terminology of the Old Ukrainian literature of the 18th century. *Uchenye zapiski*, 18(57):5--7.

N Shyshkina, G Zorko, and L Lesko. 2010. Terminology work and software localization in Ukraine. In *Problems of Cybernetics and Informatics*, pages 17--20.

В Балог. 2004. Параметрична система тлумачних словників української мови та її використання в комп'ютерній лексикографії. *Лексикографічний бюлетень*, 10:12--18.

ОВ Бугаков. 2006. Создание семантического словаря предложных конструкций на основе украинского национального лингвистического корпуса. Technical report, Украинский языково-информационный фонд НАН Украины, Киев, Украина.

АН Глибовец and ИВ Решетнев. 2014. Метод итеративного построения терминологии в коллекциях научных текстов на украинском языке. *Кибернетика и системний анализ*, 50(6):53--62.

Н. Дарчук. 2010. Дослідницький корпус української мови: основні засади і перспективи. *ВІСНИК Київського національного університету імені Тараса Шевченка*, 21:45--49.

О Демська. 2011. *Текстовий корпус: ідея іншої форми*. ВПЦ НаУКМА, Київ, Україна.

О Коссак. 2000. Українська комп'ютерна термінологія. In *Сучасні проблеми в комп'ютерних науках*, pages 39--42.

О Кочерга and Є Мейнарович. 2010. *Англійсько-Українсько-Англійський словник наукової мови. Фізика та споріднені науки*. Нова книга, Вінниця, Україна.

Р Лалаєва, Ю Сурованець, and В Тищенко. 2004. Індексація польсько-, російсько- та українськомовної логопедичної термінології. *Лексикографічний бюлетень*, 10:29--36.

ОП Левченко and ІМ Кульчицький. 2013. Технологія перетворення п’ятимовного словника порівнянь в електронну форму. In *Інформаційні системи та мережі*, pages 129--138.

ТВ Монахова. 2009. Застосування прийомів корпусної лінгвістики в лексикографії. *Наукові праці*, 98(85):55--60.

Н Пуряєва. 2004. Дещо про мову богослужіння взагалі та про словник мови богослужіння зокрема. *Лексикографічний бюлетень*, 10:36--42.

Р Рожанківський and М Кузан. 2000. Комп'ютерні проблеми стандартизації термінології. In *Сучасні проблеми в комп'ютерних науках (CCU'2000)*, pages 42--44.

І Самойлова. 2004. З історії відділу лексикології та комп’ютерної лексикографії. *Лексикографічний бюлетень*, 10:7--12.

Л Тименко. 2004. Лексико-тематичні групи української юридичної термінології початку хх століття. *Лексикографічний бюлетень*, 10:65--70.

О Тищенко. 2004. Семантичні процеси в українській термінології початку ххі століття (психолінгвістичний аспект). *Лексикографічний бюлетень*, 10:50--55.

Compilation of a Multilingual (Spanish / English / French / Portuguese) Lexicon of Rural Tourism Terms of Castile and Leon

Beatriz Méndez Cendón
Associate Professor
Departamento de Filología Inglesa
Facultad de Filosofía y Letras
Plaza del Campus Universitario s/n
Universidad de Valladolid
E-47011
cendon@lia.uva.es

Leonor Pérez Ruiz
Associate Professor
Departamento de Filología Inglesa
Facultad de Filosofía y Letras
Plaza del Campus Universitario s/n
Universidad de Valladolid
E-47011
lperezru@fyl.uva.es

Abstract

Our aim is to give an account of the process carried out in the compilation of a multilingual lexicon of rural tourism terms. This lexicon provides equivalents of Spanish local culturally-loaded terms in English, French and Portuguese, the languages spoken by the vast majority of visitors to Castile and Leon. This tool will contribute to improve the communication in the catering industry in this region and will prove to be a user-friendly and time-saving device. More specifically, we have created this terminographic tool as a contribution to: (i) the understanding and fruitful communication between foreign visitors and local workforce; (ii) the specific language needs of this workforce involved in facilitating a pleasant stay to foreign travelers; (iii) the learning of the uses, needs and preferences of tourists; (iv) the avoidance of pitfalls in written texts - web pages, menus, letters, brochures, etc. Our lexicon consists of over 4,600 terms in Spanish and their equivalents in English, French and Portuguese.

1 Introduction

Communication barriers are a big challenge for rural tourism businesses who wish to increase guest satisfaction ratings, when it comes to providing quality service to international markets. One way to overcome this difficulty is to provide catering personnel with a means to easily and quickly check up difficult terms not easy to find in a traditional dictionary.

There are several studies devoted to analyzing the terminology used in rural tourism in Spanish (Fuentes Luque, 2005, 2009; Bonomi, De Santiago & Santos López, 2014; Calvi, 2001, Fijo Léon & Fuentes Luque, 2013, Le Poder & Fuentes Luque, 2005; Kelly, 2005); however, most of them provide just a monolingual perspective. Thus, as part of an ongoing project devoted to analyzing different strategies for the wooing and catering of local rural tourism, we have designed a multilingual lexicon of rural tourism terms that we believe will contribute to enhance the foreign language ability of those who work in the rural tourism sector.

The languages included in our lexicon are Spanish, English, French and Portuguese, since these are the languages spoken by the majority of visitors to Castile and Leon (Boletín de Coyuntura Turística de Castilla y León, 2014).

2 The language of rural tourism

Rural tourism is the kind of tourism that takes place in the countryside, and is based on local resources. It includes a wide array of tourism activities and services in rural destinations, where the different businesses operating there are often owned and managed by local entrepreneurs and their families; typically they offer small-scale accommodation, homemade cuisine and close contact with nature and the host community.

Rural tourism involves the direct contact with the

culture of the area –its folklore, customs, gastronomy, etc. - and therefore, the language of tourism serves as a link between the visitors and the place they are visiting together with its culture (Durán Muñoz, 2008:31).

One significant characteristic of rural tourism is its intimate connection with ethnography, heritage, art and architecture, history and rural life. Thus, the consequence of the promotion of these destinations has been the revival of former methods of production and ways of living, which favor the re-use of terms in many cases obsolete and/or outdated. But it is also very closely linked with other areas of study, such as marketing, economics, public relations, geography and landscaping, etc. It is due to this varied mixture of disciplines that linguists play a basic and fundamental role in the supervision of the correct use of language in rural tourism communicative situations. Our interest in this work is from a lexicographic perspective and it should be noted that the emergence of rural tourism in the past few years worldwide but also in Castile and Leon, has contributed to the widening of the "repertoire of this specialized language" (Fijo & Fuentes, 2013:213). From a multilingual perspective, a problem that arises from this phenomenon is the constant divergence found between the terms used in the different languages; and this is something we intend to clarify with this empirical analysis and classification.

3 Methodology

Our multilingual lexicon is the result of a team project that includes university professors from different Linguistics and Translation departments of Portuguese, French and English languages from the University of Valladolid (Spain) and Universidad do Minho (Portugal).

To build our lexicon we have first compiled a monolingual electronic corpus of Spanish online rural tourism websites and pdf documents referring to Castile and Leon tourism. The corpus consisted of 350 texts in Spanish (50,000 words). For the extraction of the Spanish candidate terms we used a word list generated by AntConc, which is a free online corpus analysis tool. Once extracted, a selection of these terms was made, discarding those that were not appropriate for our purposes. The selected Spanish terms were classified and a conceptual tree was built using tags designating different subfields of the domain of rural tourism -*art*, *artistic activities*, *architecture*, *churches and convents*, *castles and fortifications*, *handicraft*, *popular celebrations*, *typical dishes*, *businesses* and so on-. To be included in this lexicon these subdomains had to directly apply to an issue relevant to Castile and Leon rural tourism.

Once the Spanish terms were assigned to the different subfields, the experts had to conduct a research in order to find appropriate equivalents for these often very culturally-loaded Spanish terms in the target languages. When no equivalent term was found, these specialists had to elaborate a concise and precise definition explaining the entry term.

The final product is a lexicon consisting of more than 4,600 terms in Spanish and their equivalents in English, French and Portuguese. This lexicographic tool is in the process of being published as a printed terminographic resource and also as an online source. The reason for these two versions is that, due to the specific orographical characteristics of the region of Castile and León, it is very difficult to access the Internet in various locations due to technical problems; therefore, a hard copy version of the lexicon would be the most appropriate solution. On the other hand, an online version of the lexicon will allow the authors to constantly update it with new terms. Also this terminology tool will easily be accessed through app devices.

4. Results

The lexicon compiled contains terms dealing with different subfields of rural tourism. While finding appropriate equivalents for the terms, we had to face various terminology and translation issues. Often we had great difficulty when there was a lack of equivalent in the target language due to referential opacity or if there was one it was either too general, inaccurate or confusing (Rabadán, 1991: 166). When unable to detect a correct equivalent, we decided to rather include a brief and accurate description of the pertaining term, for example this was the case of the terms 'palloza' and 'hogaza':

palloza: traditional stone thatched house typical of Leon (En) maison en pierres sèches, couverte de paille, typique du nord de Léon (Fr) casa de campo, quinta típica da Leão (Pt)

hogaza (de pan): round multi-grain peasant bread (En) miche (Fr) fogaça (Pt)

When the case was that the Spanish term has been borrowed, sometimes we decided to also include a brief explanation in the target language. The reason for this is that we believe that despite the lexicalization of this borrowing, some users may not be familiar with it yet. This is the case of 'sangría':

sangría: sangria, red wine punch (En) sangria, boisson rafraîchissante à base de vin rouge et de jus de citron (Fr) sangria (Pt)

Another important issue we had to cope with was the case of polysemous words. There are some words that are found under two different subfields since they are used with different meanings in each subfield, for example: 'muela' listed under the subfields *handicraft* and *parts of the body*, 'crucero' listed under *churches and convents* and *traditional architecture,* and 'talla' belonging to the fields *clothes and accessories* as well as *sculpture*. Since we also provide an alphabetical list of all the terms the user can refer to it if in doubt.

muela (molino): millstone (En) meule (Fr) moinho (Pt) (HANDICRAFT)
muela: molar, back tooth (En) molaire (Fr) molar (Pt) (PARTS OF THE BODY)

crucero: transept (En) croisée du transept (Fr) transepto (Pt) (CHURCHES AND CONVENTS)
crucero: stone cross (En) calvaire, croix dressée sur une plate-forme ou à un carrefour (Fr)
cruzeiro (Pt) (TRADICTIONAL ARCHITECTURE)

talla (madera): carving (En) sculpture (Fr) talha (Pt) (SCULPTURE)
talla: size (En) taille (Fr) tamanho (Pt) (CLOTHES AND ACCESSORIES)

We also took into account the most relevant collocations in the language of rural tourism in Spanish while compiling our lexicon. A collocation is a recurrent word combination consisting of a base and one or more collocates (Méndez Cendón, 2004: 196). We used a concordance tool to detect collocations. The high frequency of occurrence of a given collocate with a certain base was a key issue to identify a collocation for the lexicon. It is important to mention that sometimes there is just a verb as the equivalent in one or some of the target languages instead of an equivalent collocation. For example: 'esculpir/tallar madera/metal', 'hacer autostop', 'ir en bicicleta' and 'ir a caballo':

esculpir (madera): to carve (En) sculpter (Fr) esculpir (Pt)
esculpir (metal): to engrave (En) sculpter (Fr) gravar (Pt)

tallar (madera): to carve (En) tailler, sculpter (Fr) talhar (madeira) (Pt)

autostop (hacer): to hitch-hike (En) auto-stop (Fr) pedir, andar à boleia (Pt)

bicicleta (ir en): to cycle (En) monter à vélo (Fr) andar de bicicleta (Pt)

caballo (ir a): to ride (En) monter à cheval (Fr) montar (Pt)

On other occasions, the equivalents are other collocations in the target languages:

aire (tomar el): to get some fresh air (En) prendre l'air (Fr) ar (apanhar) (Pt)

copa (tomar una): to have a drink (En) prendre un verre (Fr) beber um copo (Pt)

ciclismo de montaña (hacer): to go mountain-biking (En) faire du cyclisme de montagne (Fr) fazer ciclismo de montanha (Pt)

5. Conclusion

The ultimate goal of our study has been to make a minor contribution to the communication between foreign visitors and local workforce in the rural catering industry. We have found that there were relevant issues that needed to be clarified when searching for the right equivalents in the target language – e.g. opacity, borrowings, lexicalization, polysemy, collocations and so on. We hope that our lexicographic tool will be of great help to those involved in the different areas of the rural tourism industry.

Acknowledgments

This paper has been written in the framework of the research projects: "Los idiomas al servicio del turismo rural de Castilla y León: Optimización de los métodos de captación, fidelización y atención al turismo de habla inglesa, portuguesa y francesa" (VA018A10-1), supported financially by the Junta de Castilla y León and "Léxico Combinatorio en Medicina: Cognición, Texto y Contexto (CombiMed)" (FFI2014-51899-R), supported financially by the Ministry of Economy and Competitiveness.

References

Cánoves, Gemma, Montserrat Villarino, Gerda K. Priestly and Asunción Blanco Romero. 2004. Rural tourism in Spain: an Analysis of Recent Evolution. Geoforum, 35: 755-769.

Consejería de Cultura y Turismo. 2014. *Boletín de Coyuntura Turística de Castilla y León.* Junta de Castilla y León, Valladolid.

Durán Muñoz, Isabel. 2012. Caracterización de la traducción turística: problemas, dificultades. Revista de Lingüística y Lenguas Aplicadas, 7: 103-113.

Fijo León, Maria Isabel and Adrían Fuentes Luque. 2013. "A Corpus-based Approach to the Compilation, Analysis and Translation of Rural Tourism Terms. Meta, 58(1): 212-226.

Fuentes Luque, Adrián. 2009. El turismo rural en España: Terminología y Problemas de Traducción. Entreculturas, 1: 469-487.

Méndez Cendón, Beatriz. 2004. Medical Language Collocations: the Case of the Verb Perform. Bravo Gozalo, Jose María (ed). A New Spectrum of Translation Studies. Universidad de Valladolid, Valladolid: 195-208.

Pérez Ruiz, Leonor and Patricia Tabarés Pérez. 2010 The Wooing of English Speaking Clients in the Rural Tourism Industry: The project Turicyl. Pérez Ruiz Leonor (ed.) Estudios de Metodología de la Lengua Inglesa V. Servicio de Publicaciones Universidad de Valladolid, Valladolid: 317-333.

Rabadán, Rosa. 1991. Equivalencia y traducción. Problemática de la equivalencia translémica inglés-español. Universidad de León, León.

Dealing with Large Corpora for Ontology Population

Yuliya Korenchuk (1,2)
(1) LiLPa (Linguistique, Langues, Parole), EA 1339, Universit de Strasbourg
(2) Rebuz SAS, Strasbourg
yuliya.korenchuk@yahoo.fr

1 Introduction

Multilingual ontology population from texts, i.e. addition of new terms in an ontology, requires a suitable parallel or comparable corpus. In this paper, we aim to check whether the corpus selected for our project suits the ontology we want to populate. The corpus for ontology population should not only reflect a specific domain and have a sufficient volume of data, as discussed in (Delpech et al., 2012), but also suit the initial ontology. Using an existing corpus can be an efficient solution used in many projects (Cimiano, 2006; Bouamor, 2014; Pinnis, 2014). However this option is less reliable in the case of a large multi-domain corpus and an ontology which might not cover all the domain concepts. The need for suitability between text corpora and ontology is expressed by (Aussenac-Gilles et al., 2006) who underlined the importance of text type in the corpus, the ontology application, the validation criteria and set up. The text layout can also play an important role: some projects aim to use extralinguistic information for ontology population (Kamel et al., 2013), while others concentrate on the comprehensiveness of the text (Faber et al., 2006).

In this case study, we set up an experiment checking whether a corpus is suitable for ontology population, based on the example of the large parallel (English, French and German) corpus PatTR[1] (Wäschle and Riezler, 2012) and the EcoLexicon[2] terminology knowledge base which we use in our project.

[1] http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/

[2] http://ecolexicon.ugr.es/en/index.htm

2 Resources

2.1 Corpus

The PatTR corpus is a large[3] collection of parallel segments from patents organized by language pairs. These segments are classified into files according to their position in the patent structure (title, abstract, description or claims) (Wäschle and Riezler, 2012). All the language pairs have their metadata files which contain essential information (the IPC[4] code, the reference, etc.) for each segment. As the different domains are mixed, the metadata play a crucial role for our project.

2.2 Ontology

The terminological knowledge base EcoLexicon is developped by the LexiCon research group at the University of Granada. The resource is designed according to the principles of Frame Based Terminology (Faber et al., 2005; Faber et al., 2006; Faber et al., 2009; Faber, 2011; Araúz et al., 2011). It contains 3,547 concepts and 19,712 terms (cf Table 1) on the topic of environment in seven languages, including English, German and French. The terms are connected by generic-specific, part-whole and non-hierarchical relations. The latter refer to the behaviour of the concepts in a domain-specific or a general semantic frame (Faber et al., 2009).

EcoLexicon was built using two types of resources: manually selected domain corpora (bottom-up approach) and a collection of domain thesauri, dictionaries and lexicons (top-down ap-

[3] 22,998,357 segments for EN-DE pair; 18,764,038 for EN-FR and 5,110,262 for FR-DE (PatTR web site)

[4] International Patent Classification, http://www.wipo.int/classifications/ipc/en/

Language	Nb of terms
FR	640
EN	3079
DE	3713

Table 1: Number of terms by language in EcoLexicon

proach) (Faber et al., 2006). The multilingual corpora were built manually from reliable domain sources, taking into account multiple criteria (quantity, quality, simplicity and documentation). The domain-specific terminological resources were compared and evaluated in order to obtain a representative dataset.

3 Main issues

The PatTR corpus represents two main challenges: its size and its domain diversity. In fact, we can hardly estimate the amount of data for each IPC category without getting into the metadata analysis. Domain diversity can also be addressed through the metadata. However, a manual analysis is required: unless being a specialist of the IPC, one needs to manually establish a list of categories potentially corresponding to the ontology domain. Since this intervention is guided by human intuition, we need to validate the sub-corpora choice. Due to its size, the corpus is not designed to be read by a human user, so it is difficult to perform any manual check on the selected domain-specific sub-corpus. We address the validation by counting the concepts occurrences in the selected sub-corpora and checking that these occurrences belong mainly to domain-specific concepts of the ontology.

4 Set up

We defined a set up based on three main steps: (i) manually matching IPC categories to select the sub-corpora, (ii) counting concept occurrences in the selected sub-corpora and (iii) performing a semi-automatic validation of the concept occurrences.

4.1 Manual selection of IPC categories

The main challenge is to select the IPC categories that are suitable for the EcoLexicon ontology population and enrichment. As the corpus is very large, we cannot take all the data to check the concepts occurences. Therefore we started by looking up the domains defined in EcoLexicon and limited our interest to the domains enumerated in Table 2. Then we selected the IPC categories which might suit the EcoLexion ones. As one can notice, this manual correlation is subjective and not transparent, so we need an automated validation.

IPC	EcoLexicon
C02F Treatment of water, waste water, sewage, or sludge	3.2.5.1 Waste treatment and 3.2.5.2 Water treatment
B09C Disposal of solid waste; reclamation of contaminated soil	3.2.5.4 Soil quality management
H01(G,M) Basic electric elements, C01G Inorganic chemistry, H02(J,M) Generation, conversion, or distribution of electric power, C25(B,C,D,F) Electrolytic or electrophoretic processes; apparatus therefor	3.5 Energy engeneering

Table 2: Manual IPC categories selection[5]

4.2 Occurrences count

We counted the occurrences of the concept labels to validate the selected sub-corpora. In fact, this approach is used to evaluate the ontology coverage regarding a domain corpus (Oostdijk et al., 2010). To do so, we lemmatized the corpus with the TreeTagger (Schmid, 1994) tool and transformed both the corpus and the concept labels to lowercase. This caused some problems, because some labels lost their domain specificity (for example, *Be@en* for *berrilium* became *be* and was found nearly in every English phrase). So we had to limit the labels to words longer than 2 characters.

We calculated the percentage of the concept occurrences in the total amount of tokens in the domain sub-corpus. For example, the English sub-corpus for the C02F category has 1,339,946 occurrences for 7,806,687 tokens, so the concept occurrences represent 17% of the tokens (the highest rate in our data collection). The least covered sub-corpus is the French H02M one with 1% of occurrences (55,803 occurrences for 4,359,434 tokens).

[5] As the category titles are too complex, we took in this table the generic IPC descriptions (i.e. *Basic electric elements* is the title of the whole H01 category)

Our hypothesis is that the sub-corpora containing more ontology concepts are more likely to be efficient for ontology population, so we will start the ontology population from the most covered sub-corpora.

The disparity in the coverage among languages observed in the Table 4 (17.16% maximum for English, 3.67% for German and 3.60% for French) can be explained by the difference in the number of EcoLexicon labels for these languages (cf Table 1). As we use a parallel corpus, we will base the suitability analysis on the occurrence percentages for English and try to find the terms translations for the other languages from the corpus.

4.3 Semi-automatic validation

The purpose of this step is to see which concepts appear in the corpus and to validate that their meaning in the corpus matches the one described in the ontology.

We noticed that a part of the occurrences belongs to quite general concepts that are quite close to the definition of transdisciplinary vocabulary (Tutin, 2007; Jacquey et al., 2013), such as *method, device, process* which is due to the fact that the corpus contains segments from patents. We want to be sure that the total occurrences count is not made only of these concepts. To do so, we definded a set of five recurrent concepts and their labels in the three languages (cf Table 3) in order to calculate their percentage in the total occurrences count.

Concept	Labels
Method	method@en, mthode@fr, Methode@de
Process	process@en, processus@fr, Prozess@de
Treatment	treatment@en, traitement@fr, Verarbeitung@de, Behandlung@de
Device	device@en, outil@fr, Mechanismus@de
System	system@en, systme@fr, System@de

Table 3: Manual concepts and labels selection

The combination of the concept occurence and the general concept percentages (cf Table 4) gives a better idea of the best sub-corpora to be used in the next steps. The highest percentage of general concepts is 19% (C25F for English), that means that almost every 5th occurrence is a general concept one. Without final results of the ontology population and enrichment, we cannot judge if this proportion is too high. The maximal percentages of the general concepts for German and French are respectively 9.14% and 1.19% of the concept occurrences.

IPC	Lang	Occurrences %	General concepts %
C02F	**en**	**17.16**	11.86
B09C	en	16.17	13.40
C25C	en	12.54	11.91
C25D	en	11.66	14.88
C01G	en	11.57	14.72
C25B	en	11.18	13.43
C25F	**en**	11.04	**19.00**
H01M	en	10.32	10.73
H02J	en	9.57	15.49
H01G	en	8.15	12.12
H02M	en	8.08	9.54
B09C	**de**	**3.67**	6.66
B09C	**fr**	**3.60**	0.99
C02F	de	3.36	7.29
C25C	fr	3.33	0.88
C25C	de	3.12	2.66
C01G	fr	3.10	0.46
C25B	fr	2.93	0.79
H01M	fr	2.69	0.55
C25D	fr	2.63	0.98
C01G	de	2.57	2.91
C25F	de	2.55	6.75
C25D	de	2.48	4.41
C25B	de	2.25	5.31
C25F	fr	2.18	1.09
H01G	de	1.94	2.70
H01M	de	1.86	4.17
H01G	fr	1.79	1.13
H02J	**de**	1.68	**9.14**
C02F	**fr**	1.63	**1.19**
H02J	fr	1.57	0.94
H02M	de	1.39	4.03
H02M	fr	1.28	0.45

Table 4: Concept occurrences and general concepts percentages

We also manually checked 5 random segments for 10 randomly selected terms, for example *surface water, waste, biomass, etc.*, to be sure that they preserve their terminological meaning. This quick validation helped us to confirm that the selected sub-corpora can be used for future treatments.

Regarding the meaning of the matched terms, the patent titles and abstracts preserve the terminological sense, while the claims part has more rigid style and uses some specific expressions, like

method as in claim X, product accord to one of the claim X, a process along the line of claim, etc. In the same time, domain specific terms contained in claims can still be used as such.

5 Conclusion

The described set up can save time while using a large corpus for the ontology population task. The combined use of metadata and occurrences count show the best sub-corpora that we should keep for further treatment. The semi-automatic validation of occurrences is a useful step which helps to ensure that we know the data used in the project.

Acknowlegments

The research project is supported by the CIFRE grant 2013/0744 delivered by the ANRT. We are grateful to the Lexicon research group[6] of the University of Granada for the access to the EcoLexicon ontology.

References

[Araúz et al.2011] Pilar Araúz, Arianne Reimerink, and Pamela Faber. 2011. Environmental knowledge in EcoLexicon. In *Computational Linguistics-Applications Conference*, number 14, pages 9–16.

[Aussenac-Gilles et al.2006] Nathalie Aussenac-Gilles, Anne Condamines, and Florence Sèdes. 2006. Evolution et maintenance des ressources termino-ontologique: une question à approfondir. *Information interaction intelligence*, HS.

[Bouamor2014] Dhouha Bouamor. 2014. *Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables*. Ph.D. thesis, Université Paris Sud - Paris XI.

[Cimiano2006] Philipp Cimiano. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Application*. Springer US.

[Delpech et al.2012] Estelle Delpech, Béatrice Daille, Emmanuel Morin, and Claire Lemaire. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora : compositional translation and ranking. In *COLING*, volume 3.

[Faber et al.2005] Pamela Faber, Carlos Márquez Linares, and Miguel Vega Exposito. 2005. Framing Terminology: A Process Oriented Approach. *Meta: journal des traducteurs*, 50(4):1492–1421.

[Faber et al.2006] Pamela Faber, Silvia Montero Martínez, María Rosa Castro Prieto, José Senso Ruiz, Juan Antonio Prieto Velasco, Pilar León Arauz, Carlos Márquez Linares, and Miguel Vega Expósito. 2006. Process-oriented terminology management in the domain of Coastal Engineering. *Terminology*, 12(2):189–213.

[Faber et al.2009] Pamela Faber, Pilar Leon, and Juan Antonio Prieto. 2009. Semantic Relations, Dynamicity, And Terminological Knowledge Bases. *Current Issues in Language Studies*, 1:1–23.

[Faber2011] Pamela Faber. 2011. The dynamics of specialized knowledge representation: Simulational reconstruction or the perceptionaction interface. *Terminology*, 17(1):9–29.

[Jacquey et al.2013] Evelyne Jacquey, Agnès Tutin, Laurence Kister, Marie-paule Jacques, Sylvain Hatier, and Sandrine Ollinger. 2013. Filtrage terminologique par le lexique transdisciplinaire scientifique : une expérimentation en sciences humaines. In *Terminologie et Intelligence Artificielle (TIA)*, Paris.

[Kamel et al.2013] Mouna Kamel, Nathalie Aussenac-Gilles, Davide Buscaldi, and Catherine Comparot. 2013. A semi-automatic approach for building ontologies from a collection of structured web documents. In *K-Cap'13 Proceedings of the seventh international conference on Knowledge capture*, pages 139–140.

[Oostdijk et al.2010] Nelleke Oostdijk, Suzan Verberne, and Cornelis Koster. 2010. Constructing a broad-coverage lexicon for text mining in the patent domain. In *LREC*, pages 2292–2299.

[Pinnis2014] Marcis Pinnis. 2014. Bootstrapping of a Multilingual Transliteration Dictionary for European Languages. In *Proceedings of the Sixth International Conference Baltic HLT*.

[Schmid1994] Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester.

[Tutin2007] Agnès Tutin. 2007. Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, XII:5–14.

[Wäschle and Riezler2012] Katharina Wäschle and Stefan Riezler. 2012. Structural and Topical Dimensions in Multi-Task Patent Translation. In *The 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 818–828, Avignon, France.

[6] http://lexicon.ugr.es/

Towards the Integration of Multilingual Terminologies: an Example of a Linked Data Prototype

Elena Montiel-Ponsoda, Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado-de-Cea, Daniel Vila-Suero

Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Campus de Montegancedo sn, Boadilla del Monte 28660 Madrid (Spain)
{emontiel, jbosque, jgracia, lupe, dvila}@fi.upm.es

Abstract

Many language resources are nowadays available in machine readable formats, but still contained in isolated silos. Current Semantic Web-based techniques enable the transformation and linking of those resources to become a navigable graph of linked language resources, which can be directly consumed by third-party applications. The prototype we have developed builds on a web user interface and SPARQL endpoint initially developed to query a single terminological database (Terminesp), now extended to navigate a set of multilingual terminologies. The vocabulary used to represent these terminologies into the linked data format is *lemon-ontolex*, a *de facto* standard for representing lexical information relative to ontologies and for linking lexicons and machine-readable dictionaries to the Semantic Web.

1. Introduction

The Linguistic Linked Open Data (LLOD) cloud[1] is a sub-cloud of linguistic resources provided in an interoperable way (using the Resource Description Framework or RDF data model), freely accessible and linked with each other. In its current state, the LLOD Cloud contains monolingual and multilingual dictionaries, lexicons, thesauri and even corpora. English is the best represented language, and some languages are underrepresented or not present at all.

With Terminesp (a multilingual terminological database created by the Spanish Association for Terminology, AETER), we aimed at validating the *lemon-ontolex* model as a representation scheme for lexical resources, specifically, the so-called `vartrans` module, a dedicated module that accounts for terminological variation and translation relations among entries (Bosque-Gil et al., 2015). Building on that experience, we have now transformed additional multilingual terminological resources, namely, a set of freely available terminology databases[2] from the Catalan Terminological Centre, TERMCAT, into linked data (LD) using *lemon-ontolex* as underlying data format, and aim to showcase the benefits of integrating terminological resources.

In this paper, we focus on the design decisions taken in the transformation and linking steps, and on the impact they have in the search and navigation of the resulting linked terminological data.

In Section 2, we introduce *lemon-ontolex* and the `vartrans` module. In section 3, we describe the design decisions taken in the transformation process. In section 4, we refer to the benefits of browsing and navigating linked multilingual terminologies.

2. lemon-ontolex

The *lemon-ontolex* model is the resulting work of the efforts made by the W3C Ontology Lexica Community Group since 2011 to build a rich model to represent the lexicon-ontology interface. It is largely based on the *lemon* model (McCrae et al., 2012) and consists of a core set of classes and several modules[3]. The `vartrans` module has been developed to record lexico-semantic relations across entries in the same or different languages (Fig. 1.): those among senses and those among lexical entries and/or forms. Lexico-semantic relations among senses are of semantic nature and include

[1] http://linguistic-lod.org/

[2] http://www.termcat.cat/es/terminologiaoberta/

[3] See *lemon-ontolex* final model specifications at http://www.w3.org/community/ontolex/wiki/Final_Model_Specification

terminological relations (dialectal, register, chronological, discursive, and dimensional variation) and translation relations. In contrast, relations among lexical entries and/or forms concern the surface form of a term and encode morphological and orthographical variation, among other aspects.

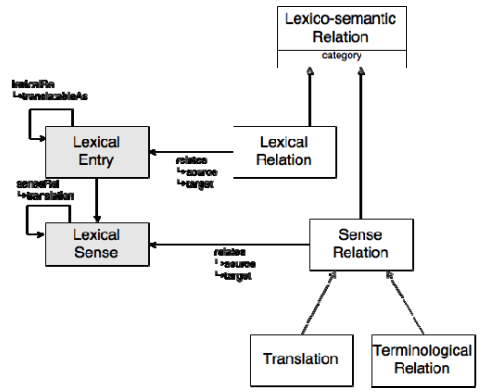


Fig. 1. Classes and properties in *vartrans*

3. Migration and linking of the resources

For the transformation of TERMCAT terminology repertoires to the LD format and linking to Terminesp we followed these steps: data exploration, URI naming strategy, data modeling, RDF generation and linking (Vila-Suero et al., 2014).

Data exploration. TERMCAT terminology repertoires are divided by domain. Each database consists of a list of entries in Catalan and their translations into Spanish, English, French, etc., along with the term type (full form or abbreviation), references to associated terms, synonyms, and, sometimes, definitions. Data for part-of-speech, gender and number in nouns, and subcategorization of verbs, is also available.

URI naming strategy. Inspired by the work in the Apertium dictionaries[4], the term itself, its part of speech and the language of the term are part of the URI of the lexical entry. For lexical senses, the domain is included in the URI.

Modeling. For the modeling process, we regard each term in a set of translations as a specific sense of a lexical entry, a sense that is mapped to a concept in a particular domain. This allows us to have a unique lexical entry *red* (network), for instance, which occurs both in the lexicon *Internet i societat de la informació* as in the lexicon *Indústria electrònica i dels materials elèctrics*, with different senses that we extract from each domain lexicon. This results in a number of RDF lexica that matches the number of languages available in TERMCAT data, and each lexical entry will have a different number of senses depending on its use across domains. In this way, the lexical entry :red-n-es will be mapped to a sense :red-n-es-Internet-sense, as well as to a :red-n-es-Industria-sense, etc. Each of these senses refers to a skos:Concept with a particular definition and domain. Regarding translations, the `vartrans` module represents them as relations across lexical senses of the entries of each lexicon. Parts of speech, subcategorization, gender and number are accounted for as well.

Generation and linking. For the transformation we used the data cleaning and transformation tool OpenRefine[5] with its extension for LD. We linked to lexinfo[6] to cover morphosyntactic information, and to Terminesp at the lexical entry level. Linking to DBpedia is also planned as a next step.

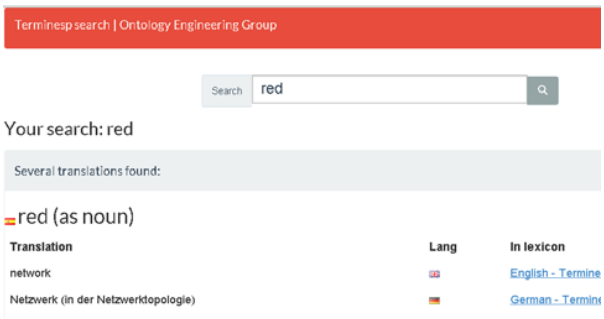


Fig. 2. Web user interface

4. Browsing multilingual terminologies

We reuse the Terminesp web user interface (see Fig. 2.) and SPARQL endpoint to browse and query this set of integrated terminologies[7]. Benefits are related to easy access and reuse of linguistic data by end users (translators, terminologists) and semantic-aware software agents.

Acknowledgements. This work is supported by the FP7 EU project LIDER (610782), and the Spanish 4V project (TIN2013-46238-C4-2-R).

References

J. Bosque-Gil et al. (2015). Applying the OntoLex Model to a Multilingual Terminological Resource. In Proc. of ESWC 2015. Springer.

J. McCrae et al. (2012). Interchanging lexical resources on the semantic web. Language Resources and Evaluation, vol. 46.

D. Vila-Suero et al. (2014). Publishing Linked Data on the Web: the Multilingual Dimension. In P. Cimiano & P. Buitelaar (Eds.) Towards the Multilingual Semantic Web. Springer.

[4] http://linguistic.linkeddata.es/apertium/

[5] http://openrefine.org/index.html

[6] http://lexinfo.net/

[7] http://linguistic.linkeddata.es/terminesp/

Author Index