Web : http://lipn.univ-paris13.fr
ISBN :978-2-9174-9025-9
Dépôt légal : octobre 2013

Proceedings
10th International Conference on Terminology and Artificial Intelligence
TIA 2013

Guadalupe Aguado de Cea - Nathalie Aussenac-Gilles
Presidents

Adeline Nazarenko - Sylvie Szulman
Commitee Organization Presidents

Université Paris 13 - Paris Sorbonne Cité
99, Avenue J.B. Clément
93430 Villetaneuse, France

28–30 octobre 2013

Preface

Terminology and Artificial Intelligence 2013
Terminology for a networked society :
Recent advances in multilingualknowledge-based resources

The Terminology and Artificial Intelligence Conference (TIA) is one of the main international events in Terminology. In 2013, the 10th edition of TIA will be held at the very cradle where it was born, Villateneuse, in 1995. Since then, this conference has been celebrated on a biannual basis and in different venues in France. The aim of these conferences is to bring together researchers in terminology, applied linguistics, knowledge engineering and natural language processing, among other fields. The title of TIA 2013, Terminology for a networked society : Recent advances in multilingual knowledge-based resources, highlights firstly the new challenges that researchers face to represent multilingual knowledge in a networked society like the new Semantic Web, by combining lexicons and ontologies, and secondly emphasizes the importance of term extraction and semantic relations in linguistic resources. Increasing interest on certain fields, such as health care environments, has been proved by the number of contributions and the organization of one of the workshops.

This year, TIA has received 41 submissions from 18 countries in Europe, Africa, Asia and North and South America. These papers were reviewed by a Program Committee formed by 30 experts in the domain from all over the world. Each paper received three reviews. Out of the 41 submissions, 16 (almost 40%) were accepted as long papers, 3 as short papers and 8 as posters, which results in a global acceptance rate of 66%. All of them have been included in this volume.

In order to create an environment of maximum participation among all

conference attendants, the conference has been organized as a single track session with two keynote speakers from the two fields on which the conference is based, Terminology and Artificial Intelligence. The first talk, entitled "Lost in concept systems - Multilingual problems of different conceptualizations" was delivered by Prof. Anita Nuopponen of the University of Vaasa, Finland, and the second talk, entitled "What a beautiful multilingual world : BabelNet 2.0 and friends", by Roberto Navigli of the Sapienza University of Rome. We believe that TIA 2013 will turn into a fruitful and enlightening meeting for all the participants interested in multilingual knowledge-based resources.

Organizing an international conference always requires the services of many people. First and foremost, we would like to thank all the members of the Program Committee for carrying out their meticulous reviews on time. Many thanks to the Organizing Committee at the University Paris 13, Villateneuse, for their invaluable help and for pouring all their efforts into preparing a successful conference. We would like to extend our thanks to the organizers of the workshops ("Optimizing understanding in multilingual hospital encounters" and "Lexical Movement and its impact on specialized resources) and the tutorial on "Frame-based Terminology : An Eco-Lexicon tutorial" for presenting relevant topics that have attracted the attention of many researchers.

Our special gratitude goes to all the authors who submitted their papers and, very specially, to all the participants for contributing to such an interesting and lively conference.

The TIA 2013 program committe chairs, Guadalupe Aguado de Cea (OEG, Universidad Politécnica de Madrid) Nathalie Aussenac-Gilles (IRIT, CNRS, Toulouse)

Invited talks

Lost in concept systems - Multilingual problems of different conceptualizations

Anita Nuopponen, University of Vaasa.

Alternative and overlapping conceptualizations that we meet in our daily life become even more tangible when two or more languages or cultures meet. This is emphasized in the recently published Guidelines for collaborative legal/administrative terminology work by LISE (Legal Languages Interoperability Services) project (2013 : 11) : "When language barriers are coupled with conceptual barriers, i.e. when a transfer takes place across languages and legal systems and so calls for a comparison of legal concepts and cultures, conveying the message across these barriers becomes particularly daunting." In specialized communication - inside or over the language borders - conceptual differences may in the worst-case scenario lead to fatal consequences if the communication partners are not aware of differences and do not find a way to overcome them. Different conceptualizations caused by culture, history, geography, climate, technology, education, economic aspects, theoretical views etc. present challenges not only for those who communicate with each other but to a great degree also for terminologists, translators, technical communicators, standardizers, information system designers, ontology builders, and other knowledge mediators. It is not enough to be competent in language and culture of the languages involved, but also expert knowledge of the special field in question is needed in order to analyze and measure the similarity and degree

of correspondence between the concepts and concept systems behind the terminology.

The point of view taken in this presentation is the one of terminology research and terminology work. ISO 1087-1 defines terminology work as "work concerned with the systematic collection, description, processing and presentation of concepts and their designations". The methods of terminology work have been designed for solving conceptual and terminological problems in specialized communication. Even though the emphasis has originally been on the creation of unified concept systems for international standardization in technological fields, today, terminological methods are applied for all kinds of terminological and conceptual problems in any field, and for various purposes. Researchers and professionals from many disciplines are joining terminology researchers and terminologists in order to develop new methods and tools for new tasks and applications for terminology work and research. As a result, borders between different approaches and traditional dichotomies (e.g. normative vs. descriptive, systematic vs. ad hoc etc.) of terminology work are fading and the overall picture of the terminological activities is getting more complicated. This presentation discusses terminology work as a multifaceted activity drawing on the latest research on multilingual terminology work.

Anita Nuopponen, PhD, is currently Acting Professor in Communication Sciences and head of the Dept. of Communication Studies at the University of Vaasa. She is university lecturer in Applied Linguistics with the teaching and research focus in Terminology Science. She is also the leader of the Master Programme for Technical Communication, for which a new area of specialization "Terminology" is to be started in the autumn 2014. Education : PhD dissertation "Concept relations and systems for terminological analysis" (in Swedish) 1994 ; MSc (Economics), BSc (Economics/Business correspondence) ; Language studies in Finnish, Swedish, English, German, Spanish, French, Icelandic, Greenlandic, and lately also in Japanese. Research and teaching interests cover terminology science, concept analysis methodology but also CMC (computer-mediated/online/digital communication), technical communication, and cognitive science. Since 2008 a member of the Advisory Board of "Terminology International Journal of Theoretical and Applied Issues in Specialized Communication", Benjamins. Anita Nuopponen, PhD, is currently Acting Professor in Communication Sciences and head of the Dept. of Communication Studies at the University of Vaasa. She is university lecturer in Applied Linguistics

with the teaching and research focus in Terminology Science. She is also the leader of the Master Programme for Technical Communication, for which a new area of specialization "Terminology" is to be started in the autumn 2014. Education : PhD dissertation "Concept relations and systems for terminological analysis" (in Swedish) 1994 ; MSc (Economics), BSc (Economics/Business correspondence) ; Language studies in Finnish, Swedish, English, German, Spanish, French, Icelandic, Greenlandic, and lately also in Japanese. Research and teaching interests cover terminology science, concept analysis methodology but also CMC (computer-mediated/online/digital communication), technical communication, and cognitive science. A list of publications : http ://lipas.uwasa.fi/ atn/AnitaNuopponen. Since 2008 a member of the Advisory Board of “Terminology International Journal of Theoretical and Applied Issues in Specialized Communication”, Benjamins.

What a beautiful multilingual world : BabelNet 2.0 & friends !

Roberto Navigli, Sapienza University.

The textual content that is available on the Web is becoming ever increasingly multilingual, providing an additional wealth of valuable information. Most of this information, however, remains inaccessible to the majority of users because of language barriers. Consequently, both humans and automatic systems need tools which will enable them to enjoy the beauty and the usefulness of this varied multilingual world.

BabelNet is a major project under way at the Linguistic Computing Laboratory of the Sapienza University of Rome, focusing on the creation of a very large multilingual semantic network. In this talk I introduce, for the first time, a major new version of the network, named BabelNet 2.0, which covers 50 languages and provides both lexicographic and encyclopedic knowledge for all the open-class parts of speech. As such, BabelNet 2.0 represents an ideal solution not only for humans, who can use it as a multilingual "encyclopedic dictionary" (http ://babelnet.org), but also for the multilingual processing of text, thanks to a new high-performance API.

Roberto Navigli is an Associate Professor in the Department of Computer Science of the Sapienza University of Rome. He was awarded the Marco Cadoli 2007 AI*IA Prize for the best doctoral thesis in Artificial Intelligence and the Marco Somalvico 2013 AI*IA Prize for the best young

researcher in AI. He is the recipient of an ERC Starting Grant in computer science and informatics on multilingual word sense disambiguation (2011-2016) and a co-PI of a Google Focused Research Award on Natural Language Understanding.

His research lies in the field of Natural Language Processing (including word sense disambiguation and induction, ontology learning from scratch, large-scale knowledge acquisition, open information extraction and relation extraction).

He has served as an area chair of ACL, WWW, and *SEM, and a senior program committee member of IJCAI. Currently he is an Associate Editor of the Artificial Intelligence Journal, a member of the editorial boards of Computational Linguistics and the Journal of Natural Language Engineering, and a guest editor of the Journal of Web Semantics.

Contents

Program Committee

Conference Chair : Guadalupe Aguado de Cea (OEG, Universidad Politécnica de Madrid, Spain) Nathalie Aussenac-Gilles (IRIT, CNRS, Toulouse, France)

Amparo Alcina	Universitat Jaume-I, Castellón de la Plana, Spain
Sophia Ananiadou	NaCTeM, Manchester, UK
Caroline Barrière	CRIM, Montréal, Canada
Paul Buitelaar	DERI, Galway, Ireland
Maria Teresa Cabré	Universitat Pompeu Fabra, Spain
Farid Cerbah	Dassault Aviation, Paris, France
Jean Charlet	AP-HP & INSERM, Paris, France
Philipp Cimiano	University of Bielefeld, Germany
Anne Condamines	CLLE-ERSS, Toulouse, France
Béatrice Daille	LINA, Université de Nantes, Nantes, France
Valérie Delavigne	Institut National du Cancer, France
Pascaline Dury	Université Lyon 2, Lyon, France
Fidelia Ibekwe-San Juan	Université Lyon 3, France
Marie-Christine Jaulent	INSERM, France
Kyo Kageura	University of Tokyo, Japan
Olivia Kwong	City University Hong Kong, Hong Kong, China
Marie-Claude L'Homme	(OLST, Université de Montréal, Canada
Elena Montiel-Ponsoda	Universidad Politéécnica de Madrid, Spain
Adeline Nazarenko	LIPN, Université Paris 13 SPC - CNRS, Villetaneuse, France
Pascale Sébillot	IRISA, Rennes, France
Monique Slodzian	CRIM-INALCO, Paris, France
Mari Carmen Suarez-Figueroa	Universidad Politécnica de Madrid, Spain
Sylvie Szulman	LIPN, Université Paris 13 SPC - CNRS, Villetaneuse, France
Annette Ten Teije	Free University Amsterdam, Netherlands
Koichi Takeuchi	Okayama University, Japan
Rita Temmerman	Erasmushogeschool, Belgium
Yannick Toussaint	LORIA, Nancy, France
Špela Vintar	University of Ljubljana, Ljubljana, Slovenia
Pierre Zweigenbaum	LIMSI-CNRS & CRIM/ERTIM-INALCO, Paris, France

Organizing Commitee

Presidents :
Adeline Nazarenko (LIPN-Université Paris 13 SPC)
Sylvie Szulman (LIPN-Université Paris 13 SPC)

Laurent Audibert	LIPN, Université Paris 13 SPC
Ines Bannour	LIPN, Université Paris 13 SPC
Sondes Bannour	LIPN, Université Paris 13 SPC
Davide Buscaldi	LIPN, Université Paris 13 SPC
François Lévy	LIPN, Université Paris 13 SPC
Jorge J. Garcia Flores	LIPN, Université Paris 13 SPC
Ehab Hassan	LIPN, Université Paris 13 SPC
Thibault Mondary	LIPN, Université Paris 13 SPC
Nada Mimouni	LIPN, Université Paris 13 SPC
Antoine Rozenknop	LIPN, Université Paris 13 SPC
Sylvie Salotti	LIPN, Université Paris 13 SPC
Nadi Tomeh	LIPN, Université Paris 13 SPC
Jonathan Van Puymbrouk	LIPN, Université Paris 13 SPC
Haifa Zargayouna	LIPN, Université Paris 13 SPC

Session : Representing Multilingual Linguistic Knowledge

Multilingual Variation in the context of Linked Data

Elena Montiel-Ponsoda
Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
elena.montiel@upm.es

John McCrae
Semantic Computing Group
CITEC
Universität Bielefeld
jmccrae@cit-ec.uni-bielefeld.de

Guadalupe Aguado-de-Cea
Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
lupe@fi.upm.es

Jorge Gracia
Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
jgracia@fi.upm.es

Abstract

In this paper we present a revisited classification of term variation in the light of the Linked Data initiative. Linked Data refers to a set of best practices for publishing and connecting structured data on the Web with the idea of transforming it into a global graph. One of the crucial steps of this initiative is the linking step, in which datasets in one or more languages need to be linked or connected with one another. We claim that the linking process would be facilitated if datasets are enriched with lexical and terminological information. Being that the final aim, we propose a classification of lexical, terminological and semantic variants that will become part of a model of linguistic descriptions that is currently being proposed within the framework of the W3C Ontology-Lexica Community Group to enrich ontologies and Linked Data vocabularies. Examples of modeling solutions of the different types of variants are also provided.

1 Introduction

In the same way that hyperlinks enable the creation of connections between documents, current semantic web technologies enable the establishment of connections or *links* between or among pieces of data, information, and knowledge, in what is known as the ***Linked Data*** paradigm[1], with the goal of better exploiting them in linked data-driven Web applications (Hausenblas, 2009).

In the context of this new paradigm, we believe that terminology has much to contribute to this field. In the past, terminology work was extensively applied to the identification of terms and relations for their subsequent transformation into concepts and conceptual relations in ontologies (Velardi et al., 2001; Aussenac-Guilles and Sörgel, 2005; Maynard et al., 2008; to mention just a few). Currently, works on terminological variation may play a significant role in the Linked Data linking step.

Linked Data refers to a set of best practices for publishing and connecting distributed data on the Web with the idea of transforming it into a *global graph*. For this purpose, data must be previously structured according to graph-based models in the form of ontologies, using the standard RDF (Resource Description Framework) syntax. Moreover, these data or information units have to use URIs (Uniform Resource Identifiers) as their names on the Web, and follow the HTTP (Hypertext Transfer Protocol) schema so that users can look up those names and find the information related to them. Finally, data have to be connected to similar data, so that users can explore those data and discover additional data.

[1] http://linkeddata.org/

Thus, the more links an RDF dataset has to other datasets, the more useful it will be.

The linking step is the key one, but also the one that involves greater difficulties. As stated in (Heath and Bizer, 2011), it is common practice to use the property or relation `owl:sameAs` to state that one data source in an "RDF dataset A" provides the same information as another data source in an "RDF dataset B". But, is it easy to identify two data sources in different RDF datasets that mean the same? Can it be done automatically, or does it require an expert to analyze and compare the datasets? Ideally, taking into account the number of RDF datasets currently published as Linked Data[2], this task should be performed automatically.

Moreover, although the initiative of transforming data into the linked data format was initially led by English speaking countries, nowadays we find an increasing amount of RDF datasets in languages other than English that need to be linked to similar or related datasets in other languages (Gómez-Pérez et al., 2013).

The analysis of terminological variation, a cornerstone in communicative and cognitive approaches to terminology (Cabré 1995, Daille 2005, Temmerman, 2000) could contribute to the identification of terms that refer to the same ontological concept, thus attempting to integrate univocity, defended by the traditional theory (Wüster, 1979) with variation in real situations. The result of such an analysis could be used in the automatic identification of concepts that *mean the same* or that hold a certain type of relation. It could also contribute to the definition of the reasons that caused that variation, and propose alternatives to the `owl:sameAs` property to capture more fine-grained relations between data sources. Finally, from a multilingual perspective, it could also help to establish cross-lingual relations between RDF datasets in various languages.

For all these reasons, we believe that Linked Data datasets should be enriched with terminological variants, as well as with other types of lexical and linguistic information as proposed in (McCrae et al., 2011; Gracia et al., 2012), so that further processes in the Linked Open Data Cloud construction – specifically the linking step – become smoother and more reliable. Such enrichment could also be very profitable due to its potential exploitation by linked data-driven Web applications.

In this contribution we provide a classification of lexical, terminological and semantic variants that has been proposed within the framework of the W3C Ontology-Lexica Community Group[3] to enhance a model of linguistic descriptions intended to enrich domain ontologies and RDF datasets. The model being designed in this framework relies on previous computational models of linguistic description, such as LMF (Francopoulo, 2013; ISO 24613), SKOS (Miles et al., 2005), or, fundamentally, the *lemon* model (McCrae et al., 2011).

Basing on works that analyse the causes of denominative variation in communicative approaches to terminology, in section 2 we revisit previous classifications of terminological variants in the light of the Linked Data paradigm. In other to justify the proposed classification, in section 3 we provide examples of modelling solutions for the different types of variants (lexical, terminological and semantic variants). We compare the mechanisms provided by available models (SKOS) to represent such variants, in contrast to the richer, more complex model of linguistic descriptions that is being proposed in the W3C Ontology-Lexica Community Group and that takes as starting point the *lemon* model. Finally, in section 4, we present some concluding remarks and discuss some further lines of research.

2 Revisited classification of term variation

As suggested in Cabré (2008), term variants that refer to one and the same concept can be divided into two types: (1) Term variants that are *semantically coincident but formally different*, i.e., terms that mean the same but are expressed by different lexical forms, generally known as synonyms (e.g., eczema vs. skin rash); and (2) Term variants that are *semantically and formally different*, since each one is highlighting one facet or dimension of the same concept (e.g., hospital waste vs. biomedical waste), so that they *do not mean exactly the same*, but refer to the same concept or real world entity. The same author refers to the latter variants as *partial synonyms* and leaves open the question of whether the two terms should point to the same concept or each

[2] http://datahub.io/

[3] http://www.w3.org/community/ontolex/

term should point to a different concept, with many assumed commonalities between the two.

This discussion becomes highly significant in view of a model that is designed to associate complex linguistic descriptions to conceptual structures (ontologies, RDF datasets), because it informs how lexical and terminological descriptions of the concepts are represented. If the conceptual structure is already given and contains that conceptual difference (let us say that it makes a distinction between *biosanitary waste*, in general, and *hospital waste*, only for the waste produced in hospitals), the two terms will most probably be associated to two different concepts. Conversely, if only one concept is represented in the ontology, we may still want to account for both terminological variants in the linguistic model, and explicitly state the motivation behind each denomination. In this way, we would also facilitate the linking of this data source to another data source contained in a different dataset and to which only the term *biosanitary waste* has been associated.

The classification we propose is motivated by the causes that provoked the variation, and has been inspired for the terminological part on the work by Freixa on denominative variation in terminology (2006). In this case, we distinguish between lexical, terminological and semantic or cognitive variants. Each type of variant will be devoted a sub-section below.

2.1 Lexical variants

For the purposes of this work, lexical variants are defined as those variants that are *semantically coincident but formally different*, and which are mainly motivated by grammatical requirements, style (*Wortklang*), and linguistic economy (helping to avoid excessive denominative repetition and improving textual coherence)[4]. As Freixa (2006: 61) maintains for acronyms and reductions of terms, this lexical variation has a high level of conceptual equivalence. Also, the use of one variant over the other does not really change the intention of the message, but it is rather caused by formal aspects of the text.

The following types can be mentioned:

- Orthographic variants
  - Diatopic variants (e.g., localize vs. localise)
  - Diachronic variants (e.g., different scripts for languages such as Azeri)
  - Ideographic variants (e.g., in Japanese both “寿司” and “鮨” are used for sushi)
- Affixal variants
  - Derivational variants (e.g., adjective -> adverb variation: quick vs. quickly; verb -> noun: activate vs. activation)
  - Inflexional variants (e.g., adjective agreement: *rojo, roja, rojos, rojas*)
- Morphosyntactic variants
  - Compounds (e.g., ecological tourism vs. eco-tourism)
  - Abbreviations (including acronyms, among others. E.g., peer to peer and p2p; WYSWYG, FAO, UNO, etc.)
  - Rephrasing variants (e.g., immigration law vs. law for regulating and controlling immigration)

2.2 Terminological variants

As for terminological variants, we understand those variants that are *not only formally, but also semantically different*, and this difference is intentionally caused. As stated in Diki-Kidiri (2000:29 and ff), in order to better understand this type of variants, it may be useful to make a distinction between concept and meaning or sense (*le signifiant, le signifié* and *le concept*), since we could say that these terminological variants refer to the same concept but they represent “the multiple specific perceptions of the same object”.

In this type of terminological variants, the denomination or term itself is a clear indicator of the reasons or causes for variation. As mentioned in Freixa (2006), these reasons can be the origins of the authors, in the case of *diatopic* variants; the different communicative registers, in the case of *diaphasic* variants (also termed functional

[4] This type of variants have been thoroughly analysed by Jacquemin (1997) mainly for French.

variants); the stylistic or expressive needs of the authors, as for the so-called *diastratic* variants; and the different conceptualizations, approaches or perspectives underlying them, in what we have termed *dimensional* variants (dubbed cognitive variants in Freixa (2006)).

Regarding the latter ones, we would like to emphasize that it is more common than not to find different conceptualizations of the same domain when different groups approach the same area of knowledge from different perspectives or with different needs. Because of that, some terms may highlight certain properties of a concept, which are not so relevant for other users. This is even more obvious in a multilingual context, in which different geographical, cultural and social groups comprehend reality in different ways. In this sense, we have included a subtype of dimensional variants called *cross-lingual dimensional variants*.

Finally, we would like to refer to the cross-lingual variants. It could be argued that these are not terminological variants strictly speaking, but translations. However, we have decided to consider them a subtype of variants with the aim of covering those scenarios in the Linked Data context in which datasets in different natural languages have to be linked and this linkage becomes essential in this new paradigm.

We have identified two types of cross-lingual variants. First, we include translations, in the general sense. It is widely accepted that original and target cultures have segmented and categorized a bit of reality in a very similar way and have a similar concept and equivalent term to refer to it. We do not account for the reasons for this similarity (it may be that one culture has imported not only the concept but also the term by providing a loan translation, etc.). However, we account for the case in which the target culture has no equivalent concept and describes the concept of the original culture and/or directly reuses the foreign term.

Finally, it is worth mentioning that we could also have *cross-lingual diaphasic variants*, if one language uses the scientific term in all registers, and the other language has two terms: one for an expert-to-expert communication situation and another term for an expert-lay communication situation (e.g., huesos metacarpianos in Spanish vs. *Ossa metacarpi* and *Mittelhandknochen* in German). The same could happen in the case of diachronic and diastratic variants.

In any case, it is important to consider that in a multilingual scenario these terminological variants would be pointing to the same concept or conceptual structure, or even share the same conceptualization. This is one of the main differences compared to the variants in section 2.3, namely, the so-called semantic or cognitive variants.

So, we have classified terminological variants as follows:

- Diatopic (dialectal or geographical variants) (e.g., gasoline vs. petrol)
- Diaphasic (register) (e.g., headache and cephalalgia; swine flu and pig flu and H1N1 and Mexican pandemic flu)
- Diachronic (or chronological variants) (e.g., tuberculosis and phthisis)
- Diastratic (discursive or stylistic variants) (e.g., man vs. bloke)
- Dimensional variants: the terms point to the same concept but highlight a different property or dimension of the concept (e.g., biosanitary waste vs. hospital waste; Novel Coronavirus vs. Middle East Respiratory Syndrome Coronavirus; obsolete technology vs. dangerous technology[5])
    - Cross-lingual dimensional variants: the concept exists in both cultures, but the terms highlight different aspects of the concept or approach it from different perspectives (e.g., *madre de alquiler* (lit. rental mother) in Spanish vs. *mere porteuse* (carrier mother) in French vs. surrogate mother in English.
- Cross-lingual variants
    - Translations (e.g., the translation *nogomet* instead of the loanword *fudbal* for soccer in Serbo-Croatian)

[5] From Freixa (2006: 68)

- Descriptions or glosses (when the concept does not exist in the target language and a literal translation or gloss is used) (e.g., École normal and French Normal School, Panetone vs. Panetone, Italian Christmas cake)

2.3 Semantic or Cognitive Variants

Semantic or cognitive variants are mainly caused by different conceptualization and/or motivations. We could say that these term variants are *semantically and formally* different, as in the case of terminological variants, but they usually point to two closely related, but different, ontological concepts, which means that they are also *conceptually different* (Aguado-de-Cea and Montiel-Ponsoda, 2012).

Such variants are commonly found at a multilingual level, but we can also found them in monolingual contexts. Let us imagine the case of an ontology or dataset that contains the concept *religious building*, and another ontology that contains the concept of *mosque*. At the linguistic level we could say that religious building and mosque are in a relation of hypernymy-hyponymy (one concept is subsumed by the other, but they are referring to two different concepts included in two different conceptual structures that have a different granularity level).

Here we also distinguish between vertical (general-specific) or horizontal variants. Vertical variants are defined as those variants that refer to concepts that share most properties, but one is more specific than the other (they are not at the same level in a classification tree, but one is more general and the other more specific. See the example of river in English vs. *rivière* and *fleuve* in French). In the case of horizontal variants, we refer to those terms that point to concepts that share most properties, but one includes properties that the other does not, and vice versa. As a result of these divergences, terms will be pointing to two different concepts in two different conceptual structures at the same level of specificity in a classification tree, but including unequal properties. Therefore, we can consider them counterparts or closest equivalents.

Within this group we find the following types:

- Vertical (general-specific) variants (e.g., benign neoplasms vs. benign mouse skin tumours)
    - Cross-lingual vertical variants (e.g., river in English vs. *rivière* and *fleuve* in French; *testamento* in Spanish vs. testament and last will in English)

- Horizontal variants (counterparts or closest equivalents):

    - Cross-lingual horizontal variants (e.g., Prime Minister in English vs. *Presidente del Gobierno* in Spanish)

3 Modelling examples

Incorporating all this terminological knowledge in ontologies is important if we aim at optimizing the linking process. Thus, distinguishing different forms of term variation turns into a key issue, when we model terminology for practical applications. In the context of linked data this means that we will model the data by means of an existing model such as SKOS or *lemon*. In this section we present practical modelling examples for kind of term variation.

3.1 Lexical variants

Lexical variants are modelled by either the multiple forms of the same entry or by means of relationships between lexical entries. For example in *lemon*, we would model *orthographic variants* as different representations of the same form[6]:

```
:myExampleLexicon a lemon:Lexicon ;
     lemon:language "en" ;
     lemon:entry :theatre_lexicalentry .

example_ontology:Theatre
   lemon:isReferenceOf [
      lemon:isSenseOf :theatre_lexicalentry].

:theatre_lexicalentry a lemon:LexicalEntry ;
    lemon:canonicalForm [
         lemon:writtenRep "theater"@en-us ;
         lemon:writtenRep "theatre"@en-gb ] .
```

In the example above, a monolingual lemon lexicon is defined which contains a lexical entry *:theatre_lexicalentry*. The concept "Theatre" in a

[6] We are using turtle RDF notation in our examples.

certain ontology is the reference of such lexical entry, which has two associated written representations: “theater” and “theatre” for American and British English respectively.

Alternatively, two different lexical entries could have been defined for each different representation. In that case, a relation can be defined between the lexical entries in *lemon* in this way:

```
:theatre_lexicalentry a lemon:LexicalEntry ;
  lemon:form [
    lemon:writtenRep "theatre"@en-gb ] .

:theater_lexicalentry a lemon:LexicalEntry ;
  lemon:form [
    lemon:writtenRep "theater"@en-us ] .

:theater_lexicalentry :diatopicVariant
  :theatre_lexicalentry .

:diatopicVariant rdfs:subPropertyOf
  lemon:lexicalVariant  .
```

In this example “theater” and “theatre” are associated to two different lexical entries. They are linked by a new relation *:diatopicVariant*, which is defined as a subtype of a *lemon* lexical variant.

Notice that *:diatopicVariant* does not exist in the *lemon* model as such, but it can be defined as in the example or, alternatively, an external category could be used, such as *"diatopical"* included in ISOCAT[7].

Morphosyntactic variants are also represented as links between lexical entries, as there may be differences in the syntactic properties of the entries. Let us consider the term “peer to peer” and its abbreviated form “p2p”:

```
:p2p a lemon:LexicalEntry ;
  lemon:form [ lemon:writtenRep "P2P"@en ] .

:peer_to_peer a lemon:LexicalEntry ;
  lemon:form [
    lemon:writtenRep "Peer-to-peer"@en ] .

isocat:fullFormFor rdfs:subPropertyOf
  lemon:lexicalVariant .
isocat:initialismFor rdfs:subPropertyOf
  lemon:lexicalVariant .

:peer_to_peer isocat:fullFormFor :p2p .
:p2p isocat:initialismFor :peer_to_peer .
```

In the previous example *:p2p* and *:peer_to_peer* are lexical entries with their respective written representations. Then, ISOCAT categories are defined as lexical variants and used to relate or link both lexical entries[8].

We can also use SKOS for representing lexical variants. In that case, we can show two preferred labels for different dialects of a language as follows:

```
:theatre skos:prefLabel “theater”@en-us  ,
                        “theatre”@en-gb.
```

However, to represent morphosyntactic variants it is necessary to use the extended label model (SKOS-XL) as follows, but, otherwise, it is similar to *lemon*, where we define a named label entity for each label and represent the link between them as a triple.

```
example_ontology:P2P
  skosxl:prefLabel :p2p_label ,
  skosxl:altLabel :peer_to_peer_label .

:p2p_label skosxl:literalForm “p2p”@en .

:peer_to_peer_label skosxl:literalForm
  “peer to peer”@en .

:P2P_label lexinfo:abbreviationFor
  :peer_to_peer_label .
```

3.2 Terminological variants

As terminological variants maintain the meaning of the term while changing the surface form, it is necessary to distinguish between the syntactic and semantic level of the term. For this reason, we characterize terminological variants as links between different senses not between lexical entries. In *lemon* this is easy to model as can be seen in the following example:

```
biontology-icd:011
  lemon:prefRef :tuberc_sense ,
  lemon:altRef :phthisis_sense .

:tuberculosis_sense lemon:isSenseOf
  :tuberculosis_lexicalentry .
:phtisis_sense lemon:isSenseOf
  :phthisis_lexicalentry .

:tuberc_sense lexinfo:dating lexinfo:modern .
:phthisis_sense lexinfo:dating lexinfo:old .
```

[7] http://www.isocat.org/rest/dc/3669

[8] For readability, we have substituted the original identifiers of the ISOCAT categories by descriptive ones. The originals are: http://www.isocat.org/datcat/DC-321 and http://www.isocat.org/datcat/DC-329

In the example above, an entity of *biontology-icd*[9] has been associated with two *lemon* senses. Such senses constitute the bridge between the ontology entities and their respective lexical entries. Then, the temporal dimension ("modern" vs. "old") can be established as an attribute of the senses (by using the LexInfo vocabulary (Cimiano *et al*, 2011) in our example).

Alternatively, a relation at the level of senses can be established in *lemon* in this way:

```
:diachronicVariant rdfs:subpropertyOf
  lemon:senseRelation .
:phthisis_sense  :diachronicVariant
  :tuberculosis_sense .
```

However, there is no way to make this distinction in SKOS and this will lead inevitably to confusion about the syntactic and semantic layer.

3.3 Semantic or cognitive variants

Cognitive variants are distinct but closely related meanings of a word. So, we can model the variation not only as a relationship between words but also as described by a semantic model, i.e., an ontology. As such, we would state OWL axioms to describe the relationship, as illustrated in the example below. There, Chancellor of Germany and Prime Minister of Spain are both subclasses of the concept Head of government.

```
dbpedia:Chancellor_of_Germany rdfs:subClassOf
  dbpedia:Head_of_government .
dbpedia:Prime_Minister_of_Spain
  rdfs:subClassOf dbpedia:Head_of_government.
```

At the lexicon level we could also establish a relation of horizontal variants between the two terms. This relation is established because we know that the two terms "Chancellor of Germany" and "Prime Minister of Spain" are not equal but can be considered similar (or counterparts) in the two cultural settings, as they have a close antecedent concept, "head of government".

4 Conclusions

As mentioned in this paper, the Linked Data initiative needs to find ways of linking the huge amount of structured datasets found on the Web in the same or in different languages. We believe that although ontologies aim at achieving univocity in as much as traditional terminology did, the more sociolinguistic cognitive approaches to terminology can also contribute to enrich the current computational models of linguistic descriptions. With this purpose, we have revisited previous classifications of term variants in the light of the Linked Data initiative so as to facilitate the process of recognition of terminological variation. We have proposed a classification of term variants in three wide groups: lexical variants, terminological variants and semantic or cognitive variants. We have also illustrated this classification with the corresponding examples at the ontology level by resorting to different ontology representation models, such as *lemon*. With the solutions proposed we also aim to enrich the linguistic ontology models as well as to make them more reliable when applied to Linked Data.

[9] http://purl.bioontology.org/ontology/ICD9CM/

Acknowledgments

This work has been supported by the BabelData (TIN2010-17550) Spanish project and the FP7 European project Monnet (FP7-ICT-4-248458).

References

Aussenac-Gilles, N. and D. Soergel. 2005. Text analysis for ontology and terminology engineering. In Journal of Applied Ontology (1), 35 – 46.

Cabré, M.T. 1995. On diversity and terminology. Terminology 2(1), 1-16.

Cabré, M.T. 2008. El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología (I). IBÉRICA 16: 9-36.

Cimiano, P., P. Buitelaar, J. McCrae, and M. 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.

Daille, B. 2005. Variations and application-oriented terminology engineering. Terminology 11(1): 181-197.

Diki-Kidiri, M. 2000. Une aproche culturelle de la terminologie. Terminologies nouvelles 21 (Terminologie et diversité culturelle): 27-31.

Francopoulo, G. (ed.) 2013. LMF Lexical Markup Framework, Wiley- ISTE.

Gómez-Pérez, A. Vila-Suero, D. Montiel-Ponsoda, E. Gracia, J. and G. Aguado-de-Cea. 2013. Guidelines for multilingual linked data. In Proceedings of the 3rd International Conference on Web Intelligence,

Mining and Semantics (WIMS'13), ACM, New York.

Gracia, J. Montiel-Ponsoda, E. Cimiano, P. Gómez-Pérez, A. Buitelaar, P. and J. McCrae. 2011. Challenges for the multilingual Web of Data. In Web Semantics: Science, Services and Agents on the World Wide Web 11: 63-71.

Heath, E. and Ch. Bizer. 2011. Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, (1)1: 1-136. Morgan & Claypool.

ISO 24613 - LMF- Lexical Markup Framework, Language Resource Management. 2006. Available from http://lirics.loria.fr/doc_pub/LMF%20rev9%2015 March2006.pdf

Jacquemin, C. 1997. Variation terminologique: reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Memoire d'habilitation a diriger des recherches en informatique fondamentale, Universite de Nantes.

Maynard, D., Y. Li, and W. Peters. 2008. NLP Techniques for Term Extraction and Ontology Population. In Buitelaar, P. and P. Cimiano, (eds.), Ontology Learning and Population: Bridging the Gap between Text and Knowledge, IOS Press, Amsterdam: 171-199.

McCrae, J., G. Aguado de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez- Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. 2012. Interchanging Lexical Resources in the Semantic Web. Language Resources and Evaluation 46, (4): 701-719.

Miles, A., B. Matthews, D. Beckett, D. Brickley, M. Wilson, and N. Rogers. 2005. SKOS: A language to describe simple knowledge structures for the web. In Proceedings of the XTech Conference.

Temmerman, R. 2000. Towards new ways of terminology description. The Sociocognitive Approach. John Benjamins, Amsterdam.

Velardi, P. Missikoff, M. and R. Basili. 2001. Identification of relevant terms to support the construction of Domain Ontologies. In Proceedings of ACL-01 Workshop on Human Language Technologies, Toulouse, France, 2001.

Wüster, E. 1979. Introduction to the General Theory of Terminology and Terminological Lexicography, Springer: Vienna.

Using ISO and Semantic Web standards for creating a Multilingual Medical Interface Terminology: A use case for Heart Failure

Elena Cardillo	Maxime Warnier	Joseph Roumier	Marc Jamoulle	Robert Vander Stichele
Fondazione Bruno Kessler, Trento, Italy	Université Catholique de Louvain Louvain, Belgium	CETIC Charleroi, Belgium	IRSS-Université Catholique de Louvain Louvain, Belgium	Heymans Institute of Pharmacology Ghent, Belgium
cardillo@fbk.eu	maxime.warnier@student.uclouvain.be	joseph.roumier@cetic.be	marc.jamoulle@uclouvain.be	Robert.VanderStichele@UGent.be

Abstract

The correct registration and encoding of medical data in Electronic Health Records is still a major challenge for health care professionals. Efficient terminological systems are lacking to enable multilingual semantic interoperability between general practitioners, patients, medical specialists, and allied health personnel. The aim of this paper is to propose an architectural structure for a Multilingual Medical Interface Terminology. We propose a dual structure with a multilingual reference terminology and a collection of unilingual end-user lexicons. Our methods rely on terminological standards, such as Terminology Markup Framework (ISO 16642) and Lexical Markup Framework (ISO 24613), and on Semantic Web technologies. We present procedures to select words, phrases, and concepts to populate these resources (manual concept extraction, automated term extraction), to link them to NLP applications and international classifications. We present the publication of these resources in Linked Open Data and show the feasibility of the approach in a use case with terms related to heart failure in several languages. We illustrate in particular the difficulties in linking real life concepts (N=168) to multiple international classifications with different functionalities, level of granularity, and scopes.

1 Introduction

In spite of the progress reached in the field of Medical Informatics, it is still difficult for Health care professionals and other health actors to register and codify clinical data in daily practice. The link between Electronic Health Records (EHRs) and international English classification systems used to codify the clinical data is often complex, not well integrated and hampered by a translation gap, both in terms of language and of world of reference (patient, general practitioner, medical specialist, etc.).

There is an increased awareness that attaining the goal of semantic interoperability entails construction of interface terminologies, defined by (Rosenbloom et al., 2006) as a ‘‘systematic collection of healthcare-related phrases (terms) to support clinicians’ entries of patient-related information into computer programs such as clinical “note capture” and decision support tools, facilitating display of computer-stored patient information to clinician users as simple human-readable texts’’. These kinds of interface terminologies can be used for problem list entry, clinical documentation in EHRs, text generation and care provider order entry with decision support.
The question is whether existing interface terminologies are sophisticated enough to support semantic interoperable communication of the clinical data between partners in a multilingual health care system. Until now, interface terminologies have been either limited to one language (often English), or either providing an

interface to only one nomenclature (e.g. SNOMED) or classification (e.g. International Classification of Diseases).

The aim of this paper is to present the architectural structure for a multilingual medical interface terminology, following both lexical and terminological ISO standards on multilingual terminologies and using Semantic Web technologies and languages (RDF/OWL, SPARQL, Linked Data, etc.). The ambition of this multilingual resource for general practitioners, patients, medical professionals, and allied health personnel is to span the gap between human language (as addressed in Natural Language Processing systems) and machine language (used to manage concepts and their lexical representations) and to map to a variety of well-respected international medical nomenclatures, thesauri and classification systems).

The paper is structured as follows: Section 2 gives an overview of the state of the art in the field of Medical Terminology. Section 3 describes the approach to building the structure for a hybrid interface terminology. Section 4 is devoted to present a use case on Heart Failure and preliminary results. Finally, Section 5 provides some discussion and conclusions.

2 Background

Over the last two decades, research on medical terminologies and classification systems has become a popular topic and much work has been done to map between several nomenclatures (UMLS[1], SNOMED-CT[2]), thesauri (MeSH[3]), and classification systems (ICPC[4], ICD[5]), different in structure and purposes and used by physicians during their patients' health care visits.

A number of studies are based, for example, on the extensive use of UMLS as a knowledge resource for extracting semantic mappings (see Fung and Bodenreider, 2005).

Recently, the use of Semantic Web[6] technologies in the biomedical domain has leaded to promising results in terms of information integration across heterogeneous resources. Examples are the use of Resource Development Framework (RDF), triple stores and SPARQL queries in integrating consumer-oriented terminologies with standard classification systems in UMLS (Cardillo et al., 2012), or for aligning standardized nomenclatures with thesauri (Bodenreider, 2008). Many tools have been created for automatic alignment between resources. Two examples are TAXOMAP[7] and ONAGUI[8], more useful to ontology alignment. Results of these kind of alignment algorithms are recorded in file in RDF or OWL[9] language.

The major use of Semantic Web technologies in healthcare is the formalization of existing medical terminologies or classification systems in ontologies. SNOMED-CT (Rector and Brandt, 2008) and the upcoming ICD-11 (Tudorache et al., 2010) represent a good example of this practice. Important is also the creation of medical ontology repositories, such as Bioportal[10], a Web-based open repository where users can search and browse biomedical ontologies as well as create mappings for them.

The use of ontologies in healthcare is very useful, above all if integrated in EHRs or in Personal Health Records (PHRs), since they give structuring and semantics to the recorded information. With predefined mappings of vocabularies used in the original data, they also allow for aggregation, reuse of knowledge, automated reasoning on data, search and retrieval of data from diverse original source systems.

Finally, a number of initiatives have been launched for the creation of consumer-oriented medical vocabularies[11] that map to standardized terminologies or classification systems. Examples of this challenge are: the Open Access Collaborative Consumer Health Vocabulary for English, developed by (Zeng et al., 2006) and available in the UMLS Metathesaurus; the Italian Consumer Medical Vocabulary (ICMV), developed by (Cardillo, 2011) using a lexi-ontological approach, and the Multilingual Glossary of Popular and Technical Medical Terms, in nine European languages[12], developed under initiative of the European Commission.

Crucial missing links in this field are the lack of application in the domain of medical terminology of two ISO terminological standards,

[1] http://www.nlm.nih.gov/research/umls/
[2] http://www.ihtsdo.org/
[3] http://www.ncbi.nlm.nih.gov/mesh/
[4] http://www.globalfamilydoctor.com/wicc/sensi.html
[5] http://www.who.int/classifications/icd/en/
[6] http://www.w3.org/standards/semanticweb/
[7] https://www.lri.fr/~hamdi/TaxoMap/TaxoMap.html
[8] http://onagui.sourceforge.net/
[9] http://www.w3.org/TR/owl-features/
[10] http://bioportal.bioontology.org/
[11] Lexical resources that reflect the way consumers/patients express and think about health topics.
[12] http://users.ugent.be/~rvdstich/eugloss/information.html

namely Lexical Markup Framework (LMF, ISO 24613, 2008) and Terminological Markup Framework (TMF, ISO 16642, 2003), the absence of a multilingual approach.

3 Building the architectural structure for a hybrid Interface Terminology

We propose a Medical Interface Terminology, that is composed of two types of domain-specific resources: a multilingual reference terminology (linked to international classifications, thesauri and nomenclatures), on the one hand, and unilingual end-user lexicons the other hand.

To create a comprehensive interface terminology to be used in the primary domain for information storage, encoding, translation, and retrieval, we used a hybrid approach that consists in the combination between onomasiological and semasiological approaches. On one hand we build the structure of the reference terminology starting from essential concepts of the domain and then look for their lexical representations in different international classifications. On the other hand, we build also the structure of an end-user lexicon following the opposite approach, so starting with a word (or phrase) and looking for its different meanings, and in particular to which concept in the onomasiological resource the word refers to. As mentioned in Section 1., to build this interface terminology we advocate our choice for standard frameworks, using the ISO norms on Terminological Markup Framework (TMF) and on Lexical Markup Framework (LMF), whose meta-models perfectly fit with our requirements. For each type of resource of the interface terminology, we explain our approach to apply the mentioned standards, and to build the two recourses with data categories and domain values from the ISOcat.org platform[13], a data category registry (ISO 12620, 1999). Finally, we describe the approach for publishing these resources as Linked Open Data (LOD).

3.1. Creating a TMF based Reference Terminology

A reference terminology comprehensively and rigorously defines reference concepts and expressions within the biomedical domain, including interrelations between concepts (Rosenbloom et al., 2006), and provides a common reference point for comparisons and aggregation of data about the entire health care process, recorded by multiple different individuals, systems or institutions. It allows the concepts to be defined in a formal and computer-processable way and to be mapped to existing standard nomenclatures and classification systems. This allows to support consistent and understandable coding of clinical concepts and so is a central feature for use in EHRs.

The first resource we created was the multilingual reference terminology, following the standard TMF (Romary et al., 2006), conceived to structurally address multilingualism. The TMF meta-model, which keeps the traditional onomasiological view of a terminological entry, decomposes the organization of a terminological database into five basic components: i) the terminological resource (Terminological Data Collection); ii) the concept (Terminological Entry); iii) the language chosen (Language Selection); iv) the term(s) in the language chosen (Term Entry); v) components of the term (Term Component Section).

The meta-model is ornated with a subset of internationally accepted data categories, relevant to the functionalities of the reference terminology (e.g., entrySource, languageIdentifier, preferredTerm) that have been selected from the already mentioned ISOcat.org platform. Detailed explanations on the TMF meta-model and on the selected data categories can be found in (Roumier et al., 2011).

In a further step, we created a Terminology Markup Language (TML) that comprises the mappings between the Meta-model, the data categories and their domain values. The resulting TML has been serialized using the XML schema language RelaxNG[14]. For this part of the work, we were inspired by the TML created in the TermSciences[15] project (Khayari et al., 2006).

Our aim is to keep the number of core concepts to a respectable minimum (between 7.000 and 15.000 concepts). For each concept, the level of granularity is assessed. Pre-coordinated concepts are kept to a minimum (based on frequency-of-use criteria).

In a first step at the Terminological Entry level (language independent), selected concepts are incorporated in the TMF resource. To each concept a definition in English is assigned and if possible, a perfect match for that concept is

[13] http://www.isocat.org/

[14] http://w ww.relaxng.org/

[15] http://www.termsciences.fr/

looked for in the SNOMED-CT nomenclature, or in the UMLS Metathesaurus. Otherwise, the concept is genuinely defined within the system. Each concept in the reference terminology is categorized according to medical categories (symptom, disease, medical procedure, body part, etc.). In addition, to ensure semantic interoperability, links to international terminologies (SNOMED-CT and UMLS) and classification systems (ICD 10th revision, ICD-10, and ICPC 2nd edition, ICPC-2) are provided, with the corresponding qualitative nature of the mapping (exact match, nearly exact, match to higher or lower level of granularity, match not possible).

In a second step at the Term Entry level (language dependent) the concept is labeled with one preferred term (the most suitable clinical term used by physicians to coin the concept in this language) and one admitted term (the most suitable lay term used by consumers and patients to coin the concept in that language) for each participating language. In case the concept has a SNOMED-CT exact match, a "standardized term" is also given (literal translation of the fully specified SNOMED-CT name).

We made a deliberate choice not to represent hierarchical (broader- narrower) links or links between related concept into our system, nor any other attempt to self-generated ontological mapping. We decided to rely on the mappings to the international nomenclatures and classifications to explore the semantic relations between the concepts in our resource, and not to invest energy in the possibly redundant activity of creating a new ontology.

3.2. Creating an LMF based end-user lexicon

The second type of resource in the interface terminology is a series of unilingual end-user lexicons that must be linked to the multilingual reference terminology described above. These end-user lexicons (one for each language) can be linked to Natural Language Processing (NLP) applications, and can be oriented to patients and to professionals.

To represent the end-user lexicons, we used the ISO standard LMF[16], a model that provides a common standardized framework for NLP lexicons. This model deals with linguistic complexities, and, as TMF, uses the ISOcat source for the association of linguistic data categories (e.g., partOfSpeech, namedEntity). Furthermore it can be linked to TMF and other concept based representation systems. Resources in LMF can also be linked to existing lexical NLP resources (such as WordNet). LMF meta-model contains a mechanism to deal with multiple senses (Francopoulo et al, 2007).

Lexical entries of the end-user lexicon are words (or phrases) that are selected from everyday interactions between doctors and patients (occurrences in medical records, in guidelines, in web consultations, etc.), based on frequency count and relevance. Various methodologies for human or automated term extraction can be used. For each of the selected words (or phrases), the possible senses (i.e. medical and non-medical ones) are clearly defined and entered as such in the LMF end-user lexicon of the originating language.

We propose a dual mechanism to link the first type of resource, the multilingual reference terminology and the end-user unilingual lexicons:

- Each concept in the reference terminology is linked to the sense part of a word in the lexicon. This mechanism preserves the conceptual integrity and is language independent.
- Each lexical representation, in a specific language, of a concept in the reference terminology, is linked to the corresponding lemma in a end-user lexicon of that language. This is language dependent.

In the unilingual end-user lexicons, links to synonyms for selected words (or phrases) in the selected sense can be provided. At this level, also subtle differences between related languages (e.g. Portuguese in Portugal and in Brazil, English in the UK and in the US, Dutch in the Netherlands and in Belgium) are addressed.

3.3. Publication in Linked Open Data

The TMF model of the reference terminology, described in Section 3.1, is implemented as an OWL-DL ontology, which is an efficient way of defining the components (represented as hierarchically organized classes) and the vocabulary (consisting of data and object properties that can be easily reused in other data sets) and ensuring the consistency of the data.

Terminological entries are represented as classes, while data categories are represented as

[16] http://www.lexicalmarkupframework.org/

OWL object properties. Classes, whenever possible, are linked to similar concepts in other international recognized classifications, available on the web as Linked Data, using the owl:equivalentClass property (which, in this case, should be preferred to owl:sameAs to avoid undesired effects when using reasoners (Halpin et al., 2010). Because a one-to-one correspondence cannot always be found between all the classifications, in some cases entries can be grouped in more general categories with the owl:unionOf property.

The conversion of the TMF resource in OWL/RDF allows for the publication of the data as Linked Data[17] on the Semantic Web that are accessible via SPARQL queries. Linked Data principles encourage reuse, reduce redundancy, maximize its real and potential interconnectedness, and finally enable network effects to add value to data (Bizer et al., 2009).

The end-user lexicons will be published according to the LEMON framework[18], which is a proposed model for modeling lexicons and machine-readable dictionaries based on LMF and, similarly to our TMF terminology, using Semantic Web technologies and ISOcat data categories (McCrae et al., 2012). The functionality to publish the content of the lexicons in Linked Open Data is an established part of the LEMON framework.

4 A use case for Heart Failure

To populate the structure of our interface terminology with a test sample, we have chosen to extract relevant concepts and words or phrases from a Belgian bilingual (Dutch and French) guideline on Heart Failure for general practitioners, published by the Scientific Associations of Primary Care Physicians (Van Royen et al., 2011). For this study we worked with the French version as a starting point. We describe a manual concept extraction and an automated term extraction, along with the procedures to populate both the TMF and LMF-LEMON resources.

4.1. Concept extraction to the TMF resource.

A general practitioner, expert in medical classifications, analyzed the French version of mentioned guideline, and after a careful reading and tagging, selected 168 concepts, relevant for the clinical domain of heart failure and pertaining to the reference world of general practice.

In a first step, all identified concepts were entered in the TMF resource, at the Term Entry section of the French language section.

A definition in French was given for each concept, together with a reference to the French guideline from which the concept was extracted. For each concept, a preferred term (representing the technical term used by physicians) and one or more admitted terms (representing the lay term used by patients for that concept) were chosen. For some technical terms the corresponding lay term was a simple description of the term itself (e.g. the French admitted term "eau dans le ventre" for the technical term "ascite").

Then, in a second step the Terminological Entry level was addressed. The corresponding concepts were looked up in ICPC-2, the reference classification for general practice. Then, the cross mapping to ICD10 were sought after. Next a search in the SNOMED-CT web browser was made and finally the relevant corresponding definitions were extracted from the UMLS Metathesaurus, using a dedicated tool. In case a perfect match with a SNOMED-CT concept was possible, its Fully Specified Name was entered (as well as its French literal translation). In case the concept could not be matched in SNOMED-CT, a definition in English was sought (and the event recorded for notification to the international SNOMED-CT governance group IHTSDO).

At the end of this process, among the 168 concepts selected from the guideline, 153 were mapped to ICPC-2, 131 to ICD-10, 161 concepts to SNOMED-CT, and 116 to UMLS definitions.

Reasons for the inability to match with SNOMED-CT were the too broad and general nature of the selected concept, and mismatches between "world of references". For instance, in the guideline, the concept "sexual problems" was repeatedly used at this broad level to convey information (and thus necessary in the communication). Some concepts (e.g. drug side effect "dry mouth") were not within the scope of SNOMED-CT, as the category "Symptom and complaint" (very important for primary care) is alien to SNOMED-CT. Both UMLS and SNOMED-CT use semantic types to categorize concepts. However, the conceptual framework of

[17] http://linkeddata.org/

[18] http://www.lemon-model.net/

these semantic types is different, which also leads to difficulties in finding an exact conceptual match for locally used medical terms in international nomenclatures. With regard to mapping to ICD-10, we noticed that 15 of the 168 concepts referred to aspects of functional status (e.g. nutritional status, exercise intolerance), which is difficult to represent in ICD. This provoked difficulties to map to this classification (oriented towards morbidity and mortality classification). Similarly, a number of terms related to medical procedures could not be mapped. Regarding ICPC-2, some of the selected concepts were mapped to more than one ICPC-2 rubric (e.g. "Alcoholic cardiomyopathy" mapped to "heart disease other" (K84) and "chronic alcohol abuse" (P15)) and in most cases they were mapped to broad rag bag ICPC-2 rubrics as in the case of the concept "maladie de Paget" mapped to the ICPC-2 rubric "T99 - Musculoskeletal other diseases" .

For each concept, also the Italian language section was populated with a translation of the preferred term and of the admitted term. This was performed with the help of Italian speaking domain experts and a terminologist, sustained by mapping to an existing Italian medical vocabulary oriented to healthcare consumers (the already mentioned ICMV). A similar work is to be made for English and Dutch language sections, and further expanded as described above.

4.2. Term extraction to the LMF resource.

The two versions (Dutch/French) of the heart failure guideline (30 pages of text) were submitted to the term extraction program TExSIS (Macken et al., 2013) , which provides tokenization, part-of-speech tagging, lemmatization, detection of phrases, named entity recognition, and bilingual sentence alignment. The term extraction resulted in an aligned bilingual glossary of 774 words and phrases.

After this automatic extraction, an exact matching to the French preferred terms of concepts in the TMF resource was performed. Surprisingly only 77 French preferred terms among 168 had a string match with the 774 word and phrases extracted by TExSIS. These 77 French terms were then entered in the French LMF - LEMON resource and linked to the identification number of the concepts in the TMF resource.

In addition, we considered all the 77 corresponding admitted (lay language[19]) terms from the TMF resource, resulting into an increase of entries in the LMF-LEMON resource to 138 lexical entries. For 16 admitted terms there was no difference with the French preferred term. Among the 138 lexical entries, 114 were phrases, and were then decomposed into single words, the total of entries to 298 in the LMF-LEMON resource. This rapid increase in the number of terms is however unlikely to be linear, since some words are part of several phrases (e.g. "insuffisance cardiaque" – "insuffisance rénale").

For each entry we added linguistic information, such as part-of-speech tag, lemma, spelling variants and inflected forms. The medical senses of the entries are linked to the reference terminology, and the other senses to DBpedia[20].

4.3. Publication of the TMF and LMF-LEMON resources in Linked Open Data

The tabular data representing the TMF resource on Heart Failure were automatically converted into 16.636 RDF triples, assigning a unique identifier to each element, and serializing it in RDF/XML. We published the resulting file (in French, English and Italian) on the Data Hub[21] under a Creative Commons License[22], so that it can be freely available for computer applications. The publication is provided using the Meriterm[23] server along with a SPARQL endpoint. By the combined use of URL rewriting techniques and SPARQL construct queries, triples are given a more convenient access.

[19] ''Lay'' language is the language used by patients to communicate between themselves or with physicians about their health problems, using for example ''heart attack'' instead of ''Myocardial Infarction'', as opposite to professional language which is the one used by doctors to communicate between themselves. For example CVA (Cardiovascular Accident) or TIA (Transient Ischemic attack) is useful between doctors but has to be "translated" in lay language for patient.

[20] http://fr.dbpedia.org/

[21] http://datahub.io/

[22] http://creativecommons.org/

[23] Medical Reference Interface Terminologies: http://meriterm.org

Also the LMF-LEMON resource is represented in RDF triples from the start, and hence automatically transferrable in Linked Open Data[24].
The following table show some statistics on the use case of Heart Failure.

	English	French	Italian
TMF Concepts		168	
Preferred Terms		168	168
Admitted Terms		168	168
Standardized Terms	161	161	161
UMLS definitions	116		
Links to ICD-10	131		
Links to ICPC-2	153		
TMF Triples	16,636		
LMF entries	298		
LMF Triples	3,400		

Table 1. Statistics on the Hearth Failure use case

5 Discussion and Conclusion

In this article, we presented the architectural structure of a multilingual medical interface terminology with an ambitious set of objectives. We presented a first attempt, within a small use case, to create a sophisticated, hybrid medical interface terminology, aiming at multilingual solutions, at semantic interoperability, and at open source availability as Linked Open Data in the Semantic Web. For resources as precious as language and international terminologies a proprietary approach would not be appropriate.

The value of our resource for medical communication is not tested yet in a clinical setting. Preliminary results indicate the need for a concerted quality control of the process of "words/concepts" selection. To improve results, a refinement of the mapping approach (manual, so time and cost consuming) is needed, trying to investigate semi-automatic approaches relying on Semantic Web technologies, as done in (Cardillo et al., 2012).

One big stimulus for our work is the increasing internationalization of health care (near border, cross border and intercontinental health care exchange including migrants health issues). In fact, in order to provide semantic interoperability between different healthcare information systems and between different health actors and patients, multilingual access to international terminologies is needed.

In conclusion, this terminology support system, relying on ISO standards and Semantic Web languages and tools, published as Linked Open Data supports: (i) the efficient use of existing medical terminologies and their legacy data in the activity of clinical encoding; (ii) links between professional language and lay language, and healthcare information system integration (e.g. between EHRs and PHRs); (iii) multilingualism in its core approach to semantic interoperability; (iv) information retrieval, and (v) creation of information through epidemiological research. Our interface terminology will enable physicians to find the right medical entry at the right moment at the point of care, and to integrate their data with standard classifications for their encoding, being consistent with the requirements of their Health Authority (e.g. the mandatory use in the General Practice of ICPC, or ICD as coding system for diagnoses or problem lists), and finally to exchange their data with other healthcare professionals and within various healthcare information systems in an interoperable way.

We are currently working on widening the coverage of our reference terminology selecting other use cases (e.g. contra-indications of medication) and to extend the range of participating languages.

We are also building a collaborative, web-based management platform for the extension and maintenance of the proposed Multilingual Medical Interface Terminology.

Acknowledgments

This study has been done under the MERITERM consortium, devoted to joined research activities on medical terminologies and classification systems. The authors wish to thank the consortium for the support.

References

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked data-the story so far, International Journal on Semantic Web and Information Systems, 5(3), 1-22.

Olivier Bodenreider. 2008. Comparing SNOMED CT and the NCI Thesaurus through Semantic Web Technologies. In proceedings of the 3rd International Conference on Knowledge Representation in

[24] The RDF/XML file with the lexicon and its metadata can be found online at the link: http://meriterm.org/heartfailure/lexicons.rdf/.

Medicine (KRMED2008). R. Cornet, K.A. Spackman (Eds), vol. 410, pp. 37-43.

Elena Cardillo, Germano. Hernandez, and Olivier Bodenreider. 2012. Integrating consumer-oriented vocabularies with selected professional ones from the UMLS using Semantic Web Technologies. In M. Szomszor, P. Kostkova, O. Akan, P. Bellavista, J. Cao, D. Jiannong, F. Dressler, D. Ferrari, M. Gerla, H. Kobayashi, S. Palazzo, S. Sahni, X. Shen, M. Stan, J. Xiaohua, A. Zomaya, G. Coulson (eds.), Electronic Healthcare, Proceedings of e-Health2010, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 69, Part 4, pp. 54-61, Springer Berlin Heidelberg.

Elena Cardillo. 2011. A Lexi-ontological Resource for Consumer Healthcare: the Italian Consumer-oriented Vocabulary. PhD thesis, University of Trento, Fondazione Bruno Kessler. http://eprints-phd.biblio.unitn.it/570/

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2007. Lexical Markup Framework: ISO standard for semantic information in NLP lexicons. In Proceedings of the Workshop on Lexical-Semantic and Ontological Resources of the GLDV Working Group on Lexicography at the Biennal Spring Conference at the GLDV, Tubingen, (13-14/04/2007).

Kin W. Fung and Olivier Bodenreider. 2005. Utilizing the UMLS for semantic mapping between terminologies. In Proceedings of AMIA Annual Symposium: 266–270.

Harry Halpin, Ivana Herman, and Patrick J. Hayes. 2010. When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. RDF Next Steps Workshop, Palo Alto, USA.

International Standard Organization. 1999. ISO 12620:1999 Computer applications in terminology – Data categories.

International Standard Organization. 2003. ISO 16642:2003 Computer applications in terminology - Terminological markup framework (TMF).

International Standard Organization. 2008. ISO 24613:2008 Language resource management — Lexical markup framework (LMF).

Lieve Macken, Els Lefever, and Veronique Hoste (2013). TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. Terminology, Volume 19(1):1-30.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner (2012). Interchanging lexical resources on the semantic web. Language Resources and Evaluation, 46(4):701–719.

Majid Khayari, Stéphane Schneider, Isabelle Kramer, and Laurent Romary. 2006. Unification of multilingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative, LREC 2006, International Workshop Acquiring and representing multilingual, specialized lexicons: the case of biomedicine. http://hal.archives-ouvertes.fr/hal-00022424.

Alan. L. Rector and Sebastian Brandt. 2008. Why do it the hard way? The case for an expressive description logic for SNOMED. Journal of American Medical Informatics Association, 15(6):744–751.

Laurent Romary, Isabelle Kramer, Susanne Salmon-Alt, and Joseph Roumier. 2006. Gestion de données terminologiques: principes, modèles, methods. Terminologie et accès à l'information. Widad Mustafa El Hadi (Ed.), Hermes science publ, Paris, France.

S. Trent Rosenbloom, Randolph. A. Miller, Kevin B. Johnson, Peter I. Elkin, and Steven H. Brown. 2006. Interface terminologies: Facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc, May-June, 13(3):277-288.

Joseph Roumier, Robert Vander Stichele, Laurent Romary, and Elena Cardillo. 2011. Approach to the Creation of a Multilingual, Medical Interface Terminology. Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence, 13-15.M.

Tania Tudorache, Sean M. Falconer, Csongor Nyulas, Natalya F. Noy, and Marc A. Musen. 2010. Will semantic web technologies work for the development of icd-11? In Proceedings of the International Semantic Web Conference (ISWC2010): 257–272.

Paul Van Royen, Pierre Chevalier, Gilles Dekeulenaer, Martine Goossens, Philip Koeck, Michel Vanhalewyn, and Paul Van den Heuvel. 2011. Recommandation de Bonne Pratique Insuffisance cardiaque. Domus Medica. SSMG. Belgium.

Qing Zeng and Tony Tse. 2006. Exploring and Developing Consumer Health Vocabularies. Journal of the American Medical Informatics Association,13:24–29.

Blending Two Kinds of Semantic Relatedness for Cross-language Matching of Lexical Concepts

Yoshihiko Hayashi
Graduate School of Language and Culture, Osaka University
1-8 Machikaneyama, Toyonaka 5600043, Japan
hayashi@lang.osaka-u.ac.jp

Abstract

This paper proposes a method of discovering conceptual mates in the lexical-semantic resources of other languages. The proposed method integrates two types of cross-lingual semantic relatedness by balancing them using a weighted sum: The first, *synonym-based relatedness*, estimates the probabilistic correspondence between sets of synonyms, while the other, *gloss-based relatedness*, measures the textual similarity between gloss texts. Our experimental results indicate that (1) the weighted-sum method is simple yet relatively robust; (2) the extended gloss texts gathered from the conceptual neighbors of objective concepts play a certain role in improving matching performances; and (3) the negative impact of machine translation, inevitably introduced to measure textual similarities in a language, can be substantially reduced by the translation redundancy resulting from the use of multiple translation engines.

1 Introduction

Given the recent trend toward the attainment of *dynamic lexical resources* (Calzolari, 2008) in service-oriented Web environments, a process that can efficiently, and in some cases on-demand/on-the-fly, interlink lexical elements across resources is in high demand. In particular, an inexpensive computational method for finding possible *conceptual mates* in another language (the target language (TL)) from a lexical concept in a given language (the source language (SL)) can contribute to the realization of a virtually combined multilingual lexical-semantic resource on top of existing monolingual lexical-semantic resources. Here, a conceptual mate in TL means a lexical concept that denotes almost the same or a closely related concept to the one in SL.

Given this motivation, this paper proposes a method of discovering conceptual mates in the lexical-semantic resources of other languages. To accomplish this, the proposed method integrates two types of cross-lingual semantic relatedness by balancing them using a weighted sum: The first, *synonym-based relatedness*, estimates the probabilistic correspondence between sets of synonyms, while the other, *gloss-based relatedness*, measures the textual similarity between gloss texts.

For synonym-based relatedness, this paper adopts a method recently proposed by us (Hayashi, 2012), which effectively employs a sense-tagged corpus in the TL and existing bilingual dictionaries. Conversely, for gloss-based relatedness, this paper relies on the extended gloss overlap method proposed by Banerjee and Pedersen (2003). However, because their method was developed to calculate monolingual textual similarities, prior to using them, we translate glosses in one language into another language using off-the-shelf machine translation engines.

This paper empirically discusses an optimum mean for integrating synonym-based and gloss-based semantic relatednesses, while examining the range of gloss extension. In addition, this paper investigates the impact of machine translation by comparing the performances with those achieved when presumably ideal gloss translations are available.

The proposed method can contribute to the realization of dynamically combined lexical resources in service-oriented environments (Hayashi, 2011) because it is computationally inexpensive.

2 Methodology

Although the proposed method is essentially language independent, we here set our current task as the discovery of English conceptual mates in Princeton WordNet (PWN) (Miller and Fellbaum, 2007), given a concept in the EDR electronic dictionary (EDR) (Yokoi, 1995). Note that even though the EDR is organized bilingually in Japanese and English, in principle, we only employed information given in Japanese, leaving the corresponding English information for reference purposes only. (See Appendix-A for a more detailed description of EDR.)

2.1 Cross-lingual semantic relatedness

In order to find the best-matched conceptual mates in a TL lexical-semantic resource for a given SL lexical concept, the proposed method establishes a ranked list of candidate lexical concepts in the TL (a partial example is shown in Figure 1) by computing cross-lingual semantic relatedness scores $score(s,t)$, which are defined by Formula (1). Just to be safe, s and t denote lexical concepts in SL and TL, respectively. As clearly indicated in the formula, $score(s,t)$ is computed as the weighted sum of $pscore(s,t)$ and $gscore(s,t)$; the former, synonym-based relatedness, refers to cross-lingual semantic relatedness based on synonymous words (i.e., words that jointly specify a lexical concept), while the latter, gloss-based relatedness, measures cross-lingual semantic relatedness based on the textual similarity between the (extended) glosses. Finally, $0.0 \leq \beta \leq 1.0$ indicates the blending ratio of these elements.

$$score(s,t) \equiv (1-\beta)pscope(s,t)+\beta gscore(s,t) \quad (1)$$

2.2 Synonym-based relatedness: $pscore(s,t)$

We adopt the method proposed by Hayashi (2012) to compute the synonym-based relatedness, $pscore(s,t)$. As defined by Formula (2), it is computed as a weighted sum of $pscore'(x_i,t)$, which gives the cross-lingual semantic relatedness between x_i, a word in the synonym word set $\sigma(s)$ of an SL lexical concept s, and a TL lexical concept t. In the formula, $\omega(x_i,s,t)$ is a generic function for assigning a weight to $pscore'(x_i,t)$; as stated, $\omega(x_i,s,t)$ can take x_i, s, and t into account.

$$pscore(s,t) \equiv \sum_{x_i \in \sigma(s)} \omega(x_i,s,t) \times pscore'(x_i,t) \quad (2)$$

2.2.1 $pscore'(x_i,t)$

A probabilistic interpretation is given to $pscore'(x_i,t)$, as shown in Formula (3).

$$pscore'(x,t) \equiv p(t|x) = p(t) \sum_{y_j \in \tau_W(x)} \frac{p(y_j|t)p(y_j|x)}{p(y_j)} \quad (3)$$

In the formula,

- $\tau_W(x)$ denotes a set of translation words for the SL word x;
- $p(y_j|t)$ dictates the posterior probability of the TL word y_j given a TL lexical concept t, which can be obtained from a sense-tagged corpus in TL by applying the maximum likelihood estimation; and
- $p(y_j|x)$ represents the translation probability of the TL word y_j given an SL word x, which can be estimated from existing bilingual dictionaries and/or parallel corpora.

2.2.2 $\omega(x_i,s,t)$

As proposed by Hayashi (2012), we also utilize the following formula to compute $\omega(x_i,s,t)$.

$$\omega(x_i,s,t) \equiv n\_idf(x_i) \times tset(x_i,s) \times toverlap(s,t) \quad (4)$$

The terms on the right hand side of the formula are as follows:

- $n\_idf(x_i)$, which, inspired from the normalized inverse document frequency, dictates the discriminative power of a particular word x_i in s.
- $tset(x_i,s)$, which measures the representativeness of x_i in s by computing the degree of overlap between the translation words of x_i and that of other words in s.
- $toverlap(s,t)$, which measures the similarity between s and t in terms of the degree of overlap in the sets of synonymous words in TL.

2.3 Gloss-based relatedness: $gscore(s,t)$

In this paper, we propose that gloss-based relatedness be computed by Formula (5), which incorporates the notion of extended gloss overlap first proposed by Banerjee and Pedersen (2003).

$$gscore(s,t) \equiv \frac{\sum_{(r_i,r_j)\in R} TextSim(\tau_T(r_i(s)), r_j(t))}{|R|} \quad (5)$$

```
EDR query (CID: 0f75be):
quasi Synonyms: 自負心; 自尊心; プライド (pride;self-esteem)
Glosses; 自負心 (self-confidence)
-----------------------------------------------------------
* (1) PWN synset:07508486-n, Score:0.081
Synonyms: pridefulness pride
Gloss: a feeling of self-respect and personal worth
- (2) PWN synset:05613170-n, Score:0.028
Synonyms: ego
Gloss: the conscious mind
* (3) PWN synset:07531536-n, Score:0.020
Synonyms: pride
Gloss: satisfaction with your (or another's) achievements
* (4) PWN synset:00758175-n, Score:0.016
Synonyms: superbia pride
Gloss: unreasonable and inordinate self-esteem (personified as one of the deadly sins)
* (5) PWN synset:01772498-v, Score:0.006
Synonyms: pride plume congratulate
Gloss: be proud of
- (6) PWN synset:09178999-n, Score:0.001
Synonyms: reason ground
Gloss: a rational motive for a belief or action
```

Figure 1: Example of the ranked candidate list (partial): the symbols "*" and "-" indicate that the corresponding candidates are relevant and irrelevant, respectively.

To apply the notion of gloss overlap even for the cross-lingual cases, we utilize machine translation, for which the relevant function is in the formula represented by τ_T. It is a translation function that translates a text string in SL to the corresponding one in TL. In our experiments reported below, we utilized four distinct Web service Japanese-to-English machine-translation engines[1], including Google Translate.

As explicitly represented in the formula, the score is computed by averaging over the text similarity scores ($TextSim(x, y)$), each resulting from a textual combination defined in R. That is, the pair (r_i, r_j), a member of R defines a textual combination considered by a calculation of text similarity. Here, r_i and r_j are chosen from the inventory $\{gloss, hyper, hypo\}$; while, $gloss(s)$ denotes the gloss text given in the lexical resource for the lexical concept s; and $hyper(s)/hypo(s)$ denotes the concatenation of the gloss texts given in the lexical resource for the hypernym/hyponym concepts of s. Note that an instance of R represents a particular strategy for extending gloss texts.

[1]All of the translation services used were provided by the Language Grid: `http://langrid.org/`. Detailed descriptions of the access interfaces are obtainable from `http://langrid.org/service_manager/language-services`.

3 Experiments

3.1 Test dataset

We utilized the same test dataset described by (Hayashi, 2012): comprising 196 query concepts and the relevant judgment annotations.

1. Query concepts: 196 concepts were randomly chosen from the EDR concepts that satisfy the following conditions: (1) A concept has to have one or more Japanese synonyms; (2) It has to have both Japanese and English glosses[2]; and (3) At least one of the senses of a Japanese synonym has to be considered substantially familiar to ordinary Japanese native speakers.

2. Candidate PWN synsets: By applying the baseline method in Hayashi (2012), at most 25 candidate PWN synsets were collected for each EDR query concept.

3. Relevance judgments: We established two relevance levels (Syn-level and Rel-level), as listed below. One of them was assigned to each candidate PWN synset by a single annotator. The annotator (not the authors of

[2]Although EDR is constructed bilingually, not all the concepts are necessarily specified in both Japanese and English.

this paper) was a native Japanese speaker with relatively high English literacy, and was asked to annotate only by referring to EDR synonyms/glosses given in Japanese and English and PWN synonyms/glosses given in English.

- Syn-level (almost synonymous): A PWN synset considered almost synonymous to the EDR query concept was assigned this label. This label was also assigned to POS variants of a thought-to-be synonymous PWN synset, because, unlike in PWN, a concept in EDR is not explicitly classified by part of speech. This led to cases in which one or more candidate synsets were assigned this label. Note also that, in some cases, none of the PWN synsets in the candidate set could be assigned this label because of a sense gap between EDR and PWN concepts and/or a lack of relevant synsets in the candidate set.

- Rel-level (related somehow): PWN synsets that may have some semantic association with an EDR query concept were assigned this label. The association could be hyponymy, meronymy, agentive argument, or even unrestricted associations. Quite frequently more than one synset in the candidate set were assigned this label.

3.2 Evaluation measures

Because of the following obvious correspondences, the experimental task itself was quite similar to that of information retrieval (IR): the whole PWN is a document set; a PWN synset can be a document; and a given EDR concept corresponds to a query reflecting a user's intent. We therefore adopted the following two IR-based measures to assess the performance of the proposed method.

- S@n (success rate at rank n): This measure determines the proportion of successful queries in the dataset. We judged a query to be successful if we could find at least one relevant candidate at the Rel- or Syn-level within the top n candidates. This measure is more relevant and important for assessing Syn-level relevance because at Syn-level only a few candidates should actually be relevant. In analyzing the experimental results, we set n to one.

- MAP (Mean Average Precision): This measure is a figure obtained by averaging over the average precision for queries in the query set. It is appreciated as being relatively reliable in assessing ranked retrieval results from an IR system (Manning et al., 2008). Our experiments introduced this measure primarily to evaluate Rel-level relevance; at Rel-level, we can find more than one relevant candidate in the candidate set, including the one(s) that are also intrinsically relevant at Syn-level.

3.3 Design of the experiments

We organized the experiments by considering the following parameters. More specifically, we conducted one experimental run using a combination of these conditions.

- Combination of extended gloss text: We considered four combinations (A through D), as shown in Table 1: Combination A only compares gloss texts without extensions; Combination B is the one presented in Banerjee and Pedersen (2003) as an illustrative example; Combination C integrates all of the ($9 = 3 \times 3$) combinations possible from the inventory of gloss extension patterns $\{gloss, hyper, hypo\}$[3]; and Combination D compares more texts, which are gathered casually from the conceptual neighbors of objective concepts. More specifically, $1\_hop(s)$ simply gathers gloss texts from the lexical concepts that can be reached in one hop beginning at s, considering more concepts linked by any relation type.

- Text similarity measure $TextSim(x, y)$: The texts to be compared (x and y) are represented as vectors, based on the bag of words notion. In an effort to obtain better text representation, we tested both the raw-frequency vector and the $tf * idf$-weighted vector, and applied the standard $cos\theta$ to measure the similarity between the vectors. Note also that we applied, what we deemed to be, standard linguistic preprocessing, including lemmatiza-

[3] EDR maintains, what is called, super/sub conceptual relationships rather than hyper/hypo; the former may include other hierarchical relationships, such as meronymy. By necessity, we collected super/sub EDR data in place of hyper/hypo in the experiments.

tion and stop-words deletion prior to the construction of the vectors.

- Blending ratio β: To assess the contribution of $pscore(s,t)$ and $gscore(s,t)$, we coarsely swept the blending ratio β in Formula (1) from 0.0 to 1.0.

	$gloss(t)$	$hype(t)$	$hypo(t)$	$1\_hop(t)$
$gloss(s)$	A, B, C	B, C	C	
$hype(s)$	B, C	B, C	C	
$hypo(s)$	C	C	B, C	
$1\_hop(s)$				D

Table 1: Relation pairs used in the experiments.

4 Results and discussion

4.1 Main results

4.1.1 S@1 measure

Table 2 summarizes the S@1 results for the textual combinations A through D, where the best scores at both Rel-level and Syn-level are listed. Displayed immediately below the scores are the type of vector (rV: raw frequency vector, wV: weighted vector) and the optimum blending ratio β.

Combination	Rel-level	Syn-level
A	**0.776** 0.0	**0.474** wV, 0.2
B	*ditto*	0.464 wV, 0.4
C	*ditto*	0.454 rV, 0.8
D	*ditto*	0.449 wV, 0.2

Table 2: The best S@1 results.

The results at Rel-level show that textual glosses, even extended ones, are ineffective for improving the S@1 measure, because the best score (0.776) is achieved when β is zero. However, the S@1 measures are more important in considering performance at Syn-level.

The results at Syn-level show that the best score (0.474) is achieved through combination A, which uses the weighted vectors with $\beta = 0.2$. This combination only employs non-extended glosses. The score declined when we applied more complicated textual combinations (B through D), but the differences are not statistically significant (p = 0.594, 0.504, 0.166, respectively; paired t-test). The results may indicate that the task of finding a synonymous conceptual mate in the first place can be affected by the noises inevitably introduced by more extended textual glosses.

4.1.2 MAP measure

Table 3 summarizes the best results obtained for the MAP measure. Unlike the results obtained using the S@1 measure, these results in general show that extended textual glosses effectively improve MAP performance.

Combination	Rel-level	Syn-level
A	0.671 rV, 0.2	0.531 rV, 0.06
B	0.682 wV, 0.2	0.525 wV, 0.4
C	**0.695** wV, 0.6	0.525 wV, 0.6
D	0.675 wV, 0.1	**0.534** wV, 0.1

Table 3: The best MAP results.

In the table, it can be seen that the best score at Rel-level (0.695) is achieved when combination C is adopted with $\beta = 0.6$, and the differences among combinations A, B, and D are statistically significant (p = 0.011, 0.035, and 0.027, respectively; paired t-test). This suggests that moderately extending gloss texts is beneficial because combination C considers more combinations of extended-gloss texts compared to combination B, but not as lavishly as combination D.

At Syn-level, on the other hand, combination D yields the best result, and combination A the second best. Although MAP performance is vastly more important at Rel-level, we still need to closely investigate the Syn-level results because they may not correlate with the results at Rel-level.

In summary, these results suggest that extended textual glosses effectively improve the MAP measure, in particular at Rel-level. However, there might also be a better solution than the current strategy: more controlled extension strategies that rather deal with the queries in a nonuniform manner by taking into account the characteristics of each individual query concept and its textual gloss can and should be developed.

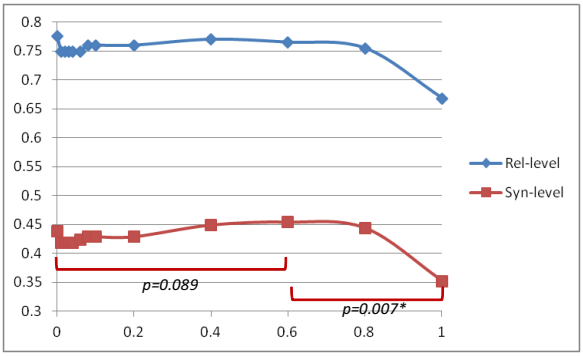


Figure 2: S@1 results with varying β.

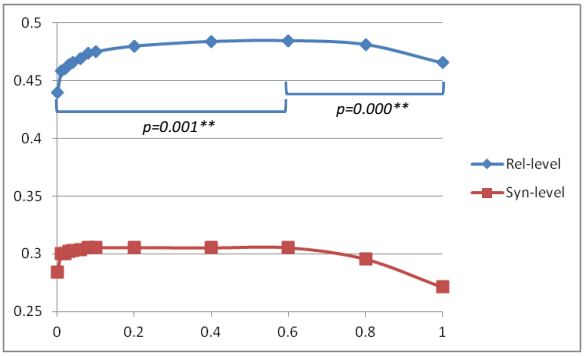


Figure 3: MAP results with varying β.

4.1.3 Blending ratio

The proposed method can be sensitive to the blending ratio β. Figure 2 displays S@1 performance and Figure 3 displays MAP performance; both sweep the value of β from 0.0 to 1.0. Note that these figures display the results only for the textual combination C with $tf * idf$-based weighted vectors.

S@1 results: Figure 2 shows that the scores at Rel-level and at Syn-level decrease sharply immediately after β departs 0.0; however, the scores are relatively stable in the range $\beta = 0.4 - 0.8$. At Rel-level, the score at $\beta = 0.4$ is almost the same as that at $\beta = 0.0$ ($0.775 \simeq 0.770$). At Syn-level, the maximum score (0.454) is obtained at $\beta = 0.6$. The scores at $\beta = 1.0$, which indicates that only gloss-based relatedness are employed, are never better than those of any other βs. Consequently, it can safely be said that adopting β somewhere between 0.4 and 0.5 is not a bad solution when we consider the simultaneous MAP performance discussed below. In the figure, it can also be seen that the difference at the Syn-level performance between $\beta = 0.0/0.6$ is not statistically significant (p = 0.089), while that at the Rel-level between $\beta = 0.6/1.0$ is statistically significant (p = 0.007).

MAP results: Figure 3 shows that scores at both Rel-level and Syn-level improve as β increases and are almost stable in the range $\beta = 0.2 - 0.6$. The best scores for both levels are achieved at $\beta = 0.6$: 0.695 at Rel-level and 0.525 at Syn-level, respectively. In the figure, it can also be seen that the differences in performance between $\beta = 0.0/0.6$ and $\beta = 0.6/1.0$ at the Rel-level are both statistically significant (p = 0.001 and 0.000, respectively). Notably, at Rel-level, the score at $\beta = 1.0$ exceeds that at $\beta = 0.0$ ($0.695 > 0.623$). This suggests that if we can adopt only one type of solution to achieve better MAP performance at Rel-level, it should be gloss-based relatedness, not synonym-based relatedness. In conclusion, we can safely say that the proposed method, balancing $pscore(s,t)$ and $gscore(s,t)$ by the blending ratio β is not only effective but also relatively robust, as the performances examined were not very sensitive to variations in β.

4.2 Impact of machine translation

To assess the impact of machine translation (MT), Table 4 compares the results achieved by the machine-translated textual glosses (MT_all column) with those yielded by the original English glosses given in the EDR (EDR column). Assuming the English glosses are reasonably adequate, they simulate a situation in which we can obtain "ideal" translations.

Measure: Level	EDR	MT_ all	MT_GT
S@1: Rel-level	**0.791**	0.776	0.739
S@1: Syn-level	**0.485**	0.474	0.449
MAP: Rel-level	0.690	**0.695**	0.670
MAP: Syn-level	**0.538**	0.534	0.510

Table 4: Impact of machine translation.

As can be imagined, the results yielded by the original glosses outperforms those of the MTed glosses in virtually all cases, but the differences are relatively minor and statistically insignificant (p = 0.257, 0.358, and 0.394, respectively; paired t-test). On the other hand, the MAP performance at Rel-level (which we consider the most important) achieved by MTed glosses is even better than that of the original EDR glosses, but again the

difference is statistically insignificant (p = 0.308). These results may suggest that the issue of vocabulary mismatch can be partially solved by incorporating as many translation texts from different translation engines as possible.

To examine this hypothesis, the MT_GT column of the table lists the results obtained when we used only Google Translate from among the four available translation engines. The figures in the related columns demonstrate that more translation engines can yield better results: the differences in the S@1 measure are not statistically significant (p = 0.089 for Rel-level and p = 0.138 for Syn-level), but the differences in the MAP measure clearly exhibited statistically significant differences (p = 0.000, 0.0426, respectively).

In summary, the use of MT to translate textual glosses does not appear to be as harmful as reported in previous ontology-matching literature Fu et al. (2009); McCrae et al. (2011) because the linguistic qualities of translation targets are different (a word/phrase label versus a short paragraph). Moreover, we have demonstrated that redundancy in the translated texts brought about by employing multiple translation engines definitely plays a role in improving performance, at least with an appropriate text-representation scheme and a proper text-similarity metric.

4.3 Extrapolated optimal performances

We believe that would be able to achieve better results if we could alter the blending ratio each time we processed *a* query concept, rather than the current uniform one. To partly examine this idea, we calculated the somewhat extrapolated results shown in Table 5, in which the best figures already shown in the previous tables are replicated in parentheses. The extrapolated figures were calculated a posteriori with the assumption that we could choose from {X:textual combination C with $\beta = 0.6$, Y:textual combination A with $\beta = 0.2$, Z:textual combination A with $\beta = 0.0$}, the best setting on a query-by-query basis.

Measure	Rel-level	Syn-level
S@1	0.847 (0.776)	0.541 (0.474)
MAP	0.741 (0.695)	0.582 (0.534)

Table 5: Extrapolated optimum performance figures.

As promisingly shown in the table, the extrapolated results assert that room for improvement exists. To further break down these results, we investigated the distribution of parameter settings adopted a posteriori: for 126 out of 196 queries, setting X was adopted, setting Y for 42 queries, and setting Z for 28. These figures again suggest that a query-dependent gloss expansion method would be promising, and worth being examined.

5 Related work and discussion

Although the present work demonstrates the potential usefulness of extended textual glosses (Banerjee and Pedersen, 2003), the range of contextual neighbors for the extension and the proper weighting of the textual similarities obtained from various combinations remain unidentified. The key to solving this issue is using a machine learning (ML) approach that can capture hidden properties in the lexical resources.

The use of machine translation and its impact in terms of performance have been explored by the research community looking at *ontology matching*. Their primary focus, however, has been on the translation of conceptual/ontological labels, as typically discussed in (Fu et al., 2009). More recently, McCrae et al. (2011) made a similar argument and presented a number of strategies that they assert could improve the performance of statistical machine translation of conceptual labels. In contrast to these research attempts, the present method applies machine translation to linguistically richer textual glosses. However, as the linguistic qualities of textual glosses may vary, at least from resource to resource, we may need to devise a method for properly assessing and adjusting to them.

From the perspective of multilingual terminology extraction and alignment, the presented work may not share research interests with terminologists, because our work targets dynamic pairwise alignment of existing lexical concepts. Nevertheless, we envision that the proposed method may be applicable to cross-lingual terminology alignment if domain-specific terms are extracted from a corpus with linguistic context, including their textual definitions. That is, we could robustly measure the cross-lingual similarity between linguistic contexts in different languages by applying multiple machine translation engines.

6 Concluding remarks

This paper demonstrated that extended textual glosses residing in lexical-semantic resources can be effectively employed even in the context of a cross-lingual task, provided multiple machine translation engines can be utilized and the resultant translation redundancy yields benefits. The applicability of the method could thereby increase, as the underlying resources (sense-tagged corpora for the synonym-based relatedness and translation engines for the gloss-based relatedness) would be more readily available and sophisticated.

However, improvements and extensions remain to be done in some areas of this work: the performance could be improved by considering the "sensitivity" (Kwong, 2012) of a lexical concept or a lexical concept type; and the multilinguality can be greatly enhanced by exploiting Web-based multilingual resources, as demonstrated by (Navigli and Ponzetto, 2010). Another direction that can be explored is the type discrimination of lexical-semantic correspondences. A promising solution would be an ML-based approach (e.g., (de Melo and Weikum, 2012)) that takes into account more features acquirable from the lexical resources at hand.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 258201170.

References

Banerjee, S. and T. Pedersen (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proc. of IJCAI 2003*, pp. 805–810.

Calzolari, N. (2008). Approaches towards a 'lexical web': the role of interoperability. In *Proc. of ICGL 2008*, pp. 34–42.

de Melo, G. and G. Weikum (2012). Constructing and utilizing wordnets using statistical methods. *Language Resources and Evaluation 46*(2), 287–311.

Fu, B., R. Brennnan, and D. O'Sullivan (2009). Cross-lingual ontology mapping - an investigation of the impact of machine translation. In *Proc. of ASWC 2009*.

Hayashi, Y. (2011). A representation framework for cross-lingual/interlingual lexical semantic correspondences. In *Proc. of IWCS 2011*, pp. 155–164.

Hayashi, Y. (2012). Computing cross-lingual synonym set similarity by using princeton annotated gloss corpus. In *Proc. of GWC 2012*, pp. 131–141.

Hayashi, Y. and T. Ishida (2006). A dictionary model for unifying machine readable dictionaries and computational concept lexicons. In *Proc. of LREC 2006*, pp. 1–6.

Kwong, O. Y. (2012). *New Perspectives on Computational and Cognitive Strategies for Word Sense Disambiguation*. Springer.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introdocution to Information Retrieval*. Cambiridge University Press.

McCrae, J., M. Espinoza, E. Montiel-Ponsoda, G. A. de Cea, and P. Cimiano (2011). Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In *Proc. of SSST-5*.

Miller, G. A. and C. Fellbaum (2007). Wordnet. then and now. *Language Resources and Evaluation 41*, 209–214.

Navigli, R. and S. P. Ponzetto (2010). Babelnet: Building a very large multilingual semantic network. In *Proc. of ACL 2010*, pp. 216–225.

Yokoi, T. (1995). The edr electronic dictionary. *Communications of the ACM 38*(11), 42–44.

Appendix-A. Information structure of the EDR Electronic Dictionary

The EDR electronic dictionary, unlike Princeton WordNet, is not a lexical database based on relational lexical semantics. Instead, it can be seen as a knowledge base that is enriched with linguistic information given in both Japanese and English. Another big difference lies in the consideration of parts of speech in the conceptual organization: as is well known, PWN maintains POS-dependent lexical semantic networks, while EDR bears only a POS-independent network.

Despite these differences, the information structure of EDR can be modeled in the same way as that of PWN (Hayashi and Ishida, 2006): the set of words associated with a common concept identifier in one or more of the sub-dictionaries can be modeled as a kind of synset.

Session : Term extraction

A Study of Association Measures and their Combination for Arabic MWT Extraction

Abdelkader El Mahdaouy
LIM, Univ. USMBA
Fès, Maroc
Univ. Grenoble Alpes
CNRS LIG/AMA
Grenoble, France
a.mahdaouy@hotmail.fr

Saïd EL Alaoui Ouatik
LIM, Univ. USMBA
Fès, Maroc
s_ouatik@yahoo.com

Eric Gaussier
Univ. Grenoble Alpes
CNRS LIG/AMA
Grenoble, France
eric.gaussier@imag.fr

Abstract

Automatic Multi-Word Term (MWT) extraction is a very important issue to many applications, such as information retrieval, question answering, and text categorization. Although many methods have been used for MWT extraction in English and other European languages, few studies have been applied to Arabic. In this paper, we propose a novel, hybrid method which combines linguistic and statistical approaches for Arabic Multi-Word Term extraction. The main contribution of our method is to consider contextual information and both termhood and unithood for association measures at the statistical filtering step. In addition, our technique takes into account the problem of MWT variation in the linguistic filtering step. The performance of the proposed statistical measure (NLC-value) is evaluated using an Arabic environment corpus by comparing it with some existing competitors. Experimental results show that our NLC-value measure outperforms the other ones in term of precision for both bi-grams and tri-grams.

1 Introduction

Automatic Multi-Word Term extraction is an important task in many Natural Language Processing (NLP) applications (Boulaknadel et al., 2008b; Wen et al., 2007). The aim of the MWT acquisition process is to extract specific domain terms from special language corpora (Korkontzelos et al., 2008). The extraction of MWTs is crucial for terminology acquisition, since they are less ambiguous and less polysemous than single word terms, and since their internal structure encodes useful semantic relations (Wen et al., 2008).

There are three main approaches to MWT extraction. The first one makes use of linguistic filters. The second one relies on statistical measures based on termhood and/or unithood. Termhood denotes *"the degree to which a linguistic unit is related to a specific domain concept"*, and unithood denotes *"the degree of strength or stability of syntagmatic combinations or collocations"* (Kageura et al., 1996). Lastly, the third approach is hybrid and combines the linguistic and the statistical approaches. Hybrid methods extract MWTs using linguistic filters and then rank the list of candidate MWTs according to statistical measures.

In this paper, we propose a novel, hybrid method for Arabic MWT extraction. Like other hybrid methods, it includes two main filters. In the first one, we use a part-of-speech (POS) tagger to extract candidate MWTs based on syntactic patterns. In the second one, we propose a novel statistical measure, the NLC-value, that unifies the contextual information and both termhood and unithood measures. We compare this measure to alternative ones in the task of MWT extraction : NTC-value (Vu et al., 2008), LLR+C-value (Al Khatib et al., 2010), C/NC-value and LLR.

The remainder of this paper is organized as follows. In the next section, Section 2, we present the related work. Section 3 describes the proposed method to extract MWTs. In Section 4, we present how MWT variation is handled in the proposed method. Section 5 describes the experimental validation and Section 6 concludes this work and presents some perspectives.

Abdelkader El Mahdaouy, Saïd El Alaoui Ouatik , Eric Gaussier

2 Related Work

Several studies have been conducted on MWT extraction for many languages. These studies have either used a linguistic approach, a statistical approach, or a combination of them (hybrid approach). Most recent MWT extraction methods rely on a hybrid approach to efficiently extract MWTs, due to its higher accuracy compared to the two other approaches (Tadic et al., 2003). The linguistic approach uses technical analysis on the current knowledge of the language and its structure. There are two subcategories : approaches based on morpho-syntactic patterns (Daille, 1994) and those based on MWT boundary detection (Bourigault, 1994).

The main purpose of applying statistical methods for MWT extraction is to rank candidate terms based on a particular measure that gives higher scores to ”good” candidate terms. Candidate terms above a particular threshold are selected for further processing. The reliance on frequency is based on the simple assumption that a frequent expression indicates an important representation of the domain in question. Therefore, frequent expressions are assumed to represent important concepts. Given a candidate multi-word term, frequency only counts how often the candidate occurs in the text, but doesn't give any information on the strength of the relationship between words composing the candidate multi-word term. Statistical approaches aim at extracting candidate terms from text corpora by means of association measures (Church et al., 1991) that concentrate on termhood and/or unithood to assign a score to candidate MWTs. These measures are based on frequency and co-occurrence information such as the T-score (Church et al., 1991), the loglikelihood ratio (LLR) (Dunning, 1994), the C/NC-Value (Frantzi et al., 1998), etc.

While linguistic approaches focus on syntactic structures, statistical methods focus on the recurrent characteristics of MWTs. Both have their advantages and limitations. As mentioned by Boulaknadel et al. (2008), statistical approaches *“are unable to deal with low-frequency MWTs“* while pure linguistic approaches are *“language dependent and not flexible enough to cope with complex structures of MWTs”* . Hybrid methods try to combine linguistic and statistical techniques to extract MWTs in order to avoid the weaknesses of the two approaches.

Boulaknadel et al. (2008) have relied on a hybrid method to extract Arabic MWTs. As a first step, candidate terms that fit syntactic patterns are extracted from the output of the part-of-speech (POS) tagging tool proposed by Diab (2004). In the second step, the list of candidate terms is ranked according to one of the following association measures : log-likelihood ratio (LLR), Mutual Information (MI), FLR, and T-score. These measures have been evaluated on an Arabic corpus and he results obtained show that LLR outperforms the other association measures.

Bounhass et al.(2009) have followed the same approach (using again Diab's (2004) POS tagger and LLR) while focusing on compound nouns and thus using a more restricted set of syntactic patterns. For the bigrams, the obtained results outperform those obtained by Boulaknadel et al. (2008).

A similar study has been conducted by Al Khatib et al. (2010), based on the POS tagger proposed by (Al-Taani et al. , 2009) and an association measure that combines both termhood and unithood through a combination of the C-value and the LLR. Experimental results show promising results for the combined measure.

Most hybrid methods presented previously have been evaluated on 100 (best) candidate MWTs and deal with bi-grams (i.e. candidate MWTs of length 2). Moreover, they rely on LRR or a combination of LRR and C-value (Al Khatib et al., 2010) and ignore contextual information in the ranking step. To overcome this limitation, we introduce a new association measure that integrates contextual information and both termhood and unithood. Our overall approach is also hybrid and relies on the same linguistic filters as the ones used in the previous studies, based on syntactic patterns applied on the output of the POS tagger developed by (Diab, 2009).

3 Proposed Method

Our method for extracting MWT candidates comprises two major steps : the linguistic and the statistical filters.

3.1 Linguistic Filter

The proposed linguistic filter extracts candidate MWTs based on two core components ; the POS tagger and the sequence identifier. In the literature, several methods for Arabic POS tagging systems have ben developed. We have used the one proposed by (Diab, 2009) as it performs at over 96% accuracy and allows a number of variable user settings. The underlying system uses Support Vector Machine (SVM). Figure (1) illustrates the global schema of our linguistic filter. As a first step, our

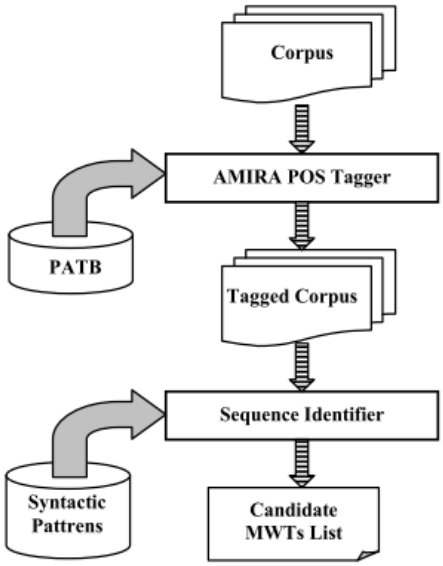


FIGURE 1: The global schema of the linguistic filter

method tags the corpus using the AMIRA toolkit (Diab POS Tagger) which is trained from the Penn Arabic TreeBank (PATB) to assign tags for each word in the corpus. Then, the sequence identifier tokenizes tagged files of the corpus and uses syntactic patterns in order to identify candidate terms that fit the rules of the grammar. We have extended the list of syntactic patterns used by Boulaknadel et al. (2008) as follows :

- $(Noun + (Noun|ADJ) + |(Noun|ADJ) + |(Noun|ADJ))$
- $Noun\ Prep\ Noun$

The second major step of the linguistic filter is handling the problem of MWTs variation to improve the effectiveness of extracted MWT candidates. Several categories of term variation are taken into account by this filter : graphical, inflectional, morpho-syntactic and syntactic variants, and are discussed in Section 4.

3.2 Statistical Filter

In this step, we apply a number of statistical measures to rank the list of candidate MWTs extracted by the linguistic filter. The main objective of our statistical filter is to consider both termhood and unithood measures.

3.2.1 The C-value

The C-value measures the termhood of a candidate string on the basis of several characteristics : number of occurrences, term nesting, and term length. It is defined as :

$$C\text{-Value}(a) = \begin{cases} \log_2(|a|) \cdot f(a) & \text{if } a \text{ is not nested,} \\ \log_2(|a|) \cdot (f(a) - g(a)) & \text{otherwise} \end{cases} \quad (1)$$

where $|a|$ denotes the length in words of candidate term a, $f(a)$ is the number of occurrences of a and :

$$g(a) = \frac{1}{|T_a|} \sum_{b \in T_a} f(b)$$

where $T(a)$ denotes the set of longer candidate terms into which a appears ($|T(a)|$ is the cardinality of this set).

As one can note, if the candidate term is not nested, its score is solely based on its number of occurrences and length. If it is nested, then its number of occurrences is corrected by the number of occurrences of the terms into which it appears.

3.2.2 The NC-value

The NC-value combines the contextual information of a term together with the C-Value. The contextual information is calculated based on the Nvalue which provides a measure of the terminological status of the context of a given candidate term. It is defined as :

$$N\text{value (a)} = \sum_{b \in C_a} f_a(b) \cdot \frac{|T(b)|}{n} \quad (2)$$

where C_a denotes the set of distinct context words of a, $f_a(b)$ corresponds to the number of times b occurs in the context of a and n is the total number of terms considered. This measure is then simply combined with the C-value to provide the overall NC-value measure :

$$NC\text{-value(a)} = 0.8 \cdot C\text{-value(a)} + 0.2 \cdot N\text{value(a)} \quad (3)$$

3.2.3 The NTC-value

The aim of the NTC-value (Vu et al., 2008) is to incorporate a unithood feature, through the T-score, to the C/NC-value to improve its performance. The T-score measures the adhesion or differences between two words in a corpus of N words as follows :

$$Ts(w_i, w_j) = \frac{p(w_i, w_j) - p(w_i) \cdot p(w_j)}{\sqrt{\frac{p(w_i, w_j)}{N}}} \quad (4)$$

where $p(w_i, w_j)$ corresponds to the probability of observing the bi-gram w_i, w_j in the corpus ; $p(w_i)$ is the probability of word w_i in the corpus and corresponds to the marginal probability of $p(w_i, w)$. The T-score is integrated in the C/NC measures through a re-weighting of the number of occurrences that privileges terms with a positive T-score :

$$F(a) = \begin{cases} f(a) & \text{if } \min(Ts(a)) \leq 0 \\ f(a) \ln(2 + \min(Ts(a)) & \text{otherwise} \end{cases} \quad (5)$$

where $\min(Ts(a))$ corresponds to the minimum T-score obtained from all the word pairs in a. Substituting $F(a)$ to $f(a)$ in Equation 1 yields the TC-value, which is then combined with the Nvalue as before, leading to the NTC-value :

$$NTC\text{-value(a)} = 0.8 \cdot TC\text{value(a)} + 0.2 \cdot N\text{value(a)} \quad (6)$$

The resulting metric (6) thus takes into account both contextual information and termhood and unithood measures.

3.2.4 The NLC-value

We follow here the same development as before but rely this time on the more accurate unithood feature LLR (Dunning, 1994), instead of the T-score, for the combination with the C/NC-value (Frantzi et al., 1998). LLR is a "goodness of fit" statistics that determines if the words in an observed n-gram come from a sample that is independently distributed (meaning they co-occur by chance) or not. The underlying measure is calculated for bi-grams by the following formula :

$$\begin{aligned} LLR(w_j, w_j) &= a \log(a) + b \log(b) + c \log(c) \\ &+ d \log(d) - (a + b) \log(a + b) \\ &- (a + c) \log(a + c) - (b + d) \log(b + d) \\ &- (c + d) \log(c + d) + N \log(N) \end{aligned}$$

with :
a : number of terms in which w_i and w_j co-occur ;
b : number of terms in which only w_i occurs ;
c : number of terms in which only w_j occurs ;
d : number of terms in which neither w_i nor w_j appear ;
N : total number of extracted terms.

For terms that consist of more than two terms, we calculate the LLR for each big-ram and then consider the minimum value obtained. The number of occurrences of a term is now re-weighted by this minimum value : $FL(a) = f(a) \cdot \ln(2 + \min(LLR(a)))$ which is used instead of $f(a)$ in the C-value, leading to the LC-value :

$$L\text{C-value(a)} = \begin{cases} \log_2(|a|) \cdot FL(a) & \text{if } a \text{ is not nested,} \\ \log_2(|a|) \cdot (FL(a) - GL(a)) & \text{else} \end{cases} \quad (7)$$

with $GL(a) = \frac{1}{|T_a|} \sum_{b \in T_a} FL(b)$

This measure is then combined with the Nvalue as before, leading to the NLC-value that integrates contextual information and both termhood and unithood :

$$NLC\text{-value(a)} = 0.8 \cdot LC\text{-value(a)} + 0.2 \cdot N\text{value(a)} \quad (8)$$

4 Term variation

As mentioned in the previous section, we have handled the problem of term variation at the linguistic step. Our method takes into account four types of variations : graphical variants, inflectional variants, morpho-syntactic variants and syntactic variants. Graphical variants concern orthographic errors occurred in writing a particular letters ("ï", "ي" and "ة") which are very common in Arabic. Furthermore, some letters go through a slight modification in writing, that doesn't necessarily change the meaning of the word. For example, the letter "ي" is replaced by another letter "ى" at the end of a MWT, as for "التلوث الكيميائي" which leads to "التلوث الكيميائى" meaning "chemical pollution". Inflectional variants are due to the use of different forms for the words constituting a MWT ; these different forms are related to gender and number of adjectives, as in "تلوث المحيط" (ocean pollution) and "تلوث المحيطات" (pollution of the oceans) and to the presence/absence

of a definite article, as in "تلوث مياه" (water pollution) and "تلوث المياه" (the water pollution). Morpho-syntactic variants affect the internal structure of term as the words it contains are related through derivational morphology. Two patterns control this type of variation in Arabic MWTs :

- $Noun1\,Noun2 \Leftrightarrow Noun1\,Adj$: "تلوث الهواء and "التلوث الهوائي" ("air pollution").
- $Noun1\,Adj \Leftrightarrow Noun1\,Prep\,Noun$: "برميل نفطي" and "برميل من النفط" ("barrel of oil").

We treat these three types of variations by using normalization method and the light stemming algorithm described in (Larkey et al., 2007) on each word of each MWT candidate.

Syntactic variants modify the internal structure of the MWT candidate by adding one or more words (as adjectives) but do not affect the grammatical categories of the content words of the original MWT candidate. Such variants can be identified, for a given MWT candidate, by searching for all the stemmed MWT candidates that contain it. All the elements that constitute an addition to the original MWT candidate are then considered as context terms.

5 Experiments and Results

5.1 The Corpus

Since there is no standard domain-specific Arabic corpus, we have built, in order to evaluate our approach, a new corpus specialized on the environmental domain with similar properties as the ones described (Boulaknadel et al. , 2008; Bounhas et al., 2009; Al Khatib et al., 2010).

The corpus built contain 1666 files comprising 53569 different tokens (without stop words) extracted from the Web site "Al-Khat Alakhdar" [1]. It covers various environmental topics such as pollution, noise effects, water purification, soil degradation, forest preservation, climate change and natural disasters.

5.2 Evaluation and Results

The evaluation of automatic MWTs extraction is a complex process and is usually performed by comparing each MWT candidate extracted to a domain-specific reference list. When there is no reference list available in the language retained, one can first translate the MWT candidates (using a machine translation system or a bilingual dictionary) and use a reference list available in another language. For the evaluation purpose, we have constituted automatically a reference list of all Arabic MWTs available in the latest version of AGROVOC [2] thesaurus and then use the stemming algorithm to remove prefixes and suffixes for each MWT in the reference list and the extracted MWT list. The next step consists of using an algorithm that considers a MWT candidate as correct if it is included in this list, noting that the MWT candidate and the term in the reference list should have the same number of stemmed words. Otherwise, we translate it and consider it as relevant whether its translation is contained in the European terminological database IATE [3]. Finally, the precision is calculated using the number of attested MWTs and the number of considred terms.

We computed the association scores (LLR, C-value, NC-value, NTC-value, LLR+C-value, NLC-value) for the MWT candidates and retain from each produced ranking for each statistical measure the k-best candidates, with k ranging from $100 - 300$ at intervals of 100. The experimental results illustrated in table 1 show that our method (NLC-value) outperforms the previous methods in term of the quality of the extracted MWTs.

	Top MWT considred		
Stat. measures	**100**	**200**	**300**
LLR	75,0%	70,5%	64,3%
C-value	71,0%	69,0%	67,3%
NC-value	74,0%	70,0%	68,3%
NTC-value	80,0%	71,5%	69,7%
LLR+C-value	73,0%	72,0%	68,3%
NLC-Value	82,0%	75,5%	73,0%

TABLE 1: Results obtained for different statistical measures

Furthermore, the combination of the context information and the C-value improves the performance of the process of MWT extraction because the NC-value outperforms the C-value for each

1. http ://www.greenline.com.kw

2. www.fao.org/agrovoc/

3. http ://iate.europa.eu/iatediff

considered MWT list. The unithood feature LLR outperforms the C/NC-value as expected from previous studies. Figure 2 illustrates the precision obtained for the C/NC-value and the LLR.

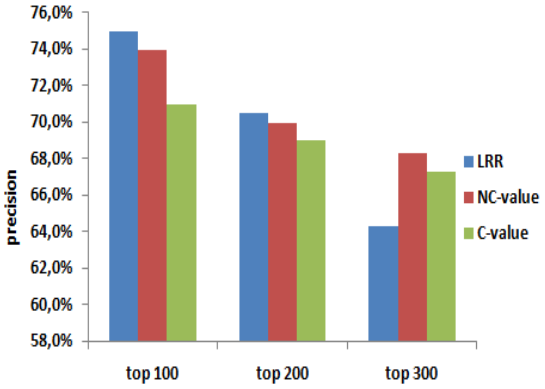


FIGURE 2: Precision obtained for the LLR and the C/NC-value

The integration of contextual information and the T-score unithood measure to the C-value improves the performance of MWT acquisition, since the NTC-value has better precision than the C/NC-value, as illustrated in Figure 3.

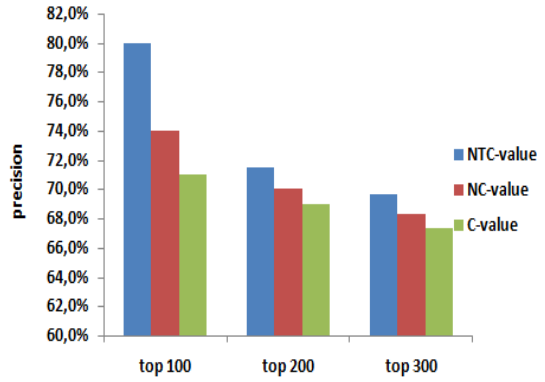


FIGURE 3: Precision obtained for the C/NC-value and the NTC-value

Lastly, the combination of termhood and unithood measures (NTC-value, $LLR + C$-value, NLC-value) is essential for MWT extraction, since all the measures based on this combination perform better than measures using only termhood or unithood (C-value, NC-value, LLR). We note that the statistical measure we have propose, NLC-value, outperforms all other measures. This measure is based on the accurate unithood feature LLR, combined with the NC-value. The NLC-value method takes advantages from previous works proposed in (Vu et al., 2008) and (Al Khatib et al., 2010) taken into account contextual information and both termhood and unithood association measures. Figure 4 presents a comparaison of the precision for different statistical measures that combine termhood and unithood.

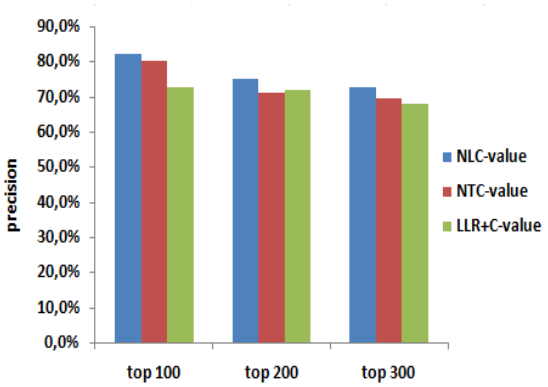


FIGURE 4: Precision obtained for different statistical measures that combine termhood and unithood

The number of different terms evaluated are 1095 amongst other 1800 terms, moreover the statical measures share 141 terms. The tables 2 and 3 represent the number of terms found in agrovoc and IATE respectively.

	Top MWT considred		
Stat. measures	**100**	**200**	**300**
LLR	35	60	80
C-value	27	59	82
NC-value	32	62	82
NTC-value	35	60	83
LLR+C-value	34	60	84
NLC-Value	41	65	86

TABLE 2: the number of terms found in agrovoc foreach measure

6 Conclusion

In this work, we have presented a hybrid method for Arabic MWT acquisition ; this method takes advantage of existing linguistic and statistical approaches. As a first step, we apply linguistic filters to extract MWT candidates based on syntactic patterns using a sequence identifier component. Then, MWT variants are identified through a morphological analysis of he extracted MWTs based on light stemming. In the statistical step, we

	Top MWT considred		
Stat. measures	**100**	**200**	**300**
LLR	40	81	113
C-value	44	79	120
NC-value	42	78	123
NTC-value	45	83	126
LLR+C-value	39	84	121
NLC-Value	41	86	133

TABLE 3: the number of terms found in IATE foreach measure

have proposed a novel statistical measure, NLC-value, that consists of ranking MWT candidates by considering contextual information and both termhood and unithood statistical measures.

Experiments are performed for bi-grams and tri-grams on an environment Arabic corpus. The experimental results show that our method outperforms the previous ones in term of quality of the extracted MWTs. In conclusion, the combination of the best association measures that integrate contextual information and both termhood and unithood statistical measures improves the performance of the MWT acquisition process.

In a near future, we plan on using the extracted MWTs in an information retrieval system as complex terms often constitute a better representation of the content of a document than single word terms.

References

Al Khatib K, and Badarneh A. 2010. *Automatic extraction of arabic multi-word terms*. In Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 411-418.

Al-Taani A, and Abu-Al-Rub S. 2009. *A rule-based approach for tagging non-vocalized Arabic words*, volume 6. The International Arab Journal of Information Technology, p 320.

Boulaknadel S, Daille B, and Aboutajdine D. 2008 a. *Multi-word term indexing for Arabic document retrieval*. In Proceedings of the The IEEE symposium on Computers and Communications, pp. 869-873.

Boulaknadel S, Daille B, and Aboutajdine D. 2008 b. *A multi-word term extraction program for Arabic language*. the 6th international Conference on Language Resources and Evaluation LREC.

Bounhas I, and Slimani Y. 2009. *A hybrid approach for Arabic multi-word term extinction*. International Conference on Language Processing and Knowledge Engineering, pp. 1-8.

Bourigault D. 1994. *LEXTER, un logiciel d'EXtraction de TERminologie, Application à l'acquisition des connaissances à partir de textes*, phd thesis. Ecole des Hautes études en Sciences Sociales, Paris.

Church K, Gale W, Hanks P, and Hindle D. 1991. *Using statistics in lexical analysis*. In Lexical Acquisition, Exploiting On-Line Resources to Build a Lexicon.

Daille B. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, Phd thesis. University of Paris 7.

Diab M Hacioglu K, and Jurafski D. 2004. *Automatic tagging of Arabic text : From raw text to base phrase chunks*. In Proceedings of North American Association for Computational Linguistics NAACL, pp. 149-152.

Diab M. 2009. *Second Generation Tools (AMIRA 2.0) : Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking*. In Proceedings of International Conference on Arabic Language Resources and Tools.

Dunning T. 1994. *Accurate Methods for the Statistics of Surprise and Coincidence*, volume 19. Computational Linguistics, pp. 61-74.

Frantzi K. T, Ananiadou S, and Tsujii T. 1998. *The C-Value/NC-Value Method of Automatic Recognition for Multi-word terms*. Journal on Research and Advanced Technology for Digital Libraries, pp. 115-130.

Kageura K, and Umino B. 1996. *Methods of Automatic Term Recognition A Review*, volume 3. Terminology.

Korkontzelos I, Ioannis P. Klapaftis, and Manandhar S. 2008. *Reviewing and Evaluating Automatic Term Recognition Techniques*. In Procedings of the 6th international Conference on Advances in Natural Language Processing, pp. 248-259.

Larkey S. Leah , Ballesteros L, and Connell E. Margaret. 2007. *Light Stemming for Arabic Information Retrieval*, volume 38. Text, Speech and Language Technology, pp. 221-243.

Tadic M, and Sojat K. 2008. *Finding multiword term candidates in Croatian*. In the Proceedings of IESL2003 Workshop, pp. 102-107.

Vu T, Aw A. Ti, and Zhang M. 2008. *Term Extraction Through Unithood And Termhood Unification*. In Procedings of IJCNLP.

Wen Z, Yoshida T, and Xijin T. 2007. *Text classification using multi-word features*. In Proceedings of the The IEEE symposium on Computers and Communications, pp. 3519-3524.

Wen Z, Yoshida T, and Xijin T. 2008. *A Study on Multi-word Extraction form Chinese Documents*. Advanced Web and Network Technologies, pp. 42-53.

Lessons from students: A pilot project to discover guidelines for creating a student-friendly, relation-rich term bank

Elizabeth Marshman
School of Translation and
Interpretation, University of Ottawa
70 Laurier Ave. E. (401)
Ottawa, Ontario, Canada K1N 6N5
Elizabeth.Marshman@uOttawa.ca

Abstract

Since the 1990s, there has been growing interest in two key types of terminological information: terminological relations (including generic-specific and part-whole, as well as various non-hierarchical relations), and terminological contexts. These come together in knowledge-rich contexts (KRCs), which both illustrate terms' behaviour in texts and reveal important connections between terms and between concepts. Such information has been integrated into prototype resources for translators, technical writers, subject-field specialists and students. As more resources integrating this information are developed, we must evaluate how to present it effectively for key user groups. In this paper, we will report on a small pilot project carried out with translation students translating between English and French. The students translated excerpts of popularized texts on breast cancer, using the CREATerminal (a terminology resource model that includes English and French KRCs describing four terminological relations), and compared the information this resource provided with that on term records in TERMIUM® Plus and the *Grand dictionnaire terminologique* (GDT). We report students' evaluations of the three resources and attempt to derive some guidelines for developing student-friendly, relation-enriched terminology resources.

1 Introduction

Although perspectives and terminology used may differ, the importance of terminological relations in terminology research and management is appreciated by many scholars. Attention has focused at various points on the classification and description of relations that are relevant for terminology work, on methods for extracting these from texts, and on the relevance of and approaches to integrating this information into terminology resources. Among the first proposals for relation-enriched terminology resources was Meyer et al.'s (1992) *terminological knowledge base* (TKB), a terminology resource that describes not only a range of concepts but also a variety of relationships that hold between them.

These relations can be identified manually or even (semi-)automatically from texts (cf. L'Homme and Marshman 2006) in the form of *knowledge-rich contexts* (KRCs) (Meyer 2001). These excerpts of texts often contain *knowledge patterns*—i.e., combinations of terms or other linguistic units that express concepts, linked by lexical markers of the relations between them—and can both provide information to assist in understanding the terms and concepts and illustrate the linguistic items in use. As excerpts of "authentic" texts, KRCs can also illustrate variation in concepts' expression and the lexical markers used in various communicative situations (e.g. Condamines 2002, 2008; Marshman and L'Homme 2008; Marshman et al. 2009).

In Meyer et al.'s footsteps followed researchers who have investigated various strategies for developing and populating TKBs (e.g. Condamines and Rebeyrolle 2000, 2001; Faber et al. 2011; Faber and San Martín 2011; León et al. 2011, 2013) in a selection of domains. Some projects have addressed the use of terminological relationships in the form of ontologies (e.g. Cabré et al. 2004; Gillam et al. 2005; Maroto and Alcina 2009), as part of an increasing movement towards the integration of terminology and ontology (e.g. Temmerman and Kerremans 2003; cf. also Roche et al. 2011). Still others (e.g. L'Homme 2013, 2013a) have described lexical relationships between terms. Most of these resources have been in electronic form, although some specialized print dictionaries (e.g. Dancette and Réthoré 2000) have included such information.

While relations are being increasingly prioritized in resources, there is still no standard model for relation choice and representation. This may be true in part because the wide variety of users of terminology resources and purposes for their use (e.g. Sager 1990) entails diverse needs in this area. Faber and San Martín (2011: 48) express the need for "customized" design of terminology resources:

[I]n order for any knowledge resource to aspire to psychological and explanatory adequacy, its underlying conceptualization and design must be in consonance with the needs and expectations of a specific user group, whose main objective is generally to acquire knowledge about the specialized area.

In some contexts, even this highly relevant observation can be questioned: translators (who are seen as the primary users of terminology databases in contexts such as Canada's) may not be as interested in domain knowledge per se as in terms, equivalents, synonyms and their use (including the contexts in which they occur). These and similar observations have led some to conclude that conventional terminology resources such as the large term banks, including TERMIUM® Plus[1] and the *Grand dictionnaire terminologique* (GDT)[2], are not adequate for translators' needs. The same can be said for other resources: ontology-based resources may also not be easy to understand and use for non-subject-field specialists such as translators and terminologists (Cabré et al. 2004: 87).

It is thus important to examine and discuss some of the resources that are available to specific groups and how (and how well) they meet the needs of these groups. In this paper, we will focus on the needs of trainee translators: individuals who are likely in need of both subject-field and linguistic knowledge to carry out a translation task, but may attribute different levels of importance to each kind of knowledge, and may evaluate the resources that supply this knowledge differently from other groups and from one another. We will gather information about trainee translators' reactions to resources from a questionnaire completed by users of three terminology resources, and try to extrapolate some guidelines for the creation of effective resources based on this feedback.

We will begin with a brief overview of some of the currently available terminology resources (section 2). We will outline the methodology used to gather information for this pilot study (section 3), and then will present and discuss some findings (section 4), before wrapping up with some brief remarks, suggested guidelines derived from the observations, and ideas for future work (section 5).

2 Approaches in terminology resources

In this section, we will provide a brief overview of the conventional term banks used in the project (2.1), as well as a few examples of relation-enriched resources and how they have complemented this basic model with terminological relations (2.2), and then describe the CREATerminal prototype used in this study (2.3).

2.1 Conventional term banks

The largest and most widely used term banks today are mainly constructed on traditional models such as those described by Pavel and Nolet (2001) and Dubuc (2002), and provide a range of information to translators, students, writers and other users.

The Government of Canada's TERMIUM® Plus term bank (Government of Canada 2013; see also Pavel and Nolet 2001) has very broad coverage, including over four million terms (most in English and French, but with a growing component of Spanish and Portuguese) from a wide range of domains. In addition to administrative information including dates of modification and record authors, its term records contain largely standard term record fields of domain and sub-domain, terms, equivalents, sources, part-of-speech labels, usage labels, definitions, contexts, observations and in some cases phraseologisms (although not all of these fields may appear on each record).

After a significant "facelift" in the last two years, the GDT now presents terms (mainly French and English, with a small complement of other languages) from a wide range of domains in a term record format that calls particular attention to French terms and to the associated usage information (particularly appropriateness for use in Quebec). In addition to (mostly French) definitions, some records include illustrations and notes to clarify meaning (including distinctions between related terms and concepts) and usage, as well as administrative fields.

Coverage of terminological relations in these resources is uneven, with any such information

[1] http://www.termiumplus.com
[2] http://www.granddictionnaire.com

generally found in definitions, contexts or observations/notes.

2.2 Enriching terminology resources with relations

In filling the gaps in this traditional term record model and developing the idea of TKBs or ontologies, a number of projects have addressed the needs of users for additional relation information. Meyer et al.'s (1992) COGNITERM project was followed by other projects including *GenomaKB*[3] (Cabré et al. 2004; Feliu et al. 2004) that integrated corpora and bibliographical information with a terminological database and an ontology to provide an integrated, multilingual resource that would meet the needs of non-subject-field specialists in the field of the genome. This type of integration reflects some of the observations of Bowker (2011), which highlighted the usefulness of access to corpus data for translators researching terms.

Similar attention has been paid to the importance of context of use and its potential for disambiguation in the description of terms and terminological relationships in the *EcoLexicon*[4] project (Faber et al. 2011; León et al. 2011, 2013). This multilingual resource in the field of environmental science provides access not only to definitions of concepts, but also to visual information in the form of both illustrations and dynamic relation maps that illustrate connections between terms and other elements (including generic-specific, part-whole and various non-hierarchical relations) based on an approach inspired by Fillmore's Frame Semantics (Faber et al. 2011; Faber and San Martín 2011). The dynamic visualization options allow the user to view a wide range of connections and to navigate by following links between concepts, in order to better understand their complex interconnections.

Another set of resources, including the *DiCoInfo*[5] and the *DiCoEnviro*[6], has been created by a team headed by Marie-Claude L'Homme at the Université de Montréal's Observatoire de linguistique Sens-Texte. Developed based on corpus data from the perspective of lexico-semantic terminology, and calling upon principles of Explanatory and Combinatorial Lexicography (Melčuk et al. 1995; L'Homme 2012) and later on Frame Semantics, these resources provide extensive descriptions of links between terms (including nouns, verbs, adjectives and phrases) in the fields of computing and the Internet and of the environment, respectively. In addition to part-of-speech labels, equivalents, contexts and definitions, terms are accompanied by an analysis of their actantial structures and typical actants, as well as a list of terminological relationships that may include synonyms, antonyms, hyponyms, hypernyms, meronyms, and holonyms, as well as a number of "custom" relations observed in the corpora. A visual interface, the *DiCoInfo visuel* (Robichaud 2012) allows users to view connections between the terms described in the *DiCoInfo*.

This small sample of resources reflects the potential for explicitly describing a wide variety of relationships relevant in terminology, as well as a range of options for making this information easily accessible to users, including increased access to a variety of contexts and options for various approaches to navigation within the resource, including a visual interface.

2.3 The CREATerminal prototype

Another in the list of relation-rich resources, but far less developed than those described above, is the CREATerminal prototype. In development since 2007, it aims to provide a useful resource for translators, built based on the content of popularized, bilingual (English-French) documents in the field of breast cancer (e.g. Marshman and Van Bolderen 2009; Marshman, Gariépy and Harms 2012). The information contained in the CREATerminal prototype was extracted from bilingual Canadian web sites (e.g. Health Canada, the Canadian Cancer Society, and the Canadian Breast Cancer Foundation).

The CREATerminal is a Microsoft Access database with three main tables: one has an entry for each of the approximately 85 concepts covered in the resource, and links the terms identified for each concept with their equivalents in the other language; one includes approximately 250 bilingual contexts showing the terms and their equivalents in use, and the third presents a total of approximately 800 bilingual KRCs that illustrate terminological relations (generic-specific, part-whole, cause-effect and entity-function) that involve the concepts and include lexical relation markers. These KRCs are anno-

[3] http://brangaene.upf.es:8080/genoma/index.jsp
[4] http://ecolexicon.ugr.es/en/index.htm
[5] http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi
[6] http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search.cgi

tated to identify the relationship present, the relation marker, the related items, and their sources.

Users can browse term records from the term record form—which shows terms and equivalents, and offers buttons to display examples and KRCs illustrating different relations—or view complete lists of KRCs for each relation type or lexical relation markers for the relations.

The database can also be searched using generic queries that allow the user to search for specific character strings in term records, examples and KRCs.

3 Methodology

This pilot project focuses on the comparison of the CREATerminal, TERMIUM® Plus and the GDT by a sample of students in translation programs (B.A. and graduate programs) at the University of Ottawa. These students were predominantly Anglophone and registered in courses that included a component of terminology and/or terminography. The students were first introduced to the concept and relevance of relations in the field of terminology in their courses and with an introductory in-class exercise, and to the CREATerminal model and how to consult and search it. (All had previously used both TERMIUM® Plus and the GDT in their coursework and were assumed to be comfortable with their use.) They were then asked to carry out a translation task and invited to complete an optional, anonymous online questionnaire summarizing their experiences after class.

The task involved translating a selection of short (1-3 sentence) excerpts of popularized texts on breast cancer. A mix of English to French and French to English translation was offered, and students were asked to try both (so that they would be translating both into and out of their L2). Students were asked to pay particular attention to highlighted terms in the excerpts and to look them up in the three terminological resources. All concepts corresponding to the highlighted terms were described in at least two of the terminology resources used in the comparison, although occasionally term forms or terms themselves varied.

Students were asked to translate as many excerpts as possible in a thirty-minute period. They then were invited to complete the questionnaire, delivered via the Survey Monkey interface. The first section of the questionnaire gathered general information on the respondents' perceptions of several subjects: the resources' usefulness for understanding concepts in the excerpts and for writing about them; what the respondents found most and least useful about each resource; and which resources they would use again for a similar task. The second section (on a new page) addressed terminological relations specifically, and asked about students' perceptions of how well terminological relations were described in each resource, as well as how useful the information about terminological relations in general was for understanding concepts and for writing about them. Respondents were also asked to evaluate the usefulness of individual record fields containing this relation-related information. Finally, the third section asked students to identify which fields they would consider useful in their own translation-oriented term records (i.e. whether they currently included them, planned to include them, would consider including them, or did not and would not include them).

Multiple-choice questions were scored on a rating scale from 1 to 4, with 1 representing a negative evaluation (e.g. "not at all useful" for questions about usefulness, and "do not and will not include" for questions about term record fields) and 4 representing a positive evaluation (e.g. "very useful", "currently include"). Average scores were computed automatically by Survey Monkey based on these scales.

Where applicable, a "don't know" or "did not consult/use" option was provided. Participants were also offered the option to list and evaluate additional resources they consulted.

In total, 24 respondents consented to participate in the survey. A very high dropout rate of almost 50% after the first question suggests that many may have first accessed the questionnaire to familiarize themselves with its contents (as the main questions could only be accessed after consenting to participate), and either returned later to complete it or were dissuaded by the nature or length of the questionnaire. Of the 13 respondents who continued to the second question, 7 continued to the final question.

3.1 Some limitations of the methodology

An important limitation of this study is the small sample size and the high dropout rate. Important ethical considerations involved in the collection of data from students required great care to avoid coercion and ensure anonymity, which unfortunately limited opportunities to encourage partici-

pation and follow up with potential respondents (in addition to imposing significant restrictions on the general methodology). Moreover, the nature of the sample itself should be taken into account, as it consists of students from a single academic setting, and those most likely to participate were doubtless those who had a particular interest in terminology in general and terminological relations in particular.

The range of term records consulted was also necessarily restricted by time limitations and coverage limitations for the three banks, and the approach used to introduce variety by giving a choice of excerpts to translate (coupled with the survey-based methodology) also made it impossible to verify exactly which term records in each resource were consulted by each individual.

The limitations inherent in the use of a purely survey-based methodology for data collection are also significant in themselves. We accessed only respondents' perceptions of their experience, and thus were not able to objectively measure aspects of this experience, or to provide a fine-grained portrait of how the various resources were actually used.

It is therefore essential that these data be taken as purely indicative clues to help in identifying key concerns in creating student-friendly, relation-rich terminology resources (and certainly not as evaluations of the quality of any specific resource). Given the limitations of the sample, no statistical evaluation of the data will be carried out beyond the comparison of average scores from multiple-choice questions and percentages of respondents within the group.

4 Findings and discussion

In the first section of the questionnaire, respondents were asked to evaluate and compare the usefulness of the three resources for two main tasks: understanding concepts (i.e. decoding the source text) and writing about concepts (i.e. encoding the target text).

In the average overall evaluation of the usefulness of the three resources, the 13 respondents found all of the resources to be between "fairly useful" and "very useful" for understanding concepts: TERMIUM® Plus had the highest average score of 3.46 out of 4, followed by the GDT at 3.17 and finally the CREATerminal at 3.00. For writing about concepts, the scores showed a wider rage and fell just slightly below "fairly useful" into the range of "somewhat useful". In contrast to the previous ranking, the CREATerminal scored highest, with an average score of 3.36, followed by TERMIUM® Plus with an average score of 2.92 and finally the GDT at 2.73.

The very different ranking of the resources for the two purposes most likely reflects the strengths of different types of data. Among the chief complaints were some gaps in information (e.g. of definitions, contexts and cooccurrents in TERMIUM® Plus and the GDT), and problems with searching and display in all three resources (e.g. having to scroll down or through various records to find the relevant one in TERMIUM® Plus and the GDT, or having to work with one query at a time and to close tabs between searches in the CREATerminal).

On the positive side, and unsurprisingly, in each resource the coverage and variety of equivalents included were valued. Among the strengths of TERMIUM® Plus, respondents cited broad coverage of terms and concepts and inclusion of bilingual information—both likely to assist with understanding—as well as ease of use and precise searching. The GDT's strengths, as identified by the students, included the notes provided about usage, origin, etc. These might fulfill a decoding or an encoding function. Finally, the numerous, bilingual KRCs in the CREATerminal seemed most helpful for writing about concepts.

We can thus observe that students value both the defining and the illustrating functions of terminology resources. This may represent an exception to the general observation that translators tend to be most concerned with equivalents and less with definitions, perhaps because these are students working in a largely unfamiliar field—or because they were asked specifically about the understanding of concepts.

On a related point, in the third section of the questionnaire, 4 of the 7 respondents reported currently storing definitions on their term records, and 2 of the others reported planning to include them (an overall score of 3.43). In contrast, none of the students reported currently storing relation-related fields, although between 3 and 5 of the respondents (depending on the field) indicated that they would consider including them. The students seemed more likely to consider including conventional term record fields (ranging from a score of 2.5 for phraseologisms to 3.29 for contexts and 3.43 for definitions) than relation-related fields (ranging from 1.5 for

sources of terminological relations to 2.33 for a context illustrating the relationships).

In the second section of the questionnaire, when asked about the usefulness of the different types of relations described in the CREATerminal, the 9 respondents indicated that they were useful to varying degrees, with the highest average score (3.5 out of 4) for the generic-specific relation, followed by part-whole (3.2), entity-function (3.0) and finally cause-effect (2.8). When asked about specific elements of the annotated KRCs that were helpful for understanding the concepts (excluding the terms themselves), the highest-ranked fields were the example source (with an average of 3.2) and the French example (3.17). The other fields, except for the French relation marker (2.67), scored 3.0, indicating that these elements were considered fairly useful. For writing about concepts, the English context explaining the relation was on average ranked most useful (3.75), followed by the English lexical relation marker (3.5), the French context (3.2) and the English related term/item (3.0). All other fields scored below 3.0. The average score from 10 responses to the final question from the section indicated that the CREATerminal provided the most useful information about terminological relations (with a score of 3.56 out of 4), followed by the GDT at 2.88 and finally TERMIUM® Plus at 2.56.

This provides an interesting contrast to the observations above, in that information about terminological relations appears to be useful, but not very likely to be stored by students in their own records (perhaps because of the complexity and labour-intensiveness of the task) and also unlikely to be thoroughly covered in conventional terminology resources. We thus see the need for "third-party" terminology resources that do integrate this information to fill the gap for trainee translators (and those with similar needs).

This need is reflected somewhat in the reactions of users when asked which resources they would use for a similar task again. Of the 12 respondents to this question, 83% indicated that they would use TERMIUM® Plus, 67% would use the CREATerminal, and 50% would use the GDT. (It should nevertheless be noted that the respondents were mostly Anglophone and that the data suggest that they were paying particular attention to information for encoding in English, which is not the primary purpose of the GDT.)

5 Conclusion

This study has elicited some encouraging reactions from students, indicating that relation-enriched resources can meet some perceived needs in carrying out a translation task. From the literature and findings described above, we can observe a high priority accorded to equivalents (which is not surprising) and to the understanding of concepts (e.g. via definitions and KRCs). This may well reflect the nature of the students' experience, in their need to interpret concepts that are almost inevitably unfamiliar (and challenging given the fact that the translations were of excerpts and not whole texts, which would provide more information to help with interpretation). There is also a positive evaluation of the usefulness of relation-related information—particularly for writing—as evidenced by the evaluation of the CREATerminal resource and the willingness to use it again.

Although this data is admittedly very limited, we can derive some preliminary, suggested guidelines for the creation of a relation-rich terminology resource that would meet the expectations of the students:

Guideline 1: Maximize user-friendliness. Students' reactions suggest that for this user group, the user-friendliness of resources is fundamental. Regardless of resources' content (and coverage was highly valued by the respondents), it seems that easy access to this information may be equally important. The inclusion of visual interfaces such as in the *EcoLexicon* and the *DiCoInfo visuel*—particularly if these are smoothly integrated into an interface that also allows for easy consultation of textual material—are promising avenues for future development.

Guideline 2: Integrate numerous KRCs. The students did find the relation information they consulted helpful, and seemed to be particularly drawn to it in the form of KRCs. This may be due to a focus on information that can be useful for writing about concepts as well as understanding them. In any case, to satisfy the needs of this user group, it seems beneficial to include as wide a range of KRCs as possible (or practical) to take advantage of the dual function of these items (while nevertheless maintaining efficient integration and organization of the material to ensure easy navigation). Despite the time investment, advantages to including selected KRCs in a resource rather than offering (only) direct access to corpus data may include both speed and ease of

access to information—particularly for users who are new to the subject field in question and may need assistance for the first stages of research—as well as the ability to exploit the data they contain, e.g. for visual representation of relevant relationships.

Guideline 3: Include parallel, bilingual information where available. A bilingual format is common to TERMIUM (which often includes definitions and contexts in both languages, usually from comparable resources) and the CREATerminal (which includes parallel bilingual contexts). As noted by Bowker (2011: 221), in spite of traditional terminology guidelines, translators increasingly tend to value translated sources (parallel corpora, translation memories) and the rapidity and ease of use these information sources offer. Although the benefits of comparable corpora in terminology work are well established, it seems that the inclusion of complementary translated information can be an asset in the eyes of the trainee translators.

Future work will allow us to investigate the use of these resources in more detail and to better understand to what extent these preliminary guidelines are relevant, and why. It will be essential to gather more data from a wider variety of users in order to identify more generalizable trends in requirements and preferences. A more in-depth study of the use of the resources by participants (e.g. using screen recording and interviews, or—resources permitting—using eye-tracking and keystroke logging tools to monitor users' activity) could allow us to obtain a more accurate and detailed picture of how students use such resources and their contents, and what factors they take into account in evaluations.

By gaining a better understanding of the design, use and usefulness of student-friendly, relation-rich resources, we will be better able not only to produce richer and more useful tools but also to better train students to use and even create them in the workplace.

Acknowledgements

The author wishes to thank the organizations who granted permission to use their texts in the CREATerminal project, as well as the numerous research assistants at uOttawa who have contributed to the project. Thanks are also extended to the Social Science and Humanities Research Council of Canada and the University of Ottawa and uOttawa Faculty of Arts for funding various stages of the project.

References

Bowker, L. 2011. "Off the record and on the fly: Examining the impact of corpora on terminographic practice in the context of translation." In J. Munday, K. Wallmach and A. Kruger, eds. *Corpus-based Translation Studies: Research and Applications*. 211-236. Manchester: St. Jerome.

Cabré, M.T., C. Bach, R. Estopà, J. Feliu, G. Martínez and J. Vivaldi. 2004. "The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities." In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa and R. Silva, eds. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*. 87-90. Lisbon, Portugal, 26-28 May 2004. https://www-new.comp.nus.edu.sg/~rpnlpir/proceedings/lrec-2004/pdf/100.pdf. Consulted 4 July 2013.

Condamines, A. 2002. "Corpus analysis and conceptual relation patterns." *Terminology* 8(1): 141–162.

Condamines, A. 2008. "Taking genre into account when analysing conceptual relation patterns." *Corpora* 3(2): 115–140.

Condamines, A. and J. Rebeyrolle. 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 and D. Bourigault, eds. *Ingénierie des connaissances, évolutions récentes et nouveaux défis*. 225-241. Paris : Eyrolles. http://w3.erss.univ-tlse2.fr/textes/pagespersos/rebeyrol/Articles/condamines_rebeyrolle_2000_b.pdf. Consulted 4 July 2013.

Condamines, A. and J. Rebeyrolle. 2001. "Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results." In Bourigault, D., M.C. L'Homme and C. Jacquemin, eds. *Recent Advances in Computational Terminology*. 127-148. Amsterdam/Philadelphia: John Benjamins.

Dancette, J. and C. Réthoré. 2000. *Dictionnaire Analytique de la Distribution*. Montreal: Les Presses de l'Université de Montréal.

Dubuc, R. 2002. *Manuel pratique de terminologie, 4e édition*. Brossard, Quebec: Linguatech éditeur.

Faber, P., P. León Araúz and A. Reimerink. 2011. "Knowledge representation in EcoLexicon." In N. Talaván Zanón, E. Martín Monje and F. Palazón Romero, eds. *Technological Innovation in the Teaching and Processing of LSPs: Proceedings of TISLID'10*. 367-385. Madrid: UNED.

Faber, P. and A. San Martín. 2011. "Linking specialized knowledge and general knowledge in EcoLexicon." In Roche, C. et al, eds. *Actes de la*

conférence Terminologie & Ontologie : Théories et Applications (TOTh) 2011. 47-61. Annency, France, 26-27 May 2013. http://www.porphyre.org/toth/files/actes/TOTh-2011-actes.pdf. Consulted 4 July 2013.

Feliu, J., J.J. Giraldo, V. Vidal, J. Vivaldi and M.T. Cabré. 2004. "The GENOMA-KB project: A concept based term enlargement system." In R. Costa, L. Weilgaard, R. Silva and P. Auger, eds. *Proceedings of the Workshop on Computational and Computer-Assisted Terminology, in association with LREC 2004*. 32-35. Lisbon, Portugal, 25 May 2004. http://www.upf.edu/pdi/dtf/teresa.cabre/docums/ca04fel.pdf. Consulted 4 July 2013.

Gillam, L., M. Tariq and K. Ahmad. 2005. "Terminology and the construction of ontology." *Terminology* 11(1): 55-81.

León Araúz, P., A. Reimerink and P. Faber. 2011. *Environmental knowledge in EcoLexicon*. In K. Jassem, P. Fuglewicz and M. Piasecki, eds. *Proceedings of the Computational Linguistics Applications Conference*. 9-16. Jachranka, Poland, 17-19 October 2011. http://lexicon.ugr.es/pdf/leonetal2011b.pdf. Consulted 4 July 2013.

León Araúz, P., A. Reimerink and A. García-Aragón. 2013. "Dynamism and context in specialized knowledge." *Terminology* 19(1): 31-61.

L'Homme, M.C. 2012. "Using Explanatory and Combinatorial Lexicology to discover the lexical structure of specialized subject fields." In J. Apresjan et al., eds. *Words, Meanings and other Interesting Things. A Festschrift in Honour of the 80th Anniversary of Professor Igor Alexandrovic Mel'cuk*. 378-390. Moscow: RCK. http://www.ruslang.ru/doc/melchuk_festschrift2012/LHomme.pdf. Consulted 5 July 2013.

Maroto, N. and A. Alcina. 2009. "Formal description of conceptual relationships with a view to implementing them in the ontology editor Protégé." *Terminology* 15(2): 232-257.

Marshman, E. and M.-C. L'Homme. 2008. "Portabilité des marqueurs de la relation causale : étude sur deux corpus spécialisés." In F. Maniez, P. Dury, N. Arlin and C. Rougemont, eds. *Corpus et dictionnaires de langues de spécialité : actes des Journées du CRTT*. 87-110. Grenoble: Presses universitaires de Grenoble.

Marshman, E. and P. Van Bolderen. 2009. "Towards an integrated analysis of aligned texts: The CREATerminal approach." In M.-C. L'Homme and A. Alcina, eds. *Proceedings of Terminology and Lexical Semantics 2009*. Montreal, Canada, 19 June 2009. CD-ROM.

Marshman, E., J.L. Gariépy and C. Harms. 2012. "Helping language professionals relate to terms: Terminological relations and termbases." *JoSTrans Journal of Specialised Translation* 18: 30-56. http://www.jostrans.org/issue18/art_marshman.pdf. Consulted 4 July 2013.

Marshman, E., M.-C. L'Homme and V. Surtees. 2008. "Portability of cause-effect relation markers across specialized domains and text genres: A comparative evaluation." *Corpora* 3(2): 141-172.

Mel'čuk, I., A. Clas and A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Louvain la Neuve: Duculot.

Meyer, I. 2001. "Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework." In D. Bourigault, C. Jacquemin and M.-C. L'Homme, eds. *Recent Advances in Computational Terminology*. 279-302. Amsterdam/Philadelphia: John Benjamins.

Meyer, I., D. Skuce, L. Bowker and K. Eck. 1992. "Towards a new generation of terminological resources: an experiment in building a terminological knowledge base." *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92)*. 956-960. Nantes, France, 23-28 August 1992.

Pavel, S. and D. Nolet. 2001. *Handbook of Terminology*. Ottawa: Government of Canada, Translation Bureau. http://www.btb.gc.ca/publications/documents/termino-eng.pdf. Consulted 5 July 2013.

Robichaud, B. 2012. "Logic based methods for terminological assessment." In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk and S. Piperidis, eds. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC) 2012*. 94-98. Istanbul, Turkey, 21-27 May 2012. http://www.lrec-conf.org/proceedings/lrec2012/pdf/1096_Paper.pdf. Consulted 5 July 2013.

Roche, C. et al., eds. 2011. *Actes de la conférence Terminologie & Ontologie : Théories et Applications (TOTh) 2011*. Annency, France, 26-27 May 2013. http://www.porphyre.org/toth/files/actes/TOTh-2011-actes.pdf. Consulted 4 July 2013.

Sager. J.C. 1990. *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

Temmerman, R. and K. Kerremans. 2003. "Termontography: Ontology building and the sociocognitive approach to terminology description." In E. Hajičová, A. Kotěšovcová and J. Mírovský, eds. *Proceedings of the XVII International Congress of Linguists*. 9-16. Prague, Czech Republic, 24-29 July 2003. http://www.starlab.vub.ac.be/research/projects/poirot/Publications/temmerman_art_prague03.pdf. Consulted 4 July 2013.

Domain-independent term extraction through domain modelling

Georgeta Bordea
UNLP, DERI
National University
of Ireland, Galway
name.surname
at deri.org

Paul Buitelaar
UNLP, DERI
National University
of Ireland, Galway
name.surname
at deri.org

Tamara Polajnar
Computer Laboratory
University of Cambridge
name.surname
at cl.cam.ac.uk

Abstract

Extracting general or intermediate level terms is a relevant problem that has not received much attention in literature. Current approaches for term extraction rely on contrastive corpora to identify domain-specific terms, which makes them better suited for specialised terms, that are rarely used outside of the domain. In this work, we propose an alternative measure of domain specificity based on term coherence with an automatically constructed domain model. Although previous systems make use of domain-independent features, their performance varies across domains, while our approach displays a more stable behaviour, with results comparable to, or better than, state-of-the-art methods.

Term extraction plays an important role in a wide range of applications including information retrieval (Yang et al., 2005), keyphrase extraction (Lopez and Romary, 2010), information extraction (Yangarber et al., 2000), domain ontology construction (Kietz et al., 2000), text classification (Basili et al., 2002), and knowledge mining (Mima et al., 2006). In many of these applications the specificity level of a term is a relevant characteristic, but despite the large body of work in term extraction there are few methods that are able to identify general terms or intermediate level terms. Take for example the following structure from the AGROVOC vocabulary[1]: *resources* → *natural resources* → *mineral resources* → *lignite*, where *resources* is an upper level term, *natural resources* and *mineral resources* are intermediate level terms, and *lignite* is a leaf. Intermediate level terms are specific to a domain but are broad enough to be usable for summarisation and classification. Methods that make use of contrastive corpora to select domain specific terms favour the leaves of the hierarchy, and are less sensitive to generic terms that can be used in other domains.

Instead, we construct a domain model by identifying upper level terms from a domain corpus. This domain model is further used to measure the coherence of a candidate term within a domain. The underlying assumption is that top level terms (e.g., *resource*) can be used to extract intermediate level terms, in our example *natural resources* and *mineral resources*. Our method for constructing a domain model is evaluated directly through an expert survey as well as indirectly based on its contribution to intermediate level term extraction. While domain modelling is tested and exemplified with English, the ideas presented here are not language dependent and can be applied to other languages, but this is outside the scope of this work.

We start by giving an overview of related work in term extraction in Section 1. Then, an approach to construct a domain model based on domain coherence is proposed in Section 2, followed by a method to apply domain models for term extraction. The experimental part of the paper starts with a direct evaluation of a domain model through a user survey (Section 3). A first set of experiments is carried in a standard setting for term evaluation, while the second set of experiments is application-driven, using corpora annotated for keyphrase extraction, information extraction, and information retrieval. We conclude this paper in Section 4, giving a few directions for future work.

[1] AGROVOC: http://aims.fao.org/standards/agrovoc/about

Georgeta Bordea, Paul Buitelar, Tamara Polajnar

1 Related work

Methods for term extraction that use corpus statistics alone are faced with the challenge of distinguishing general language expressions (e.g., *last week*) from terminological expressions. A solution to this problem is to use contrastive corpora (Huizhong, 1986). Several contrastive measures are proposed including domain relevance (Park et al., 2002), domain consensus (Velardi et al., 2001), and word impurity (Liu et al., 2005). In this work we propose an approach to compute domain specificity based on a domain model, that is less sensitive to leaf terms and is better suited for intermediate level terms.

The domain model proposed in this work is derived from the corpus itself, without the need for external corpora. An automatic method for identifying the upper level terms of a domain has applications beyond the task of term extraction. Although not named as such, upper level terms were previously used for text summarisation (Teufel and Moens, 2002). The authors manually identified a set of 37 nouns including *theory*, *method*, *prototype* and *algorithm*, without considering a principled approach to extract them. The work presented here is similar to (Barrière, 2007), but instead of re-ranking terms based on their similarity to each other we make use of domain model terms, reducing data sparsity issues.

In our experiments we employ two state of the art methods for term extraction, the NC-value approach (Frantzi et al., 2000) and TermExtractor[2] (Velardi et al., 2001). The former is a hybrid method that ranks terms using only corpus statistics, while the latter exploits contrastive corpora. NC-value is based on raw frequency counts and considers nested multi-word terms by penalising frequency counts of shorter embedded terms. Additionally, it incorporates context information in a re-ranking step using top ranked terms. Context words (nouns, verbs and adjectives) are identified based on their occurrence with top candidates. Our method is an extension of this approach that uses domain models instead of selecting context words based on frequency alone.

TermExtractor is a popular approach that combines different term extraction techniques including domain relevance, domain consensus and lexical cohesion. Domain Relevance (DR) compares the probability of a term t in a given domain D_i with the maximum probability of the term in other domains used for contrast D_j and is measured as:

$$DR_{D_i}(t) = \frac{P(t/D_i)}{max_j\left(P(t/D_j)\right)}, j \neq i \quad (1)$$

Domain Consensus (DC) identifies terms that have an even probability distribution across the corpus that represents a domain of interest, and is estimated through entropy as follows:

$$DC_{D_i}(t) = -\sum_{d \in D_i} P(t/d) \cdot log\left(P(t/d)\right) \quad (2)$$

where d is a document in the domain D_i. Finally, the degree of cohesion among the words w_j that compose the term t is computed through a measure called Lexical Cohesion (LC). Let $|t|$ be the length of t in number of words, and $f(t, D_i)$ the frequency of t in the domain D_i, then Lexical Cohesion is defined as:

$$LC_{D_i}(t) = \frac{|t| \cdot f(t, D_i) \cdot log\left(f(t, D_i)\right)}{\sum_{w_j} f(w_j, D_i)} \quad (3)$$

The weight TE used for ranking terms by TermExtractor is a linear combination of the three methods described above:

$$TE(t, D_i) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC \quad (4)$$

While general terms typically have a high domain consensus, the domain relevance measure boosts narrow terms that have limited usage outside of the domain. For example the term *system* is not identified as relevant for Computer Science because it is frequently used in general language and in other specific domains as biology. In this work we take a different approach to compute domain specificity that can be applied for general terms by using a domain coherence measure that does not use external corpora. Two general purpose corpora, the Open American National Corpus[3] and a corpus of books from Project Gutenberg[4], are used as contrastive corpora for our implementation of TermExtractor. The books selected from

[2]TermExtractor demo: http://lcl.uniroma1.it/sso/index.jsp?returnURL=\%2Ftermextractor\%2F

[3]Open American National Corpus: http://www.americannationalcorpus.org/OANC/

[4]Project Gutenberg: http://www.gutenberg.org/

Project Gutenberg include the bible, the complete works of William Shakespeare, James Joyce's *Ulysses* and Tolstoy's *War and Peace*. We consider only the default setting of TermExtractor assigning equal weights to each measure in Equation 4.

2 Constructing a domain model based on domain coherence

We begin this section by describing an approach for domain modelling based on domain coherence in Section 2.1. Then, we discuss a modification of the NC-value approach which makes it better suited for intermediate level terms (Section 2.2). We conclude this section by describing a novel method for term extraction using a domain model in Section 2.3.

2.1 Domain modelling

A domain model is represented as a vector of words which contribute to determine the domain of the whole corpus. Let Δ be the domain model, and w_1 to w_n a set of generic words, specific to the domain, then:

$$\Delta = \{w_1, ..., w_n\} \quad (5)$$

The number of words n can be empirically set according to a cutoff associated weight. Previous work on using domain information for word sense disambiguation (Magnini et al., 2002) has shown that only about 21% of the words in a text actually carry information about the prevalent domain of the whole text, and that nouns have the most significant contribution (79.4%). Several assumptions are made to identify words that are used to construct a domain model from a domain corpus:

1. **Distribution:** Generic words should appear in at least one quarter of the documents in the corpus;
2. **Length:** Only single-word candidates are considered, as longer terms are more specific;
3. **Content:** Only content-bearing words are of interest (i.e., nouns, verbs, adjectives);
4. **Semantic Relatedness:** A term is more general if it is semantically related to many specific terms.

The distribution assumption implies that rare terms are more specific, similar with the frequency-based measure previously used for measuring tag generality (Benz et al., 2011). This might not always be the case, for example a simple search with a search engine shows that *artefact* or *silverware* are more rarely used than the term *spoon*, although the first two concepts are more generic. However, in this work we are interested in extracting basic-level categories as theorised in psychology (Hajibayova, 2013). A basic-level category is the preferred level of naming, that is the taxonomical level at which categories are most cognitively efficient. A counter example can be found for the length assumption as well, as the longer term *inorganic matter* is more general than the single word *knife*, but in this case we would simply consider as a candidate the single word *matter* which is more generic than the compound term. Both length and frequency of occurrence are proposed as general criteria for identifying basic-level categories (Green, 2005).

The first three assumptions are used for candidate selection, while the fourth assumption is used to filter the candidates. A possible solution for building a domain model is to use a standard termhood measure for single-word terms. Most approaches for extracting single-word terms make use of contrastive corpora, ranking higher specific words that are rarely used outside of the domain. But our domain model is further used for term extraction, therefore it is important that we use generic words to insure a high recall.

We interpret coherence as semantic relatedness to quantify the coherence of a term in a domain. The measure used for semantic relatedness is Pointwise Mutual Information (PMI). First, we extract multi-word terms using a standard term extraction technique, then we use the top ranked terms to filter candidate words using the following scoring function for domain coherence:

$$s(\theta) = \sum_{\sigma \in \Omega} PMI(\theta, \sigma) = \sum_{\sigma \in \Omega} log\left(\frac{P(\theta, \sigma)}{P(\theta) \cdot P(\sigma)}\right) \quad (6)$$

where θ is the domain model candidate, σ is top ranked multi-word term, Ω is the set of top ranked multi-word terms and $P(\theta, \sigma)$ is the probability that the word θ appears in the context of the term σ. In our implementation, the set Ω contains the best terms extracted by our baseline term extraction method described in Section 2.2, but any other term extraction method can be applied in this step. A small sample from domain models extracted us-

Computer Science	Biomed	Food and Agriculture
development	mechanism	control
software	evidence	farm
framework	antibody	supply
information	molecule	food
system	system	forest

Table 1: Example words from domain models extracted for different domains

ing our domain coherence method for Computer Science, Food and Agriculture, and the Biomedical Domain, is shown in Table 1.

2.2 Baseline term extraction method

Our baseline approach for intermediate level term extraction is frequency-based, similar to the C-value method (Ananiadou, 1994), but we modify its ranking function. The main difference is the way we take into consideration embedded terms. In previous work, this information is used to decrease frequency counts, as shorter terms are counted both when they appear by themselves and when they are embedded in a longer term. We argue that the number of longer terms that embed a term can be used as a termhood measure. In our experiments, this measure only works for embedded multi-word terms, as single-word terms are too ambiguous. The baseline scoring method b is defined as:

$$b(\tau) = |\tau| \log f(\tau) + \alpha e_\tau \quad (7)$$

where τ is the candidate string, $|\tau|$ is the length of τ, f is its frequency in the corpus, and e_τ is the number of terms that embed the candidate string τ. The parameter α is used to linearly combine the embeddedness weight and is empirically set to 3.5 in our experiments.

2.3 Using domain coherence for term extraction

Although we proposed a method to build a domain model in Section 2.1, the question of how to use this domain model in a termhood measure remains unanswered. Again, the solution is to rely on the notion of domain coherence, which is defined in this case as the semantic relatedness between a candidate term and the domain model described above. The assumption is that a correct term should have a high semantic relatedness with representative words from the domain. This method favours more generic candidates than contrastive corpora approaches, therefore it is better suited for extracting intermediate level terms.

The same measure of semantic relatedness is used as for the domain model, the PMI measure. The domain coherence DC of a candidate string τ is defined as follows:

$$DC(\tau) = \sum_{\theta \in \Delta} PMI(\tau, \theta) \quad (8)$$

where θ is a word from the domain model, and Δ is the domain model constructed using Equation 6. Using generic terms to build the domain model is crucial for ensuring a high recall as these words are more frequently used across the corpus. In our implementation context is defined as a window of 5 words.

3 Experiments and Results

Evaluating term extraction results across domains is a challenge, because finding domain experts is difficult for more than one domain. An alternative is to reuse datasets annotated for applications where term extraction plays an important role, for example, keyphrase extraction or index term assignment. Three technical domain corpora are used in our experiments: *Krapivin*, a corpus of scientific publications in Computer Science (Krapivin et al., 2009); *GENIA*, a corpus of abstracts from the biomedical domain (Ohta et al., 2001); and *FAO*, a corpus of reports about Food and Agriculture (Medelyan and Witten, 2008) collected from the website of the Food and Agriculture Organization of the United Nations[5]. The *Krapivin* corpus provides author and reviewer assigned keyphrases for each publication. The *GENIA* corpus is exhaustively annotated with biomed terms, with about 35% of all noun phrases annotated as biomed terms. The *FAO* dataset provides index terms assigned to each document by professional indexers. It is not only the document size that varies considerably across these three corpora, but also the number of annotations assigned to each document as can be seen in Table 2.

We evaluate our measure for building a domain model in Computer Science, by identifying a list of general words with the help of a domain expert in Section 3.1. We envision two sets of experiments: a standard term extraction evaluation

[5]Food and Agriculture Organization of the United States: http://www.fao.org

Corpus	Documents	Tokens	Avg. Annotations
Krapivin	2304	$22 \cdot 10^6$	5
GENIA	1999	$0.5 \cdot 10^6$	37
FAO	780	$28 \cdot 10^6$	8

Table 2: Corpora statistics

where the top ranked terms are evaluated against the list of unique annotations provided in the evaluation datasets (Section 3.2.1), and a second set of experiments where each term extraction approach is used to assign candidates to documents in combination with a document relevance measure in Section 3.2.2.

3.1 Intrinsic evaluation of a domain model

A domain expert was asked to investigate nouns used in the ACM Computing Classification System[6]. The expert was provided with the list of nouns and their frequency in the taxonomy and was required to identify nouns that refer to generic concepts. A set of 80 nouns were selected in this manner including *system*, *information*, and *software*. Only one annotator was involved because of the complexity of the task, that implies the analysis and filtering of several hundred words. We estimate the inter-annotator agreement by analysing a subset of the selected words through a survey with 27 participants. A quarter of the selected words are combined with the same number of randomly selected rejected words and the resulting list is sorted alphabetically. The Fleiss kappa statistic for interrater agreement is 0.34, lying in the fair agreement range. 80% of the words from our gold standard domain model were selected by at least half of the participants.

We compare our method (DC) with two other benchmarks, the contrastive termhood measure used in TermExtractor, and the frequency-based method used by NC-value to select context words ($NCVweight$). Again, context is defined as a window of 5 words. A domain model has many similarities with probabilistic topic modelling, although it provides less structure. We compare our approach with a popular approach to topic modelling, Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We experimented with different numbers of topics but we report only the best results

[6]ACM Computing Classification System: http://www.acm.org/about/class/1998/

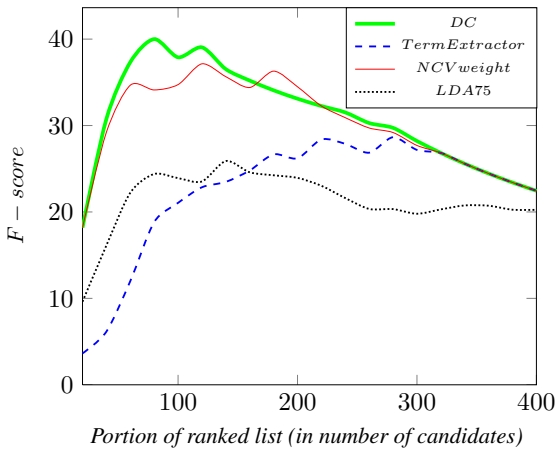


Figure 1: Methods for extracting a domain model

achieved for 75 topics ($LDA75$).

The results of this experiment are shown in Figure 1, in terms of F-score. Several conclusions can be drawn from this experiment. First, the methods that analyse the context of top ranked terms (i.e., our domain coherence measure, DC, and the weight used for context words in the NC-value, w_{NCV}) perform better than the contrastive measure used in TermExtractor, with statistically significant gains. Also, our domain coherence method outperforms the more simple frequency-based weight used in NC-value, although this result is not statistically significant. As expected, the words ranked high by TermExtractor are too specific for a generic domain model. The topic modelling approach identifies several words from the gold standard but much less than our approach and these are evenly distributed across latent topics. These conclusions will be further investigated across two other domains, using gold standard terms annotated for three different applications in Section 3.

3.2 Term extraction evaluation results

We implement and compare the baseline method presented in Section 2.2 and the method based on domain coherence described in Section 2.3, against the NC-value and TermExtractor methods, which are used as benchmarks. The same candidate selection method is used for all the evaluated approaches. Candidate terms are selected through syntactic analysis by defining a syntactic pattern for noun phrases. To assure the results are comparable, the same number of context words is used

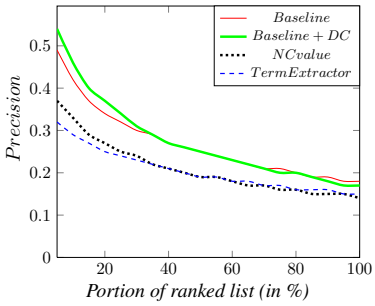


Figure 2: Precision for top 10k terms from the Krapivin corpus

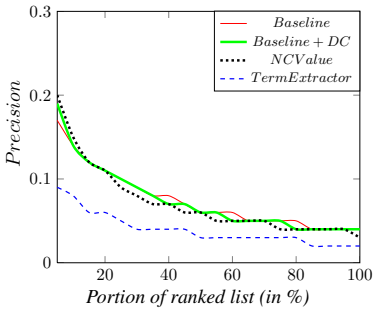


Figure 3: Precision for top 10k terms from the FAO corpus

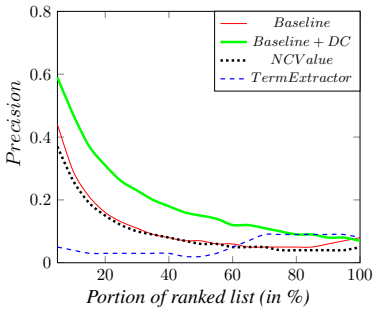


Figure 4: Precision for top 10k terms from the GENIA corpus

in our implementation of the NC-value approach as the size of the domain model. Two general purpose corpora, the Open American National Corpus[7] and a corpus of books from Project Gutenberg[8], are used as contrastive corpora for our implementation of TermExtractor. We considered only the default setting for TermExtractor, assigning equal weights to each measure.

3.2.1 Standard term extraction evaluation

While keyphrases and index terms suit well our purposes, as they are terms of an intermediate level of specificity, meant to summarise or classify documents, many of the terms annotated in GENIA are too specific. We discard the annotated terms that are mentioned in less than 1% of the documents from corpus, based on our distribution assumption. For each of the three datasets, the top ten thousand ranked terms were evaluated. We incrementally analysed portions of the ranked lists computed using the baseline approach (*Baseline*), the baseline approach linearly combined with the domain coherence measure (*Baseline+DC*), and the two benchmarks, *NC-value* and *TermExtractor*. The precision value for a portion of the list is scaled against the overall number of candidates considered. First, we observe that all methods perform better on the GENIA (Figure 4) and the Krapivin corpus (Figure 2), with the best methods achieving a maximum precision close to 60% at the top of the ranked list.

The Food and Agriculture use case is more challenging, as the best method achieves a precision of less than 20%, as can be seen in Figure 3. Also, the contrastive corpora measure employed in TermExtractor yields considerably worse results on all three domains, because the extracted terms are too specific. The baseline method, that rewards embedded terms, outperforms the NC-value method on the Computer Science domain, and in the biomedical domain, but it performs slightly worse on the Agriculture domain. The combination of our baseline method with the domain coherence measure (referred to as *Baseline + DC* in the legend) yields the most stable behaviour, outperforming all other measures across the three domains, considerably so in the biomedical domain (Figure 4) and at the top of the ranked list in Computer Science (Figure 2). In Computer Science, domain coherence significantly outperforms the best performing state-of-the-art method, NC-value (Figure 2). In Biomedicine, the improvement is statistically significant, with a gain of 106% at top 20% of the list (Figure 4).

3.2.2 Application-based evaluation

An important reason for developing termhood measures is that they are needed in specific applications, for example keyphrase extraction and index term extraction. Typically, a termhood measure is combined with different measures of document relevance in such applications, as the candidates are assigned at the document level. We make use of the standard information retrieval measure *TF-IDF* in combination with the considered term extraction scoring functions to assign terms to documents. The best results are obtained by using domain coherence as a post-processing step. In this experiment, the *PostRankDC* approach was computed by re-ranking the top 30 candidates selected using our baseline approach described in

[7]Open American National Corpus: `http://www.americannationalcorpus.org/OANC/`

[8]Project Gutenberg: `http://www.gutenberg.org/`

Top	F@5	F@10	F@15	F@20
Baseline	12.24	12.81	12.14	11.32
PostRankDC	**13.42**	**14.55**	**13.72**	**12.51**
NC-value	6.77	7.32	7.18	6.75
TermExtractor	1.41	1.77	1.95	1.97

Table 3: Keyphrase extraction evaluation on the Krapivin corpus

Top	F@5	F@10	F@15	F@20
Baseline	3.17	3.76	4.03	4.20
PostRankDC	**5**	5.8	5.62	5.29
NC-value	4.65	**5.88**	**6.09**	**5.94**
TermExtractor	0.2	0.31	0.34	0.35

Table 4: Index term evaluation on the FAO corpus

Top	F@5	F@10	F@15	F@20
Baseline	9.67	15.71	20.17	23.19
PostRankDC	**11.36**	17.63	21.52	23.55
NC-value	7.79	11.97	14.01	14.6
TermExtractor	10.77	**17.75**	**22.14**	**24.63**

Table 5: Term extraction at the document level on the GENIA corpus

Equation 7, based on their domain coherence.

The application-based evaluation proposed in this work allows us to evaluate both precision and recall, and consequently F-score can be used as an evaluation metric. The results for keyphrase extraction in Computer Science are presented in Table 3, while the results for index term extraction in the Agriculture domain are shown in Table 4. The results for document level term extraction from the Biomed corpus appear in Table 5. All three methods yield a higher performance on the GENIA corpus. The results on the Agriculture corpus are again the lowest, because a larger number of candidates has to be analysed.

Our $Baseline$ method outperforms the NC-value approach on the Krapivin corpus and on the GENIA corpus, but not on the FAO corpus. We can observe that the domain coherence approach ($PostRankDC$) improves over our baseline approach ($Baseline$) on all three domains. The improvement is statistically significant compared to the best state-of-the-art method in Computer Science, NC-value. NC-value outperforms TermExtractor in Computer Science and Agriculture, but TermExtractor performs better in Biomedicine. Although both NC-value and TermExtractor make use of domain-independent features for ranking, their performance varies across domains and applications. At the same time, combining our domain coherence approach (*PostRankDC*) with our baseline method in a post-ranking step displays a more stable behaviour, achieving the best performance on the Computer Science domain (Krapivin) and similar results with the results of the best method in Biomedicine (GENIA) and Agriculture (FAO).

4 Conclusions

In this study, we proposed an approach to identify intermediate level terms through domain modelling and a novel domain coherence measure, arguing that approaches that make use of contrastive corpora are only suitable for updating existing terminology resources with more specific terms and not for summarisation or classification tasks. The contributions described in this work are three-fold:

i) A method for extracting top level terms from a domain corpus

ii) A novel domain coherence metric based on semantic relatedness with a domain model

iii) A novel application-based evaluation for term extraction systems

Experiments discussed in this paper show that term extraction performance depends on the domain, although systems make use of domain-independent features. Our domain coherence approach based on a domain model performs well across domains, while the performance of the NC-value and TermExtractor benchmarks is more domain-dependent. The results lead to the conclusion that using a domain model is more appropriate than using statistical approaches based on contrastive corpora, for extracting intermediate level terms. Future work will include an unsupervised learning-to-rank approach for term extraction, that will allow a more principled integration of domain coherence measures with standard term extraction features. The method proposed here can be used as a specificity measure, and we currently investigate this in the context of constructing generalisation hierarchies of concepts.

Acknowledgments

This work has been funded in part by the European Union under Grant No. 258191 for the PROMISE

project, as well as by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

Sophia Ananiadou. 1994. A methodology for automatic term recognition. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, page 10341038, Kyoto, Japan.

Caroline Barrière. 2007. Une perspective interactive à lextraction de termes. In *7ème Confèrence" Terminologie et intelligence artificielle"*, pages 95–104.

Roberto Basili, Alessandro Moschitti, and Maria Teresa Pazienza. 2002. Empirical investigation of fast text classification over linguistic features. In Frank van Harmelen, editor, *ECAI*, pages 485–489. IOS Press.

Dominik Benz, Christian Krner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. 2011. One tag to bind them all: Measuring term abstractness in social metadata. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter Leenheer, and Jeff Pan, editors, *The Semanic Web: Research and Applications*, volume 6644 of *Lecture Notes in Computer Science*, pages 360–374. Springer Berlin Heidelberg.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms : the C-value / NC-value method. *Journal on Digital Libraries, Natural language processing for digital libraries*, 3 (2):115–130.

Rebecca Green. 2005. Vocabulary alignment via basic level concepts. In *Final Report, 2003 OCLC/ALISE Library and Information Science Research Grant Project*, Dublin, OH: OCLC.

Lala Hajibayova. 2013. Basic-level categories: A review. *Journal of Information Science*.

Y Huizhong. 1986. A new technique for identifying scientific/technical terms and describing science texts. *Lit. Linguist. Comput.*, 1:93–103, April.

Jörg-Uwe Kietz, Raphael Volz, and Alexander Maedche. 2000. Extracting a domain-specific ontology from a corporate intranet. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning - Volume 7*, ConLL '00, pages 167–175, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikalai Krapivin, Aliaksandr Autayeu, and Maurizio Marchese. 2009. Large dataset for keyphrases extraction. In *Technical Report DISI-09-055, DISI*, University of Trento, Italy.

Tao Liu, X Wang, Guan Yi, Zhi-Ming Xu, and Qiang Wang, 2005. *Domain-Specific Term Extraction and Its Application in Text Classification*, volume 1481, pages 1481–1484.

Patrice Lopez and Laurent Romary. 2010. HUMB : Automatic Key Term Extraction from Scientific Articles in GROBID. In *Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010)*, number July, pages 248–251.

Bernardo Magnini, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8:359–373.

Olena Medelyan and Ian H. Witten. 2008. Domain independent automatic keyphrase indexing with small training sets. *J. Am. Soc. Information Science and Technology*.

Hideki Mima, Sophia Ananiadou, and Katsumori Matsushima. 2006. Terminology-based knowledge mining for new knowledge discovery. *ACM Trans. Asian Lang. Inf. Process.*, 5(1):74–88.

Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, Sang-Zoo Lee, and Jun'ichi Tsujii. 2001. Genia corpus: A semantically annotated corpus in molecular biology domain. In *Proceedings of the ninth International Conference on Intelligent Systems for Molecular Biology (ISMB 2001) poster session*, page 68, July.

Youngja Park, Roy J. Byrd, and Branimir Boguraev. 2002. Automatic glossary extraction: Beyond terminology identification. In *19th International Conference on Computational Linguistics - COLING 02*, Taipei, Taiwan, August-September. Howard International House and Academia Sinica.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28:2002.

Paola Velardi, Michele Missikoff, and Roberto Basili. 2001. Identification of relevant terms to support the construction of domain ontologies. In *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*, Toulouse, July.

Lingpeng Yang, Dong-Hong Ji, Guodong Zhou, and Nie Yu. 2005. Improving retrieval effectiveness by using key terms in top retrieved documents. In David E. Losada and Juan M. Fernndez-Luna, editors, *ECIR*, volume 3408 of *Lecture Notes in Computer Science*, pages 169–184. Springer.

Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. 2000. Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 940–946, Stroudsburg, PA, USA. Association for Computational Linguistics.

An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri

Natalia Loukachevitch
Lomonosov Moscow State University
louk_nat@mail.ru

Michael Nokel
Lomonosov Moscow State University
mnokel@gmail.com

Abstract

Models for effective term extraction can depend on the type of a terminological resource under construction. In this paper we study term extraction models for real-working information-retrieval thesauri. The first thesaurus is the English version of EuroVoc thesaurus, the second one is the Russian Banking thesaurus. We study single-word and two-word term extraction separately to reveal the best features and feature combinations, compare best models for two thesauri. In particular, we found for this type of terminological resources that the use of association measures does not improve the quality of two-word term extraction based on combining multiple features.

1 Introduction

Automatic term extraction from texts of a specific domain is one of the well-studied applications in natural language processing and document analysis. During many years of research a lot of useful features of domain term extraction were proposed, including frequency-based and context-based features, word association measures, etc. ((Daille, 1995), (Zhang, 2008)).

Since these features characterize various properties of terms, machine-learning models based on multiple features are now increasingly used for term extraction. It was shown that such models can work considerably better than those based on single features ((Aze et al., 2005), (Loukachevitch, 2012)). Nevertheless, the significance of particular features for term extraction by machine learning depends on several important aspects concerning the domain in the target text collection, structure of extracted terms, and type of a terminological resource to be developed.

Firstly, specific domains vary in their scope (e.g., the broad social-political domain vs. the relatively narrow banking domain). Besides, domain-specific languages vary in their closeness to the general language (e.g. banking vs. immunology domain). This enhances or diminishes the role of a reference text collection required to calculate some term features (usually, a news collection or a national corpus is used).

Secondly, terms may be single-word and multi-word. To extract single-word terms, word association measures (mutual information, t-score, etc.) are not applicable; extraction of three-word and longer terms requires special forms of association measures. It means that extraction models for terms of different lengths can differ.

At last, terms are extracted for various types of terminological resources: terminological dictionaries, information-retrieval thesauri, ontologies for NLP. Dictionaries are mainly intended to supply terms with definitions, whereas information-retrieval thesauri are to provide concepts (descriptors) for domain-specific applications (Z39.19, 2005).

For example, such terms from EuroVoc information-retrieval thesaurus as *agricultural product, milk product, European party, economic consequence* denote important concepts in the contemporary socio-political life of European Union, however, it is difficult to imagine these terms as entries in terminological dictionaries. Therefore, a particular type of a terminological resource needs specialized term extraction models (Loukachevitch, 2012).

In this paper we consider the term extraction task specially for thesauri intended to be used in the information retrieval context (search, categorization, clustering and other applications),

because we suppose that such terminological resources have specific properties partially explained in specialized standards (Z39.19, 2005).

For this task we experimentally study machine-learning models based on various features for term extraction. Our study is based on two manually created thesauri and two languages: the English version of Eurovoc thesaurus and the Russian Banking thesaurus. We restrict our study to single-word and two-word terms to compare the extraction models for the most frequent types of terms.

2 Related Work

Machine-learning or combined approaches to automatic term extraction were studied in a number of works: (Vivaldi et al., 2001), (Aze et al., 2005), (Foo and Merkel, 2010), (Zhang, 2008), (Loukachevitch, 2012).

In most works automatically extracted terms are evaluated on the basis of available terminological resources or expert annotations of domain terms ((Daille, 1995), (Church and Hanks, 1990), (Dunning, 1993), (Church and Gale, 1995)). If to consider evaluation of machine-learning models for term extraction, in (Aze et al., 2005) experiments were fulfilled for texts in biological and human resources domains with expert annotation of domain terms. In the work (Foo and Merkel, 2010) two patent collections with term pre-annotation were studied. (Zhang, 2008) extracted terms from the Genia corpus, for which Genia ontology was created, and also utilized an artificial corpus of Wikipedia articles with expert annotation of terms.

In contrast to the above-mentioned works, in our study of term extraction we focus on the specific type of terminological resources – thesauri intended for information-retrieval applications. We take the well-known terminological resource EuroVoc and Banking thesaurus created for the Central Bank of the Russian Federation. Both resources are used in indexing and retrieval of documents in real information-retrieval systems.

3 Resources

3.1 EuroVoc Thesaurus and Europarl Text Collection

For the English part of our study we took EuroVoc thesaurus and Europarl parallel corpus. EuroVoc is an official thesaurus of the European Union and is intended for manual indexing of EU parliamentary documents. It is a multidisciplinary thesaurus covering the EU activities and containing terms in 22 languages of the EU. The English version of EuroVoc comprises 15161 terms (`http://eurovoc.europa.eu/drupal`).

The Europarl parallel corpus was extracted from the proceedings of the European Parliament (`http://www.statmt.org/europarl/`). The English part includes almost 54 mln. words.

In fact, EuroVoc thesaurus is intended just for the description of Europarl documents. Therefore, we can model how EuroVoc thesaurus could be developed from the Europarl corpus. EuroVoc represents a broad socio-political domain, and its language is close to general English.

3.2 Banking Thesaurus for the Central Bank and Articles from Online Magazines

For the Russian part of our study we took the Banking thesaurus created for the Central Bank of the Russian Federation. It is used in an information-retrieval system for indexing, search and vizualization of information and as a basis for text categorization. The thesaurus includes about 15 thousand terms and comprises the terminology of banking activity, banking regulation, monetary politics and macroeconomics.

As an appropriate text collection we took 10422 Russian articles from various on-line magazines: Auditor, RBC, Banking Magazine, etc. These documents contain almost 15.5 mln. words.

Since the banking thesaurus is used in real information retrieval tasks, we can model how it could be developed from the banking text corpus. In contrast to the broad socio-political domain of EuroVoc, this thesaurus represents relatively narrow banking domain, and its language is not so close to general language.

4 Features for Term Extraction

In our study we investigated single-word and two-word term extraction separately in order to have possibility to compare corresponding extraction models. As single-word term candidates we consider only *Nouns* and *Adjectives* (for Russian language) and *Nouns* (for English language); as two-word candidates we consider only *Adjective + Noun* and *Noun + Noun* (for Russian language) and *Adjective + Noun*, *Noun + Noun*, and *Noun +*

of + Noun (for English language) since they cover the majority of terms.

We use several types of enough known features for term extraction proposed in previous works and relatively new topic-based features proposed in (Bolshakova et al., 2013).

4.1 Traditional Features

The first type of traditional features is **frequency-based features**. The main assumption is that terms differ in their frequency and the distribution from other words in the target corpus. We consider the following 8 features: *Term Frequency in the collection (tf), Document Frequency (df), TF-IDF, TF-RIDF* (Church and Gale, 1995), *Domain Consensus* (Sclano and Velardi, 2007), *Term Contribution, Term Variance Quality, Term Variance* (Liu et al., 2005).

The second type of traditional features is based on the **target and reference corpora** and supposes that term frequencies in the target and reference corpora should be significantly different. We consider 9 such features, namely: *Weirdness* (Ahmad et al., 1999), *corpus-based TF-IDF* (where *TF* is taken from the target corpus, and *IDF* is taken from the reference corpus), *Relevance* (Peñas et al., 2001), *Contrastive* (Basili et al., 2001) and *Discriminative* (Wong et al., 2007) *Weights*, *Lexical Cohesion* (Park et al., 2002), *Reference Weight, KF-IDF* (Kurz and Xu, 2002), *Loglikelihood* (Gelbukh et al., 2010). In our study n-gramm statistics from British National Corpus (`http://www.natcorp.ox.ac.uk/`) and Russian National Corpus (`http://www.ruscorpora.ru`) were used as statistical data of a reference corpus for English and Russian collections correspondingly.

The third type of traditional features comprises **word-association measures** estimating mutual correlation of term candidate usage. They are primarily intended for two-word collocation extraction and are not applicable for single-word term extraction. We consider 19 word association measures: *Mutual Information (MI)* (Church and Hanks, 1990), *Augmented MI* (Zhang, 2008), *Cubic MI* (Daille, 1995), *Normalized Pointwise MI* (Bouma, 2009), *True MI, Dice Coefficient (DC)* (Smadja et al., 1996), *Modified DC, Generalized DC* (Park et al., 2002), *T-Score, Z-Score, Symmetric Conditional Probability* (Lopes and Silva, 1999), *Simple Matching Coefficient, Kulczinksy Coefficient, Ochiai Coefficient, Yule Coefficient, Jaccard Coefficient* (Daille, 1995), *Chi Square, Loglikelihood Ratio* (Dunning, 1993), *Gravity Count* (Daudarvičius and Marcinkevičienė, 2005).

The last type of traditional features is **context-based features** that account for phrases encompassing term candidates and their left and/or right context. We define a context of a term candidate as the bounds of encompassing noun phrases. In our study 11 known context-based features were considered: *C-Value, NC-Value* (Frantzi and Ananiadou, 1994), *MNC-Value, Token-LR, Token-FLR, Type-LR, Type-FLR* (Nakagawa and Mori, 2003), *Sum3, Sum10, Sum50, Insideness* (Loukachevitch, 2012).

Besides, we propose a novel context-based feature: *Modified Gravity Count (MGCount)*. It is based on Gravity Count association measure described in (Daudarvičius and Marcinkevičienė, 2005). *MGCount* for xy phrase is calculated as follows:

$$MGCount = \log\left(\frac{f(xy)l(x)}{f(x)} + \frac{f(xy)r(y)}{f(y)}\right) \quad (1)$$

where $f(x)$ is the frequency of x, $f(y)$ is the frequency of y, $f(xy)$ is the frequency of xy phrase, $l(x)$ is the number of different words to the left of x, and $r(y)$ is the number of different words to the right of y; $l(x)$ and $r(y)$ are considered only within the bounds of encompassing noun phrases. Our modification changed internal proportion $\frac{r(x)}{f(x)}$ to external proportion $\frac{l(x)}{f(x)}$ (and the same with the second component of the sum), thus the measure was transformed from the association measure to the context one.

4.2 Topic-Based Features

The next type comprises features based on so-called **topic models** (Blei and Lafferty, 2009). Topic models are intended to describe texts in terms of their topics, they determine, which topics are related to each document, and which words (or phrases) form each topic. In fact, each topic is represented as a list of frequently co-occurring words (or bigrams) ordered by descending degree of belonging to it. As an example, the first five words and bigrams from the top of four randomly selected topics of the English corpus along with

their probabilities of belonging are presented in the Table 1.

Topic #1		**Topic #2**	
Single-word	*Probability*	*Two-word*	*Probability*
Latin	0.021	European union	0.012
America	0.02	Young people	0.005
American	0.012	European council	0.004
United	0.009	United state	0.003
State	0.007	Youth program	0.002
Topic #3		**Topic #4**	
Single-word	*Probability*	*Two-word*	*Probability*
Audiovisual	0.013	Central bank	0.005
Film	0.011	European central	0.003
Television	0.01	Natural resource	0.002
Medium	0.008	Novel food	0.002
Broadcasting	0.006	Monetary policy	0.002

Table 1: Examples of revealed subtopics

Typically, there are two types of topic models: non-probabilistic ones that are based on hard clustering methods (K-Means, hierarchical agglomerative clustering, etc.) and probabilistic ones (PLSI, LDA, etc.) that represent each document as a mixture of topics and each topic is considered as a probabilistic distribution over words (Blei and Lafferty, 2009), (Bolshakova et al., 2013).

The topic-based features are relatively new and are obtained by revealing topics in the target text corpus. These features account for the idea that domain terms should usually correspond to some subtopics of the domain. As it was shown that *NMF (Non-Negative Matrix Factorization)* algorithm with KL-divergence minimization is the best topic model in terms of terminology extraction (Bolshakova et al., 2013), we applied it to reveal subtopics, as well as probabilities in them. Basically, given a non-negative term-document matrix V, this algorithm tries to find non-negative term-topic matrix W and topic-document matrix H, such that $V = WH$. We consider the version of NMF that minimizes Kullback-Leibler divergence $D(V||WH)$ (Lee and Seung, 2000).

We consider the following 7 topic-based features: *Term Frequency, TF-IDF, Domain Consensus, Maximum Term Frequency* (Bolshakova et al., 2013), *Term Score (TS)* (Blei and Lafferty, 2009), *TS-IDF, Maximum Term Score*. Most of these features are extensions of the standard frequency-based features applied to the revealed subtopics, considering probabilities of the term candidates in topics as frequencies (cf. Table 2; $P_i(w)$ denotes a probability of the term candidate w in the topic i, and K is the number of topics).

Feature	**Formula**
Term Frequency (TF)	$\sum_{i=1}^{K} P_i(w)$
TF-IDF	$TF(w) \times \log \frac{K}{DF(w)}$
Domain Consensus	$-\sum_{i=1}^{K}(P_i(w) \times \log P_i(w))$
Maximum TF	$\max_i P_i(w)$
Term Score (TS)	$\sum_{i=1}^{K} P_i(w) \log \frac{P_i(w)}{(\prod_{i=1}^{K} P_i(w))^{\frac{1}{K}}}$
TS-IDF	$TS(w) \times \log \frac{K}{TF(w)}$
Maximum TS	$\max_i TS_i(w)$

Table 2: Topic-based features

We also used 6 single-topic document features (documents are regarded as separate topics). In fact, we used all above-mentioned topic-based features except Domain Consensus, since this feature is already considered in the section of traditional frequency-based features (cf. section 4.1).

4.3 Other Features

Other features considered in our study include:

- 5 Linguistic features: *Ambiguity* (determines whether the term candidate has multiple initial forms or may belong to multiple parts of speech), *Novelty* (determines whether the term candidate is described in morphological dictionaries), *Specificity* (determines whether the term candidate exists in the reference collection), *Nouns* (determines whether the term candidate consists of only Nouns), and *Adjectives* (determines whether the term candidate contains Adjective).

- Features for term candidates that play subject syntactic role in sentences, features for term candidates beginning with a capital letter, and features for term candidates beginning with a capital letter that do not start sentences. We consider 6 features for each such group (and thus 18 features in the whole): namely, *Term Frequency, Document Frequency, TF-IDF, TF-RIDF, Domain Consensus*, and *corpus-based TF-IDF*;

- 2 features for term candidates that are in the context window of the several most frequent predefined ones: *NearTermsFreq,*

NearTermsFreq-IDF (Nokel et al., 2012). *NearTermsFreq* is defined as the number of the term candidate occurrences in the context window of the several predefined most frequent words.

- Average position of *the first occurrence in documents*, and *Term Length*.

Thus, 27 features belong to this group. To sum up, the full list of features comprises 69 features for single-word candidates and 88 features for two-word term candidates.

5 Experiments

We studied models for single-word and two-word term extraction from two above-described corpora: Russian banking electronic magazines, and English part of parallel corpus Europarl.

To extract single-word and two-word term candidates from these corpora, documents were processed by morphological analyzers. Thus, for English corpus we used Stanford POS tagger (`http://nlp.stanford.edu/software/corenlp.shtml`), while for Russian corpus we used our own morphological analyzer. Besides, from the set of extracted English term candidates we excluded words from the stop list created for the experiments (*other, another, that, this, those, mrs, sir, etc.*), and word pairs including stop-words were excluded as well.

Having extracted term candidates, we trained combined models comprising the above-described types of term features. The features were combined by Gradient Boosting machine learning algorithm, which proved to be the best one in our study. Namely, we used an open-source realization of this algorithm from `http://scikit-learn.org`. It is well-known that Gradient Boosting has a lot of parameters that need to be tuned. So, in all experiments we fixed all parameters, except the number of trees and maximum allowed depth of trees, that were tuned in each experiment individually. Besides, for training and evaluation four-fold cross validation was applied, which means that every time the training set was three-quarters of the whole list while the testing set was the remaining part.

A term extraction model has to find the best order, where real terms should be located at the beginning of the ordered list of term candidates. As an evaluation measure, we used Average Precision (AvP) often applied as a measure for term extraction (Zhang, 2008), (Bolshakova et al., 2013). It is defined for a set D of all term candidates with a subset of approved ones $D_q \subset D$ as follows:

$$AvP(D) = \frac{1}{|D_q|} \sum_{1 \le k \le |D|} (r_k \times (\frac{1}{k} \sum_{1 \le i \le k} r_i)) \quad (2)$$

where $r_i = 1$ if the i-th term $\in D_q$ and $r_i = 0$ otherwise.

At the first step of experiments we separately studied term extraction models for single-word and two-word terms. As baselines we considered several well-known features: *Weirdness, TF-IDF, C-Value* for single-word models and *TF-IDF, C-Value, Mutual Information* for two-word ones. In the Figures 1, 2, 3, 4 plots of AvP on various numbers of most frequent candidates are presented for

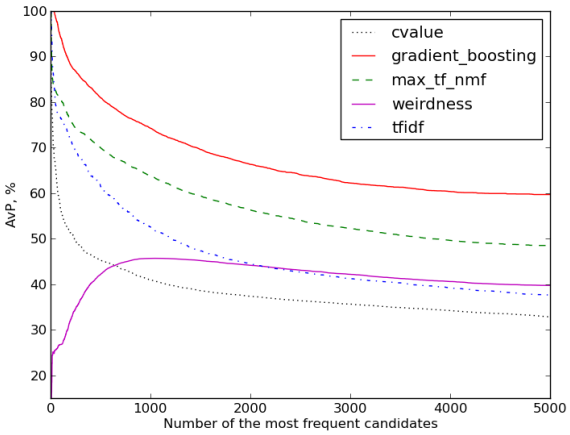


Figure 1: AvP for single-word Russian model

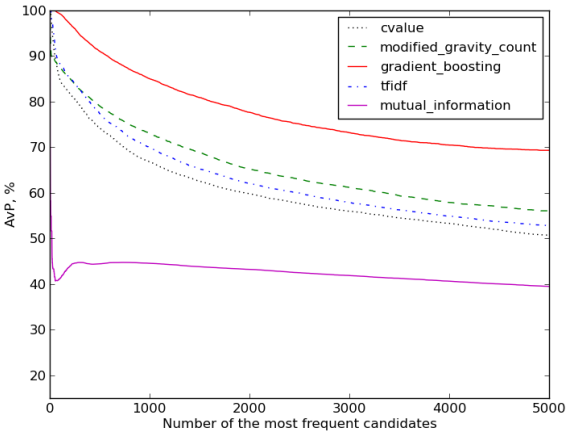


Figure 2: AvP for two-word Russian model

these baselines, the best single feature and the resulted model combined by Gradient Boosting.

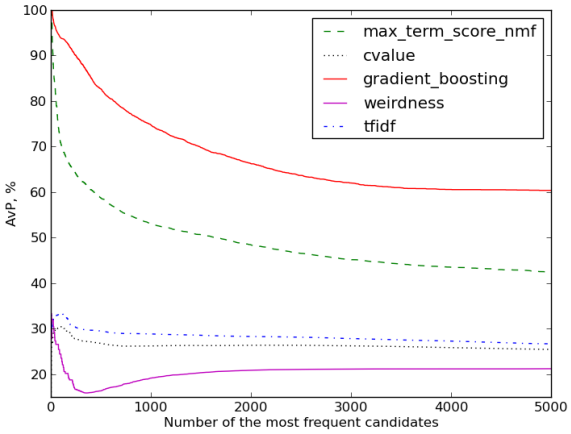


Figure 3: AvP for single-word English model

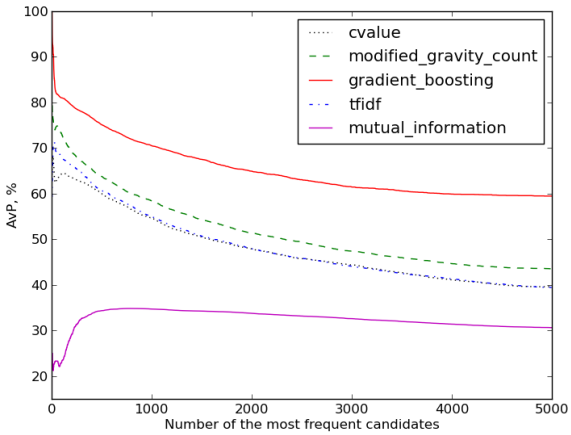


Figure 4: AvP for two-word English model

As we can see, the best single feature for single-word terms turned out to be a topic-based feature (either *Maximum Term Score* or *Maximum Term Frequency*), the performance of these features is considerably better than well-known baselines. So it seems that for single-word terms their relation to a domain subtopic is important.

The best single feature for two-word terms was found to be the novel context-based feature *Modified Gravity Count*. Besides, in all cases we can see the huge improvement of the combined model performance compared to well-known baselines and best single features.

At the second step of experiments we tried to determine the contribution of each above-described group of features to the whole combined model. We fixed the number of most frequent term candidates to 5000, excluded each of the following groups separately from the whole list: frequency-based features; features, based on the reference corpus; word association measures; context-based features, and topic-based features. The results of combining the remaining features by Gradient Boosting are presented in the Table 3.

Excluded group	**Average Precision (%)**			
	Single-word model		**Two-word model**	
	Russian	*English*	*Russian*	*English*
No (All features)	59.7	60.4	69.3	59.5
Frequency-based	59.5	59.6	68.9	58.6
Context-based	59.3	56.8	68.9	58.8
Reference corpus	57.5	59.6	68.3	55.6
Topic-based	56.8	59.4	68.9	60
Word association	–	–	69.3	59.7

Table 3: Contribution of feature groups to term extraction models

As we can see, features, which are based on the reference corpus, give the most significant contribution to the two-word term extraction models regardless of the subject domain and language.

Besides, the use of word association measures does not improve the quality of extraction of two-word terms. The latter conclusion contradicts the assumption of numerous studies that association measures should be useful for multi-word term extraction (Zhang, 2008), (Daille, 1995), (Kurz and Xu, 2002). From the other side, this conclusion can be quite evident because, for example, EuroVoc includes a lot of terms looking as compositional phrases with free separate usage of components (as *European party, European idea, economic consequence* etc.). Introduction of such terms into an information-retrieval thesaurus is possible due to multiple principles of term inclusion in information-retrieval thesauri (Z39.19, 2005).

At the last step of experiments we investigated both models for single-word and two-word term candidates together. We created a *unified model* for both types of term candidates, taking into account all features except association measures and obtaining as a result the *unified list* of candidates.

Then we created specific models separately for single-word and two-word term candidates. As the models are specialized, they can be potentially more efficient. We summed up resulted lists of extracted terms according to their probability values generated by Gradient Boosting, and in such a way obtained the *summed-up list* of term can-

didates. We should notice that in the case of the unified model there is more data to train it, so this model can be potentially very efficient too.

The comparison of AvP for these two models (for both corpora) shows that *summed-up model* slightly outperforms the *unified* one – cf. Figure 5.

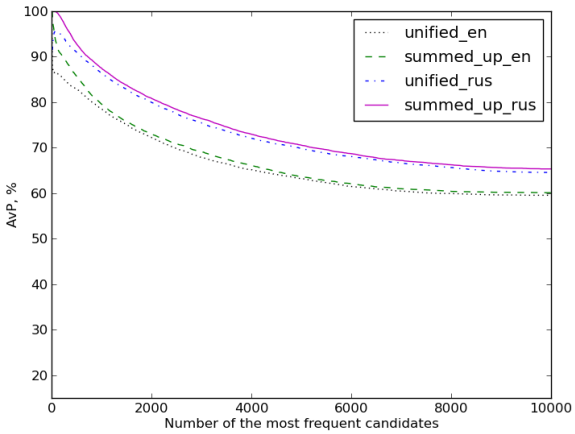


Figure 5: Unified vs summed-up models

In addition, as an example of the extracted term candidates, we present in the Table 4 the first 10 elements from the top of the term candidates lists created by unified models for Russian and English corpora (the elements in italics are real terms).

#	Russian corpus	English corpus
1	*Currency*	*Iran*
2	*Reporting period*	*Pakistan*
3	*Bond*	*Georgia*
4	Association	*India*
5	*Taxable period*	*Serbia*
6	*Reserve*	White paper
7	*Corporate governance*	*Syria*
8	*Credit history*	*Libya*
9	*Deal*	*Afghanistan*
10	*Borrower*	Member state

Table 4: Examples of term candidates extracted by unified models

The resulting unified models may be too complex in the number of applied features. Some of them may be redundant for Gradient Boosting and have no use in the models, make their training harder. In order to exclude them we applied a stepwise greedy *algorithm Add* for selecting the most significant features.

The algorithm starts with the empty set of features, and then at each step it adds the feature that maximizes the overall Average Precision, until there is any improvement between successive iterations. As a result, the combinations of only **13** features (out of total 69 features) were found for both corpora (see Table 5). We grouped similar features in the same rows of the table.

#	Russian corpus	English corpus
1	TF-RIDF Subjects	TF-RIDF Subjects
2	MGCount	MGCount
3	Lexical Cohesion	Lexical Cohesion
4	Nouns	Nouns
5	First Occurrence	First Occurrence
6	Weirdness	Weirdness
7	Corpus-based TF-IDF Non-Initial Words	TF-IDF Non-Initial Words
8	Sum3	Sum10
9	Term Score NMF	Maximum Term Score NMF
10	Single-topic TF-IDF	Single-topic Term Score
11	TF-RIDF	NearTermsFreq-IDF
12	KF-IDF	Term Variance Quality
13	TF-IDF NMF	Document Frequency

Table 5: Results of feature selection for unified models

Since there are representatives of all above-described groups in both found subsets of features, we conclude that each such group is significant for unified models of term extraction regardless of the scope and language. Besides, we can see that short models for both thesauri are quite similar.

6 Conclusion

In this paper we modelled single-word and two-word term extraction for the specific type of terminological resources – information-retrieval thesauri. Our experiments revealed features significant for extraction of single-word and two-word terms in the broad EuroVoc and relatively narrow banking domains. We showed that the best features for single term extraction in both cases are relatively new topic-based features, based on preliminary clustering of words in the target text collection. The context-based features are the most important for two-word term extraction.

The interesting result of our study is that the use of association measures does not improve the quality of term extraction models intended for information-retrieval thesaurus construction. It was also proved that the unified model can be applied to both single-word and two-word term extraction.

Acknowledgements

The work is partially supported by Dmitrii Zimin Dynastia Foundation with financial support of Yandex founders.

References

Ahmad K., Gillam L., Tostevin L. 1999. *University of Surrey Participation in TREC8: Indexing for Logical Document Extrapolation and Retrieval (WILDER)* Proceedings of TREC 1999.

Aze J., Roche M., Kodratoff Y., Sebag M. 2005. *Preference Learning in Terminology Extraction: A ROC-based Approach*. Proceedings of ASMDA'05, 209–219.

Basili R., Moschitti A., Pazienza M., Zanzotto F. 2001. *A Contrastive Approach to Term Extraction*. Proceedings of the 4th Terminology and Artificial Intelligence Conference.

Blei D. and Lafferty J. 2009. *Topic Models. Text Mining: Classification, Clustering and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Bolshakova E., Loukachevitch N., Nokel M. 2013. *Topic Models Can Improve Domain Term Extraction*. Proceedings of ECIR 2013.

Bouma G. 2009. *Normalized Pointwise Mutual Information in Collocation Extraction*. Proceedings of the Biennial GSCL Conference, 31–40.

Church K. and Gale W. 1995. *Inverse Document Frequency IDF: A Measure of Deviation from Poisson*. Proceedings of the Third Workshop on Very Large Corpora. MIT Press, 121–130.

Church K. and Hanks P. 1990. *Word Association Norms, Mutual Information, and Lexicography*. Computational Linguistics, vol. 16, 22–29.

Daille B. 1995. *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering* PhD dissertation. University of Paris, Paris.

Daudarvičius V. and Marcinkevičienė R. 2005. *Gravity Counts for the Boundaries of Collocations*. Corpus Linguistics, 9(2): 321–348.

Dunning T. 1993. *Accurate Metrics for the Statistics of Surprise and Coincidence*. Computational Linguistics, 19(1).

Foo J. and Merkel M. 2010. *Using Machine Learning to Perform Automatic Term Recognition*. Proceedings of the LREC 2010 Acquisition Workshop, Malta.

Frantzi K. and Ananiadou S. 1994. *The C-Value/NC-Value Domain-Independent Method for Multi-Word Term Extraction*. Journal of Natural Language Processing, vol. 6, no 3, 145–179.

Gelbukh A., Sidorov G., Lavin-Villa E., Chanona-Hermandez L. 2010. *Automatic Term Extraction using Log-likelihood based Comparison with General Reference Corpora*. Proceedings of the Natural Language Processing and Information Systems, 248–255.

Kurz D. and Xu F. 2002. *Text Mining for the Extraction of Domain Retrieval Terms and Term Collocations*. Proceedings of the International Workshop on Computational Approaches to Collocations.

Lee D. and Seung H. 2000. *Algorithms for Non-Negative Matrix Factorization* Proceedings of NIPS, 556–562.

Lopes G. and Silva J. 1999. *A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units*. Proceedings of the 6th Meeting on the Mathematics of Language, 369–381.

Loukachevitch N. 2012. *Automatic Term Recognition Needs Multiple Evidence*. Proceedings of LREC'12.

Liu L., Kang J., YU J., Wang Z. 2005 *A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering*. Proceedings of NLP-KE'05, 597–601.

Nakagawa H. and Mori T. 2003. *Automatic Term Recognition Based on Statistics of Compound Nouns and their Components*. Terminology, vol. 9, no. 2, 201–219.

Nokel M., Bolshakova E., Loukachevitch N. 2012. *Combining Multiple Features for Single-Word Term Extraction*. Proceedings of Dialog 2012, 490–501.

Park Y., Byrd R., Boguraev B. 2002. *Automatic Glossary Extraction: Beyond Terminology Identification*. Proceedings of the 19th International Conference on Computational Linguistics.

Peñas A., Verdejo V., Gonzalo J. 2001. *Corpus-based Terminology Extraction Applied to Information Access*. Proceedings of the Corpus Linguistics 2001 Conference, 458–465.

Sclano F. and Velardi P. 2007. *TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities*. Proceedings of the 7th Conference on Terminology and Artificial Intelligence.

Smadja F., McKeown K. R. and Hatzivassiloglou V. 1996. *Translating Collocations for Bilingual Lexicons: A Statistical Approach*. Computational Linguistics, 22(1), 1–38.

Vivaldi J., Marquez L, Rodriguez H. 2001. *Improving Term Extraction by System Combination using Boosting*. Proceedings of the 12th European Conference on Machine Learning, 515–526.

Wong W., Liu W., Bennamoun W. 2007. *Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency*. Proceedings of the 6th Australasian Conference on Data Mining, 47–54.

Zhang Z., Iria J., Brewster C., Ciravegna F. 2008. *A Comparative Evaluation of Term Recognition Algorithms*. Proceedings of LREC 2008.

Zhang Z., Yoshida T., Ho T. B., Tang X. 2008. *Augmented Mutual Information for Multi-Word Term Extraction*. International Journal of Innovative Computing, Information and Control, 8(2), 543–554.

Z39.19 2005. *Guidelines for the Construction, Formal and Management of Monolingual Thesauri* NISO.

Session : Short papers

Multilingual Problems in Navigation Terminology

Ayşe Yurdakul
Technische Universität Braunschweig
Institute for Traffic Safety and Automation Engineering
Langer Kamp 8
D-38106 Braunschweig
Tel.: 0531/391-3306
FAX: 0531/391-5197
yurdakul@iva.ing.tu-bs.de

Eckehard Schnieder
Technische Universität Braunschweig
Institute for Traffic Safety and Automation Engineering
Langer Kamp 8
D-38106 Braunschweig
Tel.: 0531/391-3317
FAX: 0531/391-5197
e.schnieder@tu-bs.de

Abstract

In modern times, technical progress accelerates the development of new disciplines and subdisciplines. That is why new and specific terminology becomes more necessary. The consequences of this interdisciplinarity are multilingual communication problems between non-professionals and experts of a special field or between experts of navigation domains of different transportation modes. Especially, these semantic problems include synonymy, antonymy, hypernymy-hyponymy relations and ambiguity etc. The main target of the *iglos*[1] terminology management system as terminological tool of the 21st century is to avoid the multilingual misunderstanding between special languages of different domains by comparing the definitions of technical terms in heterogeneous languages.

Keywords – terminology management, terminology, multilingual, semantical problems, localisation, positioning, position, location

1 Introduction

The main goal of *iglos* is to create a glossary for different technical languages or varieties at international level. Multilingual communication problems such as synonymy, antonymy, translation problems, ambiguity, risk of confusion etc. between experts of special fields intensify by the transforming technology. People extend and specialise their acquirements and tasks in the fields of technology. In general, the correct translation of a technical term is only possible if one knows the terminology of the special field.

The word pairs "location" and "localisation" and "position" and "positioning" are in the focus of our terminological contemplation. At this point, we want to analyse which relation exists between these terms in English, German and French. Finally, we want to display the equivalences in three languages.

2 Terminological Approaches in Multilingual Problems of Navigation with *iglos*

On the whole, the idea for *iglos* resulted from a cooperation of the Institute for Traffic Safety and Automation Engineering and the Department of German Linguistics of the Braunschweig University of Technology. "The project is interdisciplinary and it consistently grows in complexity and richness of perspective by our dialogue with linguists, terminologists, computer scientists, engineers, translators and users." (Arndt; Schnäpp; Schnieder; Yurdakul, 2013).

The main target of *iglos* is to develop a software platform on a linguistic basis. With this terminology management system, it is intended

[1] "iglos" is the intelligent glossary or terminology management system of the Institute for Traffic Safety and Automation Engineering of the Braunschweig University of Technology

to accelerate and facilitate a consistent, multilingual and unambiguous development of technical terminology for optimising the scientific and commercial communication. The foundation of the *iglos* system consists in a development of the variety-based trilateral sign model.

This collectively consists on the one hand of two lexemes and on the other hand of one relational lexeme which is placed between these both (Schnieder, 2012).

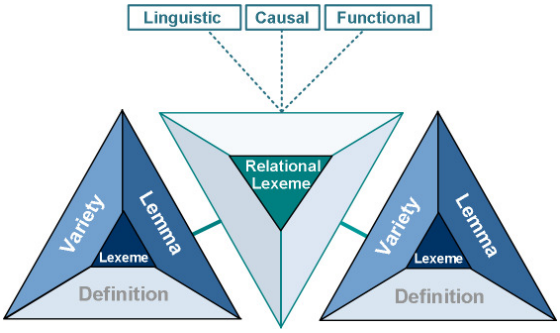


Figure 1: The *iglos* sign model

A lexeme is the abstract morphological unit or the unit of a word paradigm. In general, a lexeme is concretised by its grammatical word forms. In our case, terms are special lexemes.

Each lexeme consists of a lemma (namely a denomination such as "navigation"), a definition which relates to the context and a variety as a technical language (e.g. linguistics, technics, etc.). There is at least one relation between two lexemes. This is specified by at least one relational lexeme or a relation type.

Among these relation types, there are: **risk of confusion** (*isMixesUpWith*), **translation** (*hasTranslation*), **output** (*hasOutput, isOutputOf*), **input** (*hasInput, isInputOf*), **holonymy** (*hasPart, isPartOf*), **meronymy** (*isPartOf, hasPart)*, **antonymy** (*hasAntonym*), **synonymy** (*isSynonymOf*), **polysemy** (*isPolysemOf*), **homonymy** (*hasHomonym*)
Basically, the *iglos* sign model enables the specification of terminologies by avoiding terminological haziness and creating and visualising concrete relations between terms in a systematic context (variety). These relations are unobstructedly typable.

Besides, the merit is also to avoid synonymy and ambiguity (disambiguation) of terms. On the basis of the variety-based *iglos* sign model, there is a communication between different languages (multilingualism) on the one hand and between different domains (multidisciplinarity) on the other hand. In our consideration, we want to present the multilingual problems in relation to terminology in the domain of traffic engineering and which solution approaches can be proposed for them on the basis of *iglos*.

Within the framework of *iglos*, there are several methodological approaches for avoiding the linguistic problems e.g. between the terms "position", "positioning", "location" and "localisation" in English, German and French:

At least one definition of terms in English, German and French in **general linguistic usage**, in **etymologic, grammatical, normative technical and relational** perspective.

In the next step, it is important to find out the kinds of these semantical problems between the different terms in the navigation domain.

2.1 Kinds of Problems in Navigation Terminology

In the domain of navigation domain, we have raised the four terminological questions:

1. Are "position" and "location" or rather "positioning" and "localisation" synonyms in the languages mentioned above?
2. What is the causal or functional difference between "position" and "positioning" or rather between "location" and "localisation" in the languages mentioned above?
3. Are there exact translations for the four terms in each language mentioned above?
4. Are the denominations of the terms similar in the three languages?

2.2 Results of the Approaches

Multilingual problems between technical terms can be avoided or minimised by relating them with each other. According to normative sources (e.g. DIN standards), we tried to create relations between these terms. There are various relation types such as "synonymy", "antonymy", "part-whole relation", "hypernymy-hyponymy relation", "ambiguity", "translation" "sequence", "function", "risk of confusion", "converseness" etc.

On the basis of the methodological approaches, we have found out:

- There are several definitions for each term in dictionaries. Besides, the definition of the word pairs "position"/"positioning" and "location"/ localisation" are different (Oxford Dictionaries Online).
- Furthermore, both word pairs have an own etymologic origin. "Position" is borrowed from the Latin term "ponere" (English: to put, to set, to lay) whereas "location" is borrowed from the Latin term "locus" (English: place) (Online Etymology Dictionary).
- The analyse of "positioning" and "localisation" in relation to the grammatical aspect has shown that both distinguish by their own word forms which represent them (e.g. "positioning has a noun "position", a verb "to position" and "localisation" has a noun "location" and a verb "to locate".)
- In normative context, the four terms are also differently defined on several occasions (DIN EN 13848-4, DIN EN ISO 19148, IEC 60351-1: 1976, ISO 19134: 2007).
- Finally, the definitions of terms are analysed in the domain of traffic engineering and it could be found out that there are also different definitions for the four terms (Schnieder 2012, Schnieder & Becker 2007).
- The four terms are related with each other by relation types (e.g. by *iglos* relation types).

On the one hand, our contemplation provide the result that "positioning" and "localisation" or rather "position" and "location" are not synonyms. Firstly, they are sequent and secondly there is a risk of confusion between both. Whereas, "positioning" describes the process for achieving the place in a three-dimensional reference system, "localisation" is a large place in a two-dimensional reference system and enables local orientation on the surface of the earth (see Yurdakul; Schnieder; Hodon 2013).

"Position" is accordingly defined as the "data type that describes a point or geometry potentially occupied by an object or person" (ISO 19132, 2007) and "location" is the "identifiable geographic place" (ISO 19134, 2007) which must be localised by functional resources.

On the other hand, "position" and "location" are state nominations. In contrast to them, "positioning" and "localisation" are specific functions. A function is a specific process that a system is able to perform whereas a state is the entirety of all physical quantities of a physical system. In general, we can notice that there is a causal and functional coherence between these functions and states: "localisation" leads to "location" and "positioning" to "position" or "location" isOutputOf "localisation" and "position" of "positioning". In addition, "position" isInputOf "localisation" and "location" of "positioning". Finally, all terms together (as hyponyms) are related to "navigation" (figure 2).

2.3 Translation of Terms into German and French

Besides, English terms are often translated into German and French. On the basis of this schedule, it becomes obvious that the terms "position" and "positioning" have similar denominations in French and German whereas "location" and "localisation" are differently denominated. The term "localisation" has the same denomination in French, but not in German. In relation to "location", there are different denominations in each of the three languages.

In conclusion, the terms "position" and "positioning" are internationalisms which is a borrowed term and which occurs in several languages. "Location" and "localisation" are typically English terms but not especially German. Instead of "Ortung", there will be the possibility to use "Lokalisierung" as internationalism in German.

English	German	French
position	Position	position
location	Ort	lieu
positioning	Positionierung	positionnement
localisation	Ortung/ Lokalisierung	localisation

Table 1: Equivalences of terms in German and French

3 Conclusion

By defining and relating terms by unambiguous and consistent relation types, a terminology building can be constructed. There are various relations between different terms in this building.

Therefore, the *iglos* sign model can be described as an ontology structure with single lexemes are bound together by relations. Furthermore, *iglos* is a terminology management system of the next generation which collects and integrates different (technical) languages and guides terminologists to construct terminology.

Besides, there are various multilingual communication problems not only between experts of traffic engineering. Linguistic, normative and scientific definitions and the *iglos* sign model offer several advanced methodological solutions and approaches for identifying and avoiding these problems.

In this case, we tried to analyse the relation between the terms “position”, “location”, “localisation” and “positioning”. These differentiate in causal and functional relation.

In addition, translations from English into German and French are not very unproblematic because e.g. “position” is translated in German both as “Lage” and as “Standort”. At this point, we have a risk of confusion with “location” because of the translation problem in German or in other languages. Finally, we could find out that the four analysed terms exist in all three languages and that the relation between them is the same.

The main difficulty was the analysis and the determination of relations between these four terms in each three languages.

References

Ayşe Yurdakul, Eckehard Schnieder, Michal Hodon. 2013. Standardisation of international and interdisciplinary terminology in the language of transportation and automation engineering. EURO-ZEL 2013 – 21st International Symposium. University of Zilina. Zilina.

Eckehard Schnieder. 2012. Qualität dynamischer Satellitenortung im Eisenbahnverkehr. In: Technisches Messen (tm 4/2012), München, Oldenbourg: 210-219.

DIN EN 13848-4: Railway applications - Track - Track geometry quality - Part 4: Measuring systems - Manual and lightweight devices. Berlin, Beuth, 2011.

DIN EN ISO 19148: 2012-06: Geographic information - Linear referencing. Berlin, Beuth.

IEC 60351-1: 1976: Expression of the properties of cathode-ray oscilloscopes. Part 1: General. Berlin, Beuth.

ISO 19132: 2007. Geographic information - Location-based services - Reference model. Berlin, Beuth.

ISO 19134: 2007. Geographic information - Location-based services - Multimodal routing and navigation. Berlin, Beuth.

Online Etymology Dictionary. http://www.etymonline.com, (citated on September 6th, 2013).

Oxford Online Dictionaries. oxforddictionaries.com (citated on September 6th, 2013).

Susanne Arndt, Dieter Schnäpp, Eckehard Schnieder & Ayşe Yurdakul 2013. *iglos*. The intelligent Glossary – Terminology Management as Knowledge Network. (poster presentation). Braunschweig University of Technology, Institute for Traffic Safety and Automation Engineering, Braunschweig. EURO-ZEL 2013 – 21st International Symposium. University of Zilina. Zilina.

Attachement

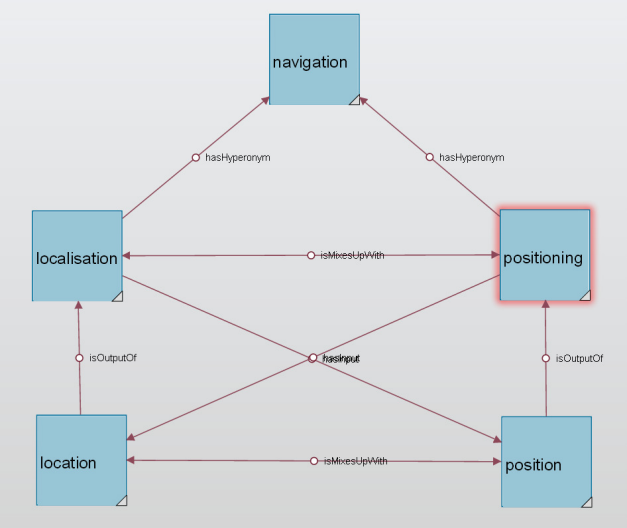


Figure 2: State-function relations of navigation terms in the *iglos* graph

Reusing existing conceptual structures and lexica for neology characterization in the field of Neurosciences: the NeuroNEO project

Nava Maroto
CES Felipe II
Universidad Complutense de Madrid
C/ Capitán, s/n, Aranjuez,
Spain, 28300
mmaro01@ucm.es

Abstract

The NeuroNEO Project aims to observe, compile and analyze newly-created terms in the realm of the neurosciences, with a special focus on the moment neologisms are born and to study these new lexical units in the very context where they appear. As part of this effort we intend to organize the concepts underlying the new terms into a structure that can be easily shared across applications. In order to do so we would like to reuse already existing resources in the field of neurosciences. In this preliminary study we describe existing ontologies and lexica and focus on the challenges and opportunities of reusing them, as well as the pitfalls of integrating tools originally created for the English language into a Spanish conceptual structure. We will also consider new models for interchanging lexical resources on the Semantic Web, such as the *lemon* model proposed by McCrae et al. (2012).

1 Introduction

In this article we present a preliminary study of the existing ontologies and lexica which could be reused in order to characterize neologisms in the realm of the neurosciences.

There is no doubt nowadays that in order to manage terminology in an efficient way it is essential to reuse existing resources and tools, rather than starting from scratch with every new project (McCrae et al 2012). Until quite recently, this effort was an almost impossible one, due to the lack of interoperability of resources, which are generally inaccessible for a number of different reasons such as the incompatibility of formats or the inaccessibility of data, among others.

However, the advent of the Semantic Web and the Linked Data paradigm opens up a new panorama in which data can (and should) be linked and shared across projects and applications. In this paper we focus on the possibilities of applying the principles of Linked Data to the characterization of new concepts within the NeuroNEO project and describe existing resources and technologies that might be suitable for our purposes.

The rest of the paper is structured as follows: First, we present the main objectives of the NeuroNEO project and the role played by conceptual organization (section 2). Then we describe the ontological and lexical resources in the neurosciences that could be reused for the purposes of the NeuroNEO project (section 3). In section 4 we focus on the principles of Linked Data and their possible application to terminology projects. In section 5 we review the *lemon* model proposed by McCrae et al. (2012) for interchanging lexical resources over the semantic web. Finally, we point out the main challenges and pitfalls that lay ahead in the NeuroNEO project.

2 The NeuroNEO project

The NeuroNEO Project (García Palacios et al, in press) aims to observe, compile and analyze newly-created Spanish terms in the constantly-evolving realm of the neurosciences, with a special focus on the very moment and context in which neologisms are born.

In order to achieve this end, the project aims to collect new lexical units in close collaboration with specialists in the field and with specialized translators as necessary collaborators and decisive agents in the dissemination of neologisms. For further details on the process of neology detection and handling within NeuroNEO refer to García Palacios et al. (in press).

As new terms are collected and described, we need to observe the place they occupy within the conceptual structure of the discipline. With a view to corroborating whether new terms are semantically motivated, the conceptual structure from the specialists' point of view needs to be considered, that is, we should observe how neuroscientists organize knowledge about their field.

In this way, if new terms are semantically motivated and can therefore be easily related to other already existing close concepts in the field, the chances for these new terms to be accepted by experts will be higher.

3 Ontologies and lexica in the neurosciences

Over the past years, several joint efforts have been made to collaborate in the exchange of ontologies in the biomedical sciences. One of them is the BioPortal of the National Center for Biomedical Ontology (NCBO[1]), which provides access to commonly used biomedical ontologies and to tools for working with them. Within the portal two ontologies dedicated to the neurosciences are available: the Computational Neuroscience Ontology[2] and the NIFSTD ontology, described in more detail later. Both can be downloaded in OWL and therefore reused with applications such as ontology editors.

Another interesting approach to sharing ontologies in the biomedical domain is that of the OBO Foundry[3], which aims at making a core of ontologies fully interoperable by virtue of a common design philosophy and implementation.

These two projects are just a sample of the interest shown by life scientists in sharing ontologies and conceptual structures across the web.

A third initiative worth mentioning is the EHTOP, European Health Terminology/Ontology Portal (Grosjean et al, 2011). The EHTOP is a repository that provides access to 32 health terminologies and ontologies, both for humans and computers. It focuses on French and English, although some data are available in other languages, too.

EHTOP is a very useful tool for accessing multiple resources of the medical sciences, including the neurosciences. However, as far as we know, it is not prepared to link data using RDF or the Linked Data principles and Spanish is not one of the supported languages.

3.1 Neuroscience Information Framework Standard Ontology (NIFSTD ontology)

One of the most outstanding efforts to develop domain ontologies in the field of the neurosciences is being carried out within the Neuroscience Information Framework (NIF[4]). NIF is a dynamic inventory of web-based neuroscience resources that gives access to data, materials, and tools via the Internet, bringing together the efforts of 16 centers for neuroscience research.

In order to overcome the need for a shared semantic framework for neuroscience, NIF has developed NeuroLex[5], a lexicon of common neuroscience terminology built from NIFSTD in a modular fashion, with separate modules covering major domains of neuroscience (e.g., anatomy, cell types, techniques). NIF Standard ontology (NIFSTD) is constructed according to the set of best practices established by the OBO Foundry project. In this way it was designed to avoid duplication of effort and ensure that work performed under one domain is of maximum utility to the broader community by conforming to standards that promote reuse (Bug et al., 2008). All data are open access and codified in RDF and the entire ontology can be downloaded in OWL. This feature makes it especially useful for our purposes, as we could download parts of the ontology and include our newly discovered concepts within the already existing NIFSTD structure.

The main drawback for our purposes is that both Neurolex and NIFSTD are codified in English and we want to describe Spanish neologisms. The need for "localization" of the ontology has to be assessed before using the NIFSTD ontology in a Spanish context (Espinoza et al, 2009).

4 Linked data and terminology

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web. It also refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is

[1] http://bioportal.bioontology.org/
[2] https://github.com/INCF/Computational-Neurosciences-Ontology--C.N.O.-/wiki
[3] http://www.obofoundry.org/
[4] http://www.neuinfo.org/. Not to be confused with the NLP Interchange Format (also abbreviated as NIF) used in Linguistics.
[5] http://neurolex.org/wiki

linked to other external data sets, and can in turn be linked to from external data sets (Bizer, Heath & Berners-Lee, 2008).

The result of linking data on the Web, referred to as Web of Data, is described as "*a web of things in the world, described by data on the Web*" (Bizer, Heath & Berners-Lee, 2008).

Berners-Lee (2006) set the following rules for publishing data on the Web as Linked Data, known as Linked Data principles:

1- Use URIs[6] as names for things
2- Use HTTP URIs so that people can look up those names
3- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4- Include links to other URIs, so that they can discover more things.

Linked Data relies on documents containing data in RDF (Resource Description Framework) format. RDF provides a generic, graph-based data model with which to structure and link data that describe things in the world. In RDF, data are represented as a Subject-Object-Predicate structure, where the objects and subjects are either resources or literals (van Erp, 2012). Several RDF extensions are being designed with the goal to formalize knowledge bases like terminology databases and lexical-semantic resources. As Chiarcos et al. (2012) point out, "if (linguistic) resources are published in accordance with these rules, it is possible to follow links between existing resources to find other, related data and exploit network effects".

However, the main challenges in achieving this are **interoperability** of language resources (both at the structure and the conceptual level) and **information integration**, that is, how to combine heterogeneous information from different sources in an efficient way (Chiarcos et al., 2012).

5 The *lemon* model for interchanging lexica over the Semantic Web

In our quest for a model that enables ontologies to be reused and to overcome the language barrier, we are considering the *lemon* model (McCrae et al., 2012a) for interchanging lexica over the Semantic Web.

Lemon (Lexicon Model for Ontologies) is an RDF model that allows lexica to be specified for ontologies and allows these lexica to be published on the Web (McCrae et al., 2012b). The main features of the model are:

- **Semantic by reference:** linguistic descriptions are separate from the ontology, but their semantics are defined by pointing to the corresponding semantic objects in the ontology.
- **Modular architecture:** the model consists of a core model and a set of complementary modules. Linguistic descriptions are grouped into linguistic properties, lexical and terminological variation, decompositions of phrase structures, syntactic frames and their mappings to the logical predicates in the ontology and morphological decomposition of lexical forms.
- **Openness:** *lemon* is a descriptive model that does not prescribe the usage of specific linguistic categories. The data categories or linguistic annotations used to define lexical information have to be specified by reusing URIs from other dictionaries and repositories such as ISOcat or the GOLD ontology.

Since *lemon* builds on the RDF data model, URIs are used to name and dereference linguistic annotations, and links can be easily created between lexicons using RDF triples.

Multilingualism is also foreseen in *lemon*, as several lexica in different languages can be associated to one and the same ontology.

That is why we consider that this model could be of use in our building of a Spanish lexicon upon NIFSTD ontology.

6 Conclusion

In this paper we have looked at different possibilities to reuse existing resources in the characterization of neologisms.

The Linked Data approach seems to offer a suitable scenario for reusing resources widely accepted by neuroscientists. NIFSTD ontology is readily available in OWL format, making it suitable for the exchange with applications such as ontology editors.

One of the challenges of the project is the reuse of resources originally created for the Eng-

[6] A URI (Uniform Resource Identifier) is a string of characters used to identify a name or a web resource. Such identification enables interaction with representations of the web resource over a network.

lish language for characterizing neologisms in Spanish. In the NeuroNEO project we are going to intervene when the neologism has just been created or has not been created yet. In a certain sense, the Spanish-speaking neuroscientist is working in a bilingual situation (English-Spanish). One consequence of this special kind of bilingualism is that there will probably be a single conceptual structure both for English and Spanish. This is why we believe that conceptual and linguistic resources created for the English language could be reused for the purposes of organizing Spanish neologisms in this field.

However, the issue of multilingualism needs to be addressed. We have seen that models such as *lemon* could be used to share ontology lexica across the web in different languages.

The NeuroNEO project still has to consider the trade-off between effort and quality of the result. The main issues that still lay ahead are the following:

- We need to evaluate tools and models that allow for the integration of new concepts within an ontology and at the same time allow for the development of a Spanish lexicon of the neurosciences that can easily be mapped to a commonly used ontology. In this regard, we need to consider how existing resources have been conceived, as this may affect their structure.
- We need to design a protocol for the integration of neologisms into a Spanish lexicon supported by NIFSTD ontology.
- We need to pay attention to data categories and try to use standardized models such as ISOcat (Windhouwer and Wright, 2012).

If we adhere to the Linked Data paradigm we will be able to benefit from previous efforts and contribute to the scientific community with our findings.

Acknowledgments

This work is supported by the Spanish Ministry of Economy and Competitiveness within the national project NeuroNEO "Regulación de los procesos neológicos y los neologismos en las áreas de neurociencias" (code FFI2012-34596).

References

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1-22.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data and Linguistics: Representing and Connecting Language Data and Language Metadata.* Berlin: Springer:57-64.

Joaquín García Palacios, Jesús Torres del Rey, Nava Maroto, Daniel Linder, Goedele De Sterck, and Miguel Sánchez-Ibáñez. In press. NeuroNEO, una investigación multidisciplinar sobre la neología terminológica.

John McCrae, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink et al. 2012a. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation.* Springer.

John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. 2012b. Integrating WordNet and Wiktionary with *lemon*. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data and Linguistics: Representing and Connecting Language Data and Language Metadata.* Berlin: Springer:25-34.

Julien Grosjean, Tayeb Merabti, Nicholas Griffon, Badisse Dahamna, and Stefan Darmoni. 2011. Multiterminology cross-lingual model to create the European Health Terminology/Ontology Portal. *9th International Conference on Terminology and Artificial Intelligence, TIA 2011, Paris, 8–10 November 2011:*119–122.

Marieke van Erp. 2012. Reusing Linguistic Resources: Tasks and Goals for a Linked Data Approach. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data and Linguistics: Representing and Connecting Language Data and Language Metadata.* Berlin: Springer: 57-64.

Mauricio Espinoza, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez. 2009. Ontology Localization. *K-CAP2009 - The Fifth International Conference on Knowledge Capture, California, September 1st-4th.*

Tim Berners-Lee. 2006. Linked Data – Design Issues. Available online at: http://www.w3.org/DesignIssues/LinkedData.html [Last accessed: June 28, 2013].

William J. Bug, Giorgio A. Ascoli *et al*. 2008. The *NIFSTD* and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics* 6:175-194.

Menzo Windhouwer and Sue Ellen Wright. 2012. Linking to Linguistic Data Categories in ISOCat. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data and Linguistics: Representing and Connecting Language Data and Language Metadata.* Berlin: Springer:99-108.

The Spanish Travel Subjective Lexicon (STSL)

Liliana Ibeth Barbosa Santillán
Information Technology
University of Guadalajara, México
ibarbosa@cucea.udg.mx

Inmaculada Alvarez de Mon y Rego
Mercedes Rodríguez Villareal
Lingüistica aplicada a la ciencia y la tecnología
Universidad Politécnica de Madrid
ialvarez@euitt.upm.es

Abstract

This paper presents a proposal for a recognition model for the appraisal value of sentences. It is based on splitting the text into independent sentences (full stops) and then analysing the appraisal elements contained in each sentence according to the previous value in the appraisal lexicon. In this lexicon, positive words are assigned a positive coefficient (+1) and negative words a negative coefficient (-1). We take into account word such as "too", "little" (when it is not "a bit"), "less", and "nothing" than can modify the polarity degree of lexical unit when appear in the nearby environment. If any of these elements are present, then the previous coefficient will be multiplied by (-1), that is, they will change their sign. Our results show a nearly theoretical effectiveness of 90%, despite not achieving the recognition (or misrecognition) of implicit elements. These elements represent approximately 4% of the total of sentences analysed for appraisal and include the errors in the recognition of coordinated sentences. On the one hand, we found that 3.6 % of the sentences could not be recognized because they use different connectors than those included in the model; on the other hand, we found that in 8.6% of the sentences despite using some of the described connectors could not be applied the rules we have developed. The percentage relative to the whole group of appraisal sentences in the corpus was approximately of 5%.

1 Introduction

The lack of studies relative to polarity analysis in the Spanish language is one of the motivations for this research because it is the fourth largest language spoken in the world with 6.9 million speakers. The automatic analysis of the appraisal in text corpora has significantly advanced in recent years. There are proposals of language codification that have been developed based on relatively reliable tools. Unfortunately, most developments are carried out only in English. Translation is not always a good solution and that is specially evident in the case of translating polarity. For example, it would be necessary to face a translation of every polarity word in English with the uncertainty that the cultural weight of each language does not suppose a modification of the polarity expression in the same way. Likewise, it would not be possible to translate some elements such as phrases, sayings, and popular expressions in a particular culture. (Gelbukh et al., 2002) said that analyzing a traditional corpus of texts has the disadvantage that few words occur many times thus using most of the processing time. In addition, the majority of interest words when appraisal or polarity are concerned have few or no representation in some of the existing corpora. This limits the study of word contexts. One solution to this problem is the use of the web which can be regarded as a large virtual corpus. It has sufficient information in order to study the properties of a large number of words. But this virtual corpus has some disadvantages compared to the traditional ones such as: the network response time, unstable results in time, etc. The corpus that has been examined in this research combines the advantages of both types, virtual and locally stored. Since it is created from the web where we find all kinds of material that can be stored in order to analyze the properties of a particular type of texts, in this case travel blogs. It is possible to gain access to stories narrated by many people simply by choosing the topic. For this research we chose to extract texts from travel blogs since they are an inexhaustible source of good material for the study of the language of appraisal. As by their nature they are personal experiences, these stories have a very strong subjectivity. But they also have the component that the authors want not only to convey their travel experiences, but also evaluate the places visited. This includes the things they have seen or the people they have met in order to advise future travellers about what to do or not do, what to see and not to see, where to eat or not to eat, where to sleep or not to sleep, etc. We see it as an ideal source to extract polarity because almost everything is based on opinion. These stories are very different compared to a travel guide where the narrative is much more objective, full of facts and with very few evaluations (González Rodríguez, 2011). The aim of this project is double; on the one hand, the creation of a corpus of travel blogs that can be by itself a useful tool for a linguistics corpus and on the other hand, a study of the value patterns in the sentences, using the corpus of text previously created for this purpose. In this way, we can propose a computational model for

to estimate the polarity of the sentence. The proposed model is based on the following phases: (1) To recognize the polarity elements (words or groups of words) in a sentence within the corpus, (2) To recognize certain grammatical structures also indexed in the lexicon, (3) To establish rules that switch the polarity elements or polarity structures, (4) To obtain a semantic sum of the polarity elements. The remainder of the paper is structured as follows: Section 2 describes the Spanish Travel Subjective Lexicon (STSL) approach. Section 4 presents details of our data sets, evaluation metrics and the result. Finally, Section 5 presents our conclusions and future research.

2 Background

We understand appraisal theory as the discursive construction of attitude and intersubjective posture (Pérez Nieto and Redondo Delgado, 1997). This approach is a term of wide scope, that includes all evaluative uses of the language through which speakers and writers not only bring particular value attitudes but also negotiate these positions with their actual and potential interlocutors (Kaplan, 2004).

Following the work of (Kaplan, 2004) and using her examples and terminology, we will discuss the way in which appraisal theory is divided into three semantic domains: attitude, compromise, and gradation.

2.1 Attitude

The attitude domain includes the meanings by which the texts or speakers attribute a value or an intersubjective appraisal to the participants and processes. These can be related to both emotional answers and with value systems that are culturally determined. All statements are classified as attitudinal if they convey an evaluation both positive and negative, or can be interpreted as an invitation to the reader to provide their own negative or positive evaluations

This category is divided into affect, judgment and appreciation subsystems.

Affect Affection is the evaluation of how the writer indicates his or her emotional disposition towards persons, things, situations or events. The emotions are concentrated in three major groups that deal with happiness or unhappiness; safety and insecurity; and satisfaction or dissatisfaction. The linguistic indicators of affection can be verbs of emotion that refer to mental processes (eg.: to love / hate); adverbs that indicate circumstances of mood (e.g.: happily and sadly); adjectives that express emotion (e.g. : happy/sad), and nominalizations, i.e. transformations of verbs and adjectives into nouns (e.g.: happiness / desperation).

Judgment Judgment can be understood as the institutionalization of emotions in the context of rules on how people should and should not behave. The social norms that act into these appraisal judgements take the form of regulations or social expectations.

Judgments of social esteem are subdivided into: (a) relative to normality, (b) the capacity or the determination demonstrated in the conduct; all are evaluated in order to know how normal is a person, how competent or how decisive and determined he or she is, and (c) judgments of social sanction related to the veracity and moral integrity.

Appreciation

Appreciation can be considered as the system where human assessment is expressed toward products, processes and entities that are valued positively or negatively. Artefacts, texts, abstract ideas, plans and policies, and objects are evaluated according to polarity. Individuals can also be evaluated through appreciation, but only when they are perceived as entities and not as humans.

3 The approach

The STSL approach has seven stages: (1) search for texts, (2) text selection, (3) building of textual documents, (4) tagging of travel blogs on the web, (5) classification of data, (6) indexation, and (7) analysis of appraisal patterns. Then a database is created in order to save or delete lexical elements. As a result, we have a corpus of text, appraisal sentences, and an appraisal lexicon that will be later used for the analysis of appraisal patterns.

Search for Texts Our text is a travel forum, with a section for blogs. There are many entries, although not all meet the characteristics of inclusion in our corpus, either due to incomplete or general ideas or any other type of feature. STSL allows access to blogs and the possibility to search by title, user or by continent. Next we found that the blog is structured in two parts: (a) the most recently published diaries, and (b) most popular daily entries depending on the number of visits. The method of search was random, since texts did not fulfill all the selection features of the blogs. It was necessary to perform a tracing process, examining many blogs and performing a preliminary inspection of all of them one by one. Then STSL selected and discarded irrelevant blogs.

Text Selection The quality of the corpus is measured by the degree of compliance of the documents that meet the purpose for which the corpus is compiled. Thus it was necessary to take special care in the selection of documents attempting to maintain homogeneity. Therefore, it was necessary to establish the following criteria that govern the selection and inclusion of documents. (a) Quantity: It was decided to include 24 blogs of different dimensions. The total number of words collected was 201.678. (b) Quality of text: Given that the selection was manual, special care was taken in that the texts were written in the correct language, without spelling mistakes, in clear writing. (c) Published in travel blogs: Due to the nature of the project, we only included published blogs, discarding blogs in restricted

personal pages. (d) Type of travel: the trip must have been carried out as a tour, following a route or path, i.e. through a single country or region as a whole (for example we recognize Scandinavia as a region although it includes three countries, or the westerm United States given its extensive area). (e) Text form: The texts must be written in the form of logs or diaries, discarding the texts of general impressions of a journey for its lack of detail. (f) Style: The texts must be comprehensive, describing the journey from beginning to end, discarding free or incomplete texts introduced in unfinished or abandoned blogs. (g) Additional information: Each sample must be marked with a series of additional data, which gives extra information and allows for identification. These marks are the: web page from which it has been extracted, country or area where the trip has been realized, language, date of the trip, date of creation of the blog, and name of the author or nickname.

Building of Textual Documents Once the 24 blogs were selected, they were copied and included in documents for accommodation in the database. We found additional difficulty with some personal blogs where each day of the trip was on different web pages and we needed to browse all the links by using an index.

Tagging of Travel Blogs on the Web Sentences with appraisal value we marked with colours, blue for positive and pink for negative.

Clasification of data Each sentence was extracted from the word or group of words that function as an appraisal element, classifying it into existing categories using the drop-down categories. Grammatical categories were defined in advance but are subject to changes.The classification included the concept of "gradation" where our approach allows to select a word and those associated with it that can modify its value either intensificating or weaken.

Indexation To accommodate the corpus and have the versatility and functionality required, it was decided to create a relational database containing: (a) The texts of each of the blogs. (b) All appraisal sentences classified as positive or negative. (c) All the appraisal elements of each sentence and its grammatical classification. (d) The established object, person or situation from the emitted evaluations.

Analysis of Appraisal Patterns First we studied the lexical items that indicate the ability to infer rules that permit automated recognition of the evaluation. This allowed us to recognize polarity, i.e. positive or negative elements. We can divide the lexical elements of the corpus in two large groups. On the one hand, words or groups of words that have a fixed structure and can be easily recognized. In this group are adjectives, verbs, adverbs, nouns and phrases. We call these elements "explicit appraisal elements". On the other hand are the elements, they are included in sentences with a clear polarity but that are subjective or complex to recognize, because sentence structure is variable or the nature of the assessment is not evident in any of its elements. We call these elements "implicit appraisal elements" are items that we cannot identify due to its semantic complexity . Irony is that property of speech by which speakers understand the opposite of what is said; this only is identified by using context, which makes it impossible to know the elements that indicate irony. We study the explicit appraisal elements based on their frequency of occurrence in sentences. We see that the greater weight of the appraisal is in adjectives, which is logical since the adjective is an element specifically designed to assess. Next are the nouns, verbs and adverbs.

4 Experiment and Results

Our approach evaluated 24 blogs consisting of 345 pages that contain a total of 201,678 words. We extracted from them 4,183 sentences and 6,295 lexical elements as shown in Fig. 1.

Place	*Words*	*Appraisal*	*Lexical*
Scotland	2488	33	43
Senegal	5296	109	129
Italy	9013	198	312
Poland	3105	59	76
Morroco	8414	165	256
Italy	3058	49	58
Japan	14747	290	427
Vietnam	10743	245	364
Cuba	9550	188	273
USA-Canada	15110	391	604
USA	5030	109	180
China	11251	194	290
Austria-Germany	12798	255	364
India	8827	201	332
Lapland	10976	234	351
Morroco	5802	147	274
Egypt-Lebanon	19512	336	506
Syria	5670	154	240
Japan	6452	92	142
Iceland	14152	301	443
Scotland	6169	125	167
Morroco	4179	70	104
Thailand	3597	128	213
Egypt	5745	110	138
TOTAL	201678	4183	6295

Figure 1: Number of appraisal and lexical elements by country

The explicit appraisal elements are shown in Table I, they constitute a total of 95% of all items that contain polarity, in contrast to 5% implicit appraisal elements.

explicit appraisal elements	implicit appraisal elements
Adjectives	Rhetoric figures
Nouns	Ironies
Verbs	Exclamations
Adverbs	Theoretical interrogation
Phrases	Quotes
Suffixes	Suspension points
Interjecciones	Change of record

Table 1: The explicit and implicit appraisal elements.

The adjectives represent 48% of the total explicit appraisal elements. Some examples of adjectives in superlative degree which appear in the corpus are adjectives that already appear in the list of appraisal adjectives and include the prefix "super". Considering gra-

dation when it accompanies explicit adjectives it does not change polarity since it only reinforces the positive or negative value of the specific adjective. The most common graders are: ”Más, mas o menos, mucho, mucho más, muy o bien, tan, bastante, algo”, etc. The nouns represent 48% of the total explicit appraisal elements. The verbs represent 14% of the total explicit appraisal elements. The adverbs represent 7% of the total explicit appraisal elements. The most used adverbs or greater number of repetitions are shown in Table II.

adverb	frequency
bine	175
mal	59
perfectamente	13
tranquilamente	24
others	206

Table 2: The most used adverbs

They represent 4% of the total and include the identify margin of error. In the same way, we found in the corpus, evaluative nouns that are created by adding a suffix to the corresponding adjectives like ”-idad” or ”-ez”. In this way, the appraisal noun ”majestuosidad” would be the evaluative adjective ”majestuoso” or of the appraisal noun ”rapidez” would be the adjective ”rápido” by adding the corresponding suffixes. It would be possible to extend the list of nouns, by utilizing the appraisal adjectives, with their corresponding endings of nouns, identifying all the possible endings. However, given the variety of sufixes and not all adjectives can be transformed into nouns. For this study, we have extended the list of nouns only for those adjectives that appear in the corpus sample. The appraisal of the entire corpus by country with its polarity is shown in Fig. 2.

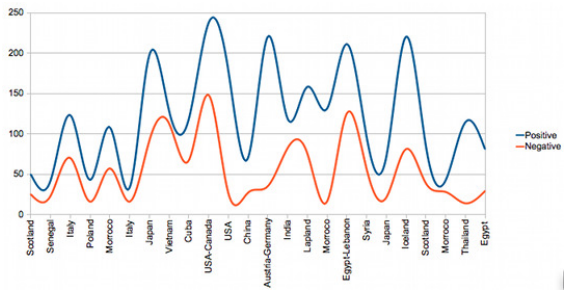


Figure 2: The appraisal by country with its polarity

In this study, the blog writers value the following the most:

$$IV :\doteq \begin{cases} amplitud, apariencia, aprovechamiento, \\ cantidad, comodidad, distancia, \\ eficacia, historia, importancia, \\ limpieza, mobiliario, olor, organización \\ precio, proximidad, puntualidad, sabor, \\ situacion, sonido, tamano, tolerancia. \\ valor\ añadido, velocidad, visitas \end{cases} \quad (1)$$

5 Conclusions

Appraisal theory can be considered as related to sentiment analysis. The application of this type of analysis has many possibilities, but all of them are based on sentence polarity recognition. Our model also accounts for the possibility of modifying the original value by means of a negative word. In this case, the possible negations must be located within the context of the sentence: ”no,” ”nor,” ”neither,” ”without,” ”none,” and ”never,” as well as any double negations that are dealt with as one: ”no/nor neither,” ”no/nor none,” ”no/nor never,” ”without nothing,” and ”without any.” If there are negations then the coefficient will change its sign by being multiplied by menus one (1). Then, the links with coordinated sentences should be sought: ”but,” ”despite” and ”although.” If there are more than one of these, then the following rules are applied: positive (P) ”but” negative (N) = negative; N ”but” P = P; P ”despite” N = P; N ”despite” P = N; P ”although” N = P; N ”although” P = N; ”although” P, N = N; ”although” N, P = P; ”although” P, N = N; ”although” N, P = P. It is possible to extend the number of appraisal words in our lexicon through the transformation of lexical categories into others by applying the relevant suffixes. We have observed that in most cases it is the root that has the appraisal meaning; thus, we can conclude that if ”glad” is an appraisal adjective then ”gladder” is an appraisal adjective too. Extension is also possible using techniques that were based on expanding lists of basic words to full lexical terms with the recursive consultation of synonymous words using electronic dictionaries.

Acknowledgments

We are grateful to the Sciences Research Council (CONACYT) and Multilingüismo en ontologías y linked data (BabelData), TIN2010-17550, funded by the Ministry of Science and Technology, 2011-2013.

References

Gelbukh, A., Sidorov, G., and Chanona-Hernández, L. (2002). Computational linguistics and intelligent text processing. 2276:285–288.

González Rodríguez, M. J. (2011). La expresión lingüística de la actitud en el género de opinión: el modelo de la valoración. *Revista de lingüística teórica y aplicada*, 49:109 – 141.

Kaplan, N. (2004). Nuevos desarrollos en el estudio de la evaluación en el lenguaje: La teoría de la valoración. *Revista Boletin de Lingüística*, 22:52–78.

Pérez Nieto, M. A. and Redondo Delgado, M. M. (1997). Procesos de valoración y emoción: Características, desarrollo, clasificación y estado actual. *Revista Electrónica de Motivación y Emoción (R.E.M.E.)*, IX(22).

Using parallel corpora to deal with unlexicalised concept for bilingual lexicon building: A case study of 'identity' in Chinese

Vincent X. Wang
Department of English, FAH
University of Macau
MACAO
vxwang@umac.mo

Abstract

One of the difficulties for the construction of bilingual lexicons comes from the fact that some concepts are lexicalised in one language but not so in the other language. This study focuses on a commonly-used word – 'identity' – which relates to a concept deeply rooted in the Western tradition but is nevertheless much newer to Asian cultures. Its meaning of unique, quintessential and defining character can hardly be reproduced by a single lexical term in the Chinese language. In this research, we use web-based English-Chinese parallel corpora as linguistic resources to retrieve candidates for the rendition of identity in Chinese. Our sample allows the main senses of 'identity' to emerge, and reveals a range of most frequently used terms in Chinese to render each of the senses. The findings are discussed in relation to concept formation, semantic frame and network.

1 Introduction

Bilingual lexicons are essential for translation practice, and fundamental for example-based machine translation (Piao, 2002). They are also valuable resources for comparative language studies. However, the instances of unlexicalised concepts in a target language present particular difficulties to the building of bilingual lexicons. One of the most appealing instances involving the Chinese language concerns the concept 'identity'. This concept is deeply trenched in western tradition and closely associated with other key notions such as 'person', 'individuality', 'human rights'. However, personal identity does not operate similarly in Eastern cultures. In a Chinese social context, a person's identity very often depends on his/her social network – e.g., who his/her family members are, who he/she knows well, which social group he/she belongs to. The concept of personal identity is not readily lexicalised in Chinese (cf. 2.2). There is therefore no clear counterpart of identity in the Chinese language. However, the present study makes use of parallel corpora to gather the instances of translation of 'identity' into Chinese, in an attempt to make suggestions for potential candidates that render the concept into Chinese.

2 Identity: its meanings and translation

In this section, we will survey the definitions of identity in dictionaries and in the research on WordNet, and review previous research on the problem to translate 'identity' into Chinese.

2.1 Entry in dictionaries and WordNet

Different dictionaries define 'identity' in different scopes and varied fine-grainedness. We surveyed several commonly-used dictionaries – *Collins Cobuild Advanced Dictionary of English (seventh Edition)*, *Macmillan English Dictionary for Advanced Learners*, *Oxford English Dictionary*, *English-Chinese Dictionary* 英漢大詞典 (by Lu Gusun), and *New Time English-English English-Chinese Dictionary* (by Yan Yuanshu). Each dictionary denotes two to five senses of the entry 'identity' (we excluded obsolete, technical, regional usage). The typical senses include (a) the distinctive characteristics of a person or a place, (b) who a person is, and (c) sameness between two entities. The online dictionary.com provides a visual thesaurus that displays four major senses of 'identity' in the form of four branches – personal identity, identicalness, identification, and (mathematical) identity operator (http://www.visualthesaurus.com). Such a visual representation closely resembles the four senses used in WordNet. We examine all the senses of

'identity' in our research, except the mathematic one, e.g., identity operator.

2.2 Difficulty to perceive the concept

Previous research has documented that Chinese learners of English tend to find the western concept of personal identity unfamiliar to their own practice and feel it is difficult to perceive language use associated with this concept (Richter and Song, 2005). The problem may stem from the Confucius tradition – it attaches importance to collectivism, universal virtues, common good of a community, but disfavours personal endeavours for self realisation (Yum, 1988). In the social context of China today, individuals' power and identity heavily derive from their social network, family ties, and social in-groupness. Richter and Song's (2005) study demonstrates that Chinese speakers encounter much difficult to translate the key concepts such as identity, self, myself into Chinese. Their informants provided a range of solutions, which are understandably often disagreeable and dissatisfactory.

3 This study

We examine the major senses of 'identity', excluding the mathematic use of the word. We collected instances of 'identity' and their translations into Chinese from parallel corpora.

3.1 Parallel corpora

There are four resources used for data in the present study:

1. The Babel Parallel Corpus (English into Chinese) returned eight matches of 'identity' from 244,696 words. http://124.193.83.252/cqp/babel1/
2. Chinese-English Online (CEO) 中英双语在线 – an online corpus 英汉对应语料库检索系统 (more ten million words) developed by the Beijing Foreign Studies University – gave nine hits. http://fleric.org.cn/ceo/index1.html
3. E-C Concord developed by The Hong Kong Institute of Education: the English novels part (807-thousand words) yielded eleven search results including one instance of reduplication. http://ec-concord.ied.edu.hk/paraconc/index.htm
4. Ju Hai 'sea of sentences', a large collection of bilingual sentences, functioning as a part of an online dictionary 海詞 dict.cn. Although there is no mentioning of the volume of words of Ju Hai, it returns 4850 solutions for the search of 'identity'. http://juhai.dict.cn/

We used all the search results (27 in total) of the first three resources and 73 results from the fourth one to build a sample of 100 pairs of English-Chinese sentences for this research. The instances of reduplications and conspicuous typos and mistranslations were laid out.

3.2 Data analysis method

A close examination of the 100 instances of 'identity' in our sample enabled its major senses to emerge and be classified in main groups. We will then investigate the possible equivalents in Chinese for each of the senses.

3.3 Results

There are four major senses that emerged in our sample: (A) self-identity, (B1) public identity, (B2) an entity's distinctive characteristics, and (C) sameness between two things. We can now examine each of the senses and the most frequently used items in Chinese to render the sense.

A. Self-identity. This sense points to a person's defining characteristic. It relates to his/her individuality, personality, and manhood. It does not have to be his/her public image, or titles and labels given by his/her social affiliations and networks. It is something deep inside, at the heart of his/her being. People tend to be aware of their own identity and consider it extremely important and dear to them. Our sample contains nine instances of identity using Sense A. The translations into Chinese do not display any readily equivalent item. This echoes the finding of previous research that it is difficulty to express this sense of 'identity' in Chinese. The translations in our sample tend to use paraphrase or specify most probable contextual meaning of the word (Baker, 2011). Here are some examples (underlines added):

(1) a . I am conscious of my own identity.
    b. 我意识到我自己的存在。
(2) a. It's a serious situation if one loses his identity.
    b. 失去个性可是件大事。
(3) a. You can lose your identity when you join the army.
    b. 你参军会失去自己的个性。

(4) a. They inevitably diminish the new individual's sense of esteem and identity because they may consider themselves to be the product of an assembly line.
b. 这样不可避免地会降低新个体的人格和尊严的意识，因为这些克隆人可能认为自己只是一条装配线上生产出来的产品。

The translators used their own interpretations in the context to explicate the meaning expressed by 'identity' – e.g., 存在 *cun zai* 'existence' or 'being' in (1b), 个性 *ge xing* 'an individual's character or disposition' in (2b) and (3b), and 人格 *ren ge* 'manhood' or 'personality' in (4b). This indicates the absence of a cover term in Chinese to render Sense A of 'identity'. Our sample also suggests other alternatives to render this sense – e.g., 个人存在 *geren cunzai* 'existence of an individual' – as well as more adaptive translations such as 失落感 *shiluo gan* 'sense of loss' to convey the feeling of losing of one's identity, and 隐姓埋名 *yinxing maiming* 'to hide one's family and given names' for the situation in which one purposely conceals of his identity.

B1. Public identity. It tends to serve an instrumental function – to single out an entity from a population by recognising its unique characteristics. Such characteristics are often physical, measurable, and operationalisable to a large scale of population. The typical collocations include identity card, proof of identity, digital identity, and identity parade. Unlike self-identity, such an identity can be designed, assumed or even stolen:

(5) a. A credit card is not a valid proof of identity.
b. 银行发的支票保付卡不是有效的身份证明文件。

(6) a. The journalist did not want to reveal the identity of his informant?
b. 那个新闻工作者不想透露消息提供人的身分。

(7) a. How do I report online fraud and identity theft?
b. 如何举报网络诈骗和盗窃身份行为?

(8) a. Managing digital identity is a critical issue for both consumers and businesses.
b. 管理数字身份鉴别不管对消费者还是商家都是极其重要的问题。

(9) a. This lets you manage identity ranges manually.
b. 这允许您手动管理标识范围。

(10)a. but for feeling certain that the man had no suspicion of my identity.
b. 幸亏这个犯人没有对我产生怀疑，没有认出我来。

Of the four senses of 'identity', this sense (B1) is most readily translatable into Chinese. There are 74 instances of Sense B1 in our sample. The most frequently used translations are 身份 *shen fen* (n= 32) and 身分 *shen fen* (n= 12) – both are pronounced identically and often used interchangeably, although the first term tends to be used for identity card and documents (5), while the second for social status of a person or a legal person (6). A similar translation – 身份鉴别 *shenfen jianbie* 'authenticating an identity' (8) – is used twice in our sample, which emphasises more on the act of authenticating or appraising. Another translation – 标识 *biao shi* 'marking' (9) – occurs seven times in our sample and expresses more technical meaning of marking out distinctive features. There are also several more adaptive translations in our sample, which express the contextual meaning of 'identity', e.g., 认出 *ren chu* 'recognising' (10), 是谁 *shi shui* 'is who', and 是他 'is him'.

B2. Distinctive characteristics. This sense, somehow related to Sense B1, reveals more distinctive characteristics of an entity. The eight instances of Sense B2 tend to be translated into 特性 *te xing* 'special features', 特征 *te zheng* 'special characteristics' (n= 3 for the two terms) and 形象 *xing xiang* '(special) image' (n= 2).

(11) a. The countries have kept their own distinct political and cultural identities.
b. 这些国家保持了自己独特的政治和文化特征。

(12) a. Language and identity in Caribbean literature..
b. 加勒比海文学的语言和特性。

(13) a. I want to design the corporate identity.
b. 我想设计企业形象。

The terms 特征 (11) and 特性 (12) appear to be close equivalents in Chinese for Sense B2.

C. Sameness. This sense of 'identity' points to sameness and identicalness between two entities through rather objective or factual comparison. It is not difficult to be translated into Chinese with several commonly used terms:

(14) a. Identity of interests is the bond that unites them.
b. 利害的一致是把他们联合起来的纽带。

(15) a. As to the cloning of dictators and celebrities, or the manufacture of a "super race", we all understand that genetic identity does not guarantee identical personality and behaviour.

b. 至于独裁者和名人的克隆或是一个"高级种族"的制造，我们都清楚，拥有相同的基因并不能保证拥有完全相同的个性和行为。

(16) a. Identity and continuity are not the same.

b. 等同与连续并非一样。

Several terms are able to convey Sense C into Chinese: 一致 *yi zhi* 'in agreement' (14), 相同 *xiang tong* 'sameness' (15), 等同 *deng tong* 'equality (16). Other alternative translations into Chinese include 统一 *tong yi* 'unity' and 视...为一体 *shi … wei yiti* 'taken as a unity'.

Table 1 summarises our findings in and shows the most frequently used Chinese translations:

sense	no	frequently used translation
A	9	个性, 人格, 个人存在, 存在
B1	74	身分, 身份, 鉴别, 鉴定, 标识, 认出
B2	8	特性, 特征, 形象
C	9	等同, 相同, 一致, 统一, 视...为一体

Table 1: Senses of 'identity' and translations.

4 Discussion

This research demonstrates the value of parallel corpus for bilingual lexicon building. From parallel language materials, the main senses of 'identity' emerged. Our results lend support to distinction between personal identity and public identity, a position advocated by John Locke (1632-1704) (see discussion by Richter and Song, 2005, p. 92). Our data showed that the former (Sense A in this study) points to a semantic vacuum in Chinese – an unlexicalised concept – while the latter can be readily translated into Chinese. From the perspective of frame semantics, when Chinese speakers have difficulty to establish a core (keyword) to a semantic network using Sense A of 'identity', they would have problem to perceive its associated words and collocations, e.g., self, myself, individuality (Fillmore, Johnson, & Petruck, 2003). However, based on our evidence from parallel corpora, although Sense A is not lexicalised in Chinese, parallel corpora provide a variety of solutions that translators have used to solve the problem (cf. Table 1). Our investigation also reveals some constrains of using use parallel corpora. We have observed a few cases of typos, mistranslations, rather questionable adaptions in translation.

5 Conclusion

The present study investigates the problem of unlexicalised concept in the target language in relation to the building of bilingual lexicon. Parallel corpora demonstrate advantages of using real-life language materials to allow the main senses of the word in question to emerge. Our results show the distribution pattern of the main senses of 'identity' and present the most frequently used terms in Chinese to render each sense. Such terms are candidates for English-Chinese lexicons. Our findings reinforce the previous research that self-identity (Sense A in this study) is unlexicalised in the Chinese language and therefore can hardly be captured by a single cover term. It requires paraphrase or contextualised renditions as commonly used translation methods. By contrast, the other three senses of 'identity' are more readily translatable into Chinese. A variety of lexical items serve as close equivalents (cf. Table 1). The potential of using parallel corpus deserves to be explored in future studies in relation to the problem of bilingual corpus construction, machine translation and translator training.

References

Baker, Mona. (2011). *In Other Words: A Coursebook on Translation* (2nd ed.). Abingdon Oxon. New York: Routledge.

Fillmore, Charles J., Johnson, Christopher R., & Petruck, Miriam R. L. (2003). Background to FrameNet. *International Journal of Lexicography, 16*(3), 235-250.

Piao, Scott Songlin. (2002). Word Alignment in English–Chinese Parallel Corpora. *Literary and Linguistic Computing, 17*(2), 207-230.

Richter, Eva, & Song, Bailin. (2005). Translating the concept of identity. In E. Hung (Ed.), *Translation and Cultural Change: Studies in History, Norms and Image-Projection* (pp. 91–110). Amsterdam: John Benjamins.

Yum, June Ock. (1988). The impact of Confucianism on interpersonal relationships and communication patterns in East Asia. *Communication Monographs, 55,* 374-388.

PLATO : un outil de facilitation des métiers du droit

Peraldi Sandrine
ISIT
s.peraldi@isit-paris.fr

Kotowicz Jean-Philippe
INSA Rouen - LITIS
jean-philippe.kotowicz@insa-rouen.fr

Abstract

L'objectif de PLATO (Platform for Legal Applications and Technology Online), récemment soumis à un financement européen, est de développer un outil multimédia multilingue, à destination des acteurs des métiers du droit. Construit sur le modèle des architectures orientées services (SOA), porteur d'une base de connaissances interactive et d'une interface adaptative intelligente, cet outil de facilitation vise à gérer et optimiser l'accès aux connaissances juridiques pour les praticiens du droit, interprètes et traducteurs juridiques dans leur pratique quotidienne, dans un environnement international et multilingue.

1 Introduction

Depuis quelques années déjà, les institutions juridiques nationales et européennes accordent une importance grandissante à la question de la justice en ligne. En effet, les technologies de l'information et de communication modernes apparaissent désormais comme essentielles afin de promouvoir l'efficacité de la justice et de faciliter l'accès au droit, en particulier dans un contexte multilingue.

Dans cette optique, une interface juridique européenne, le portail E-Justice, a été créée en 2010 afin de permettre aux citoyens, aux entreprises et aux praticiens du droit d'accéder à l'information juridique (aide juridictionnelle, formation judiciaire, registres fonciers, etc.) grâce à la mise à disposition de 22 000 pages consultables dans les 22 langues officielles de l'Union européenne (UE). Néanmoins, l'accès aux informations juridiques ne constitue au mieux qu'une première étape. En effet, avoir accès à l'information est une chose, pouvoir mettre différents systèmes juridiques européens en vis-à-vis et en comprendre les subtilités en est une autre.

Un nombre considérable d'études portant sur le droit européen et transnational comparé a depuis longtemps mis en évidence les innombrables difficultés linguistiques et techniques liées à la comparaison des systèmes juridiques (De Groot, Van Laer 2007 ; Bocquet 2008).

De même que les cultures, les traditions et la langue d'un pays ont évolué et divergé de celles de leurs voisins, les systèmes juridiques ont également développé leurs propres spécificités (par exemple, Common Law anglais versus système continental français). Ces différences interculturelles et donc juridiques sont inhérentes à la façon dont une communauté de locuteurs d'une langue donnée conçoit le monde, son environnement et crée des concepts pour appréhender, catégoriser et nommer cette réalité.

Mais aussi précieuses soient-elles, ces différences, associées à la complexité des textes de loi et à une terminologie et phraséologie marquées, sont aussi à l'origine des principales difficultés rencontrées par l'UE pour harmoniser ses démarches et procédures judiciaires ; des difficultés liées notamment à l'absence d'équivalences entre concepts de droit ou à la superposition entre concepts européens et nationaux. Nombre de juriliguistes (De Groot 1996 ; Grossfeld 2000 ; Magris & al. 1999) estiment que la complexité des concepts juridiques interdit toute possibilité de transposition linguistique. Qui plus est, l'exercice requiert de la part des traducteurs et des législateurs, une connaissance approfondie du droit comparé et des systèmes juridiques concernés, toute erreur traductologique ou terminologique pouvant mener à une application ou une interprétation de la loi erronée ou problématique au sein des pays membres (Sacco 1999).

2 Objectifs

C'est pourquoi, en complément du portail E-Justice, il devient indispensable de proposer un outil de facilitation des métiers du droit, sous la forme d'une base de connaissances multilingue et interactive, destinée aux praticiens du droit, traducteurs et interprètes juridiques qui sont les artisans du multilinguisme. Ces trois catégories d'utilisateurs ont en commun le besoin d'accéder à des informations contextualisées, associé à une compréhension conceptuelle et interculturelle fine des documents juridiques multilingues. La mise à disposition de simples lexiques d'équivalences, de traductions alignées ou d'informations juridiques tronquées et non hiérarchisées ne permet plus de gérer la complexité grandissante des textes de loi nationaux, transnationaux et européens.

PLATO a donc pour objectif de développer un système intelligent permettant un accès multi-niveaux à un vaste corpus multilingue composés d'articles, de textes de loi et de la jurisprudence récente. Plus spécifiquement, l'outil permettra :

• Une mise en parallèle intelligente de ces corpus par langue et sous-domaines, via une indexation et annotation sémantique de chaque document.

• La possibilité d'exécuter des requêtes, via un système d'interrogation multilingue et orienté utilisateurs, afin de proposer des données croisées sur des thématiques juridiques (par exemple, une mise en parallèle des jurisprudences de plusieurs pays européens sur un sujet ponctuel).

• Une terminologie et une phraséologie multilingue exhaustive afin de faciliter l'élaboration, la traduction et l'interprétation des textes et procédures juridiques. Outre la nécessité de comprendre parfaitement le message du législateur, il est essentiel de restituer un texte authentique, en respectant la terminologie, le style archétypique et les combinaisons lexicales propres au discours juridique (Heylen & al. 2010).

• Des analyses et commentaires d'ordre conceptuel et interculturel visant 1) une meilleure identification et gestion des difficultés notionnelles ; 2) une juste compréhension, interprétation et application de la loi dans les différents systèmes juridiques concernés. Les utilisateurs pourront s'appuyer sur un ensemble d'explicitations juridiques, d'analyses en droit comparé, voire de nouvelles propositions dénominatives.

• Des ontologies juridiques par (sous-)domaines, associées à des représentations graphiques des concepts afin de permettre aux utilisateurs de saisir l'articulation hiérarchique des concepts juridiques. Les ontologies serviront aussi à la construction d'un modèle de données fondamental qui interagira de manière permanente avec l'ensemble de la structure de l'outil et permettra la construction, l'échange, l'interrogation, la mise à disposition et l'interprétation des données.

Une fois identifiée dans le système, chaque catégorie d'utilisateurs sera en mesure de mener des recherches approfondies dans le corpus grâce à un assistant de recommandation électronique, l'e-advisor. Ce dernier, grâce à l'utilisation de technologies avancées en intelligence artificielle, aidera l'usager dans ses recherches, en prenant en compte son profil, ses objectifs et besoins, ainsi que l'historique des interactions précédentes avec la base de connaissances et les raisons motivant la recherche en cours. L'interaction entre l'e-advisor et la base de connaissance sera rendue possible par la mise en place d'une interface homme-machine spécialement conçue pour PLATO. Le rôle de cette interface est de s'adapter aux spécificités ainsi qu'aux besoins réels de chaque utilisateur et au contexte d'usage. L'interface aura à la fois pour rôle d'afficher de manière personnalisée les informations recherchées par l'usager, tout en collectant le plus de données possible s'agissant des actions menées par ce dernier afin d'alimenter les algorithmes d'intelligence artificielle.

Quel que soit le type de requête saisie, les utilisateurs auront la possibilité de creuser les éléments de réponse fournis par le système grâce à un ensemble de liens menant vers de nouveaux concepts, de nouvelles analyses interculturelles, des convergences notionnelles ou des applications juridiques. Ils pourront obtenir progressivement des données de plus en plus spécifiques (par exemple, en partant de textes généraux vers des législations ou jurisprudences locales) ; naviguer entre plusieurs niveaux d'informations, tout en ayant la possibilité de sélectionner, filtrer, rassembler ou rejeter des données spécifiques.

La base de connaissance PLATO est conçue pour évoluer dans un univers virtuel de connaissances complexes et multidimensionnelles, en gardant systématiquement une trace du raisonnement juridique de chaque utilisateur.

3 Spécifications techniques

Le corpus est un élément déterminant de ce projet dans la mesure où les textes constituent à la fois des vecteurs de connaissances spécialisées

et une source majeure de représentation conceptuelle.

Parallèlement à l'extraction terminologique semi-automatisée, une annotation sémantique automatique (Federated knOwledge eXtraction (FOX)), fondée sur un système de reconnaissance d'entités (nommées) et d'identification des relations sémantiques (via l'utilisation de patrons linguistiques), sera déployée afin de repérer au sein du corpus les informations d'ordre terminologique, contextuel, notionnel et interculturel. Cette activité d'indexation est essentielle afin d'être en mesure de fournir des réponses contextualisées et ajustées aux besoins de chaque catégorie d'utilisateurs.

En se fondant sur l'indexation et le traitement du corpus, l'e-advisor pourra exploiter plusieurs techniques d'intelligence artificielle afin de mettre en place un assistant de recherche qui aidera à la navigation au sein des documents. L'e-advisor ne remplit pas une simple fonction de recommandation ; il interagit directement avec l'utilisateur pour fournir des réponses personnalisées. Il s'agit donc de mettre en place un véritable dialogue entre l'assistant de recommandation et l'utilisateur afin de permettre au premier de « comprendre » véritablement la requête du second et ainsi d'ajuster ses suggestions. Cela signifie également que l'e-advisor ne peut se contenter de fonder ses recommandations sur un simple système de filtrage collaboratif ou de filtrage basé sur du contenu. Il doit en l'occurrence intégrer des technologies qui soient en mesure de prendre en compte les objectifs et les circonstances d'utilisation du traducteur/interprète/praticien à un moment spécifique. C'est pourquoi l'outil combinera des systèmes de recommandation traditionnels aux dernières technologies d'intelligence artificielle à l'instar des mécanismes d'argumentation et des modèles d'agent cognitif.

Le domaine de la recommandation a en effet suscité un intérêt grandissant ces dernières années. Les technologies de filtrage collaboratif ou basé sur du contenu (Su et al. 2009; Pazzani & Billsus 2007) sont désormais suffisamment mûres et largement employées dans des sphères telles que l'e-commerce. Néanmoins, les systèmes de recommandation fondés sur ces méthodologies tendent à recourir à des modèles d'utilisateur simplifiés, en plus d'être cumulatifs. Cette approche cumulative ne prend pas en compte la notion d' « action située », autrement dit le fait que l'utilisateur interagit avec le système dans un contexte particulier. La définition et la prise en compte de ce contexte ont un impact direct sur la nature des recommandations.

Aussi, nous proposons de mettre en place un agent virtuel qui aura une connaissance approfondie de l'utilisateur, de ses préférences, de sa situation professionnelle, des tâches qu'il doit mener (dans un sous-domaine particulier). Sur la base de ces informations, l'agent pourra sélectionner les sources les plus adéquates et la façon de les exploiter afin de fournir les meilleures recommandations possible. Ces données doivent s'appuyer sur des modèles cognitifs des comportements utilisateurs afin d'activer la « compréhension » de l'agent s'agissant des besoins des utilisateurs et du contexte d'usage.

Afin de fournir un accès simplifié, interactif et naturel aux données juridiques dans un contexte interculturel, nous dotons l'assistant de capacités conversationnelles (Loisel et al. 2012). Les interfaces de dialogue en langage naturel offrent de nombreux avantages (Androutsopoulos et al. 1995) extrêmement intéressants dans le cadre de ce projet : (i) les interfaces de dialogue sont plus intuitives ; (ii) les requêtes des utilisateurs sont transformées en représentation système ; (iii) l'utilisation de dialogues interactifs permet de corriger les incompréhensions.

Toutefois, concevoir un tel agent est d'une grande complexité. Le système doit non seulement être en mesure d'extraire les données pour les transformer en langage informatique, mais il doit également pouvoir comprendre les processus en jeu. Les hommes ont naturellement le don de communiquer et de raisonner. Notre hypothèse consiste donc à penser que la prise en compte du contexte et des stratégies discursives des utilisateurs permettra d'adapter la machine à l'homme et à ses besoins. En effet , le dialogue est un moyen de communication efficace nécessitant peu ou pas de formation (Allen et al. 2000).

S'agissant de l'interface utilisateur, nous sommes également dans un domaine en perpétuelle évolution : les systèmes sont développés sur internet, ce qui a pour effet de démultiplier le nombre et la variété des profils utilisateurs et des contextes d'usage. Á cela s'ajoute, le fait que les données juridiques sont aussi structurellement en flux constant. Cela implique par conséquent de mettre en place une toute nouvelle approche en matière de conception logicielle et graphique, et un processus facilitant l'apprentissage profes-

sionnel « à la volée » (Consiglio et Van der Veer, 2011). La multiplicité des contextes et des cultures d'utilisation incite donc au développement d'interfaces extrêmement flexibles et adaptables (Van der Veer, 2011), autorisant des contenus dynamiques et multidimensionnels.

4 Conclusion

PLATO est agencé autour de deux structures spécifiques : un premier axe orienté contenu visant l'élaboration et l'exploitation d'un corpus sémantiquement annoté à des fins d'analyse linguistique et textuelle et de constitution d'ontologies multilingues. Une seconde structure, orientée ingénierie et nouvelles technologies, visant le développement d'un e-advisor (accès aux données), d'une interface homme-machine (lien entre données, assistant de recommandation et utilisateur) et bien sûr de l'architecture globale qui soutiendra l'ensemble, en structurant les informations d'un point de vue informatique.
L'interdépendance de l'ensemble de ces activités et cycles de travail constitue un défi de taille, nécessitant de mettre en place une approche par granularité fine, associée à des étapes de contrôle récurrentes. Mais elle reflète également le caractère pluridisciplinaire extrêmement fort et innovant de PLATO, en particulier, et des ressources onto-terminologiques du futur, en général. C'est en effet ce travail collaboratif entre des spécialistes de la structuration des outils, des spécialistes de la structuration des contenus, associés à une équipe de juristes spécialisés en droit comparé et dont l'expertise est essentielle pour l'interprétation des données juridiques, qui permettra la conception et à terme l'élaboration d'un outil véritablement adapté aux besoins des utilisateurs. Et, c'est dans cette action de mise en commun des savoir-faire au cours même de la recherche que réside l'innovation.

References

Allen, J., Ferguson, G., Miller, B., Ringger, E., Sikorski-Zollo, T.: Dialogue systems : From theory to practice in Trains -96. Handbook of Natural Language Processing, 2000, 347–376

Androutsopoulos, I., Ritchie, G., and Thanisch, P., Natural language interfaces to databases–an introduction, Natural Language Engineering, 1(01), 1995, 29–81.

Bocquet, C, La traduction juridique : fondement et méthode, de Boeck, collection Traducto, 2008, Paris/Bruxelles.

Consiglio and Van der Veer. Designing an interactive learning environment for a worldwide distance adult learning community. In Anke Dittmar & Peter Forbrig (Eds) Designing Collaborative Activities - Proceedings of ECCE, 2011. ACM Digital Library, 225-228

De Groot, G. Law, Legal Language and the Legal System: Reflections on the Problems of Translating Legal Texts. In Volkmar-Gessner-Holland-Varga (eds.) European Legal Cultures. Aldershot, Darmouth, 1996, pp. 155-159.

De Groot, G., Van Laer, C. The dubious quality of legal dictionaries. In Translation and Meaning. University of Maastricht, 2007, pp. 173-187

Grossfeld, B. Comparative Law as a Comprehensive Approach: A European Tribute to Professor J.A. Hiller. Richmond Journal of Global Law and Business, vol. 1, 2000, pp.1-33.

Heylen, Kris, Hendrik Kockaert & Frieda Steurs. 2010. "The TermWise Knowledge Platform: an efficient translation and terminology management suite for legal translation in Belgium". TKE 2010 - Presenting terminology and knowledge engineering resources online: models and challenges: Fiontar, Dublin City University, 11-14 August.

Loisel A., Dubuisson Duplessis G., Chaignaud N., Kotowicz J-P, Pauchet A. : A Conversational Agent for Information Retrieval based on a Study of Human Dialogues. ICAART (1) 2012: 312-317

Magris, M., Musacchio, M.T. La terminografia orientata alla traduzione tra pragmatismo e armonizzazione. Terminologie et Traduction, 1999, p. 148-181.

Pazzani & Billsus. 2007. "Content-based recommendation systems". Michael J. Pazzani and Daniel Billsus. Lecture Notes in Computer Science (LNCS). Volume 4321, pages 325-341.

Sacco, R. Langue et Droit. In Sacco-Castellani (eds.) : Les multiples langues du droit européen uniforme. L'Harmattan Italia, Turin, 1999, p. 163-179.

Su X. and Taghi M. K. A survey of collaborative filtering techniques. Adv. in Artif. Intell. 2009, Article 4, 19 pages. DOI=10.1155/2009/421425.

Van der Veer. 2011. Culture Centered Design. In Patrizia Marti, Alessandro Soro, Luciano Gamberini and Sebastiano Bagnara (Eds) Facing Complexity - Proceedings CHItaly. ACM Digital Library. 7-8

Van der Veer and Vyas. 2011. Non-formal Techniques for Requirements Elicitation, Modeling, and Early Assessment for Services. In : Anke Dittmar & Peter Forbrig (Eds) Designing Collaborative Activities - Proceedings of ECCE 2011. ACM Digital Library, 285-286.

A Proposal for the Representation of the Relations between Concepts, Terms and Lexical Data used in Knowledge Organization Systems

Thierry Declerck
DFKI GmbH
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken, Germany
declerck@dfki.de

Abstract

This short paper describes a proposal for the representation of the relations that can exist between concepts, terms and other natural language expressions used in knowledge systems, such as thesauri, taxonomies or ontologies. Taking as an example a multilingual thesaurus in the field of social sciences, we propose to adapt its underlying design and suggest an explicit representation of the relations between concepts/classes and properties of the domain knowledge represented in the thesaurus, their associated terminology and the lexical data used for expressing the terms. As a preliminary step, we advocate for modularizing the thesaurus in three knowledge organisation systems: a conceptual, a terminological and a linguistic one, each of those using distinct but related representation schemes.

1 Introduction

In the context of projects dealing with Knowledge-Driven Information Extraction (KDIE), we investigated in details how terms and natural language expressions are used in knowledge organization systems. A goal of this study was to propose a strategy for using those combined information types (concepts, terms and lexical items) in the automated analysis of semi-structured and unstructured texts for extracting domain relevant facts (Declerck & Buitelaar, 2012).

This work led naturally to consider issues related to the representation of those combined information types, aiming at a formal representation of the used terms and other natural language expressions in order to improve interoperability of this type of language data in the context of multilingual KDIE systems. As basis for our work, we have been looking at approaches by Reymonet et al. (2009), McCrae et al. (2012) and Declerck & Lendvai (2010). As a field study for our on-going work, we dealt with a large multilingual thesaurus in the field of social science[1].

In this short paper, we present briefly the thesaurus, discuss our proposal for the modification of its organization and outline future work.

2 The Thesaurus for the Social Sciences (TheSoz)

The thesaurus for the social sciences is a knowledge source under continuous development (we are currently using version 0.92). The list of keywords used in TheSoz contains about 12,000 entries, of which more than 8,000 are descriptors (accepted keywords, which act as domain terms).

It is encoded in RDF and SKOS. While the main conceptual elements of the thesaurus are encoded in the core syntax of SKOS, the resource makes also use of the SKOS-XL properties [2] for including labels containing natural language expressions that are attached to the

[1] See http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/, or (Sapliko etal., 2012)

[2] See http://www.w3.org/TR/skos-reference/skos-xl.html

conceptual elements, using the "prefLabel" and "altLabel" annotation properties, allowing thus to describe main terms and their variants. The advantage of using SKOS-XL labels, compared to rdfs:label, consists in the fact that the value of such labels is not a literal, but an object, which can thus be related to other knowledge objects. The natural language expressions corresponding to the labels are represented by using the SKOS-XL annotation property "literalForm".

In order to give a (human readable) idea of the content of the thesaurus[3], we extracted with a Perl script the main elements from the RDF and SKOS code and present those in a tabular fashion, an example of which is given below, displaying also the terms in the languages covered by TheSoz (English, French and German):

concept id "10034303"

term "10034303"

- prefLabel id "10034303"
- lang=de "Abbrecher"
- lang=en "drop-out"
- lang=fr "drop-out"
- altLabel id "10034307"
- lang=de "Studienabbrecher"
- lang=en "university drop-out"
- lang=fr "étudiant qui abandonne ses études"

notation „3.2.00"

- lang=de „Schule und Beruf (berufliche Qualifikationselemente im Bereich der schulischen Ausbildung)"
- lang=en "School and Occupation (Elements of Occupational Qualification in School Education)"
- lang=fr « École et profession (éléments de qualification professionnelle dans le domaine de l'enseignement scolaire) »

broader notation „3.2"

- lang=de „Beruf und Qualifikation"
- lang=en „Occupation and Qualification"
- lang=fr « profession et qualification »

broader notation „3"

- lang=de „Interdiszipinäre Anwendungsbereiche der Sozialwissenschaften"
- lang=en "Interdisciplinary Application Areas of Social Sciences"
- lang=fr « domaines interdisciplinaires d'application des sciences sociales »

In the example above the reader can see how the English preferred label "drop-out" is associated with the concept "School and Occupation", which is itself a subclass of the concept "Occupation and Qualification", classified itself as a field of the broader concept "Interdisciplinary Application Areas of Social Sciences". All the language material contained in the labels or used for naming the "notations" can be re-used in the context of multilingual KDIE applied to text.

As the reader can see, in this thesaurus all conceptual, terminological and language data are present in the same Knowledge Organization System (KOS). Descriptors are introduced using the RDF language, specifying them as being an owl#Class, being itself an rdfs subClassOf a skos Concept.

```
<rdf:Description
rdf:about="http://lod.gesis.org/thesoz/ext/Descriptor"
>
    <rdfs:label
xml:lang="en">Descriptor</rdfs:label>
    <skos:definition xml:lang="en">Descriptors of
the TheSoz</skos:definition>
    <rdf:type
rdf:resource="http://www.w3.org/2002/07/owl#Class
"/>
    <rdfs:isDefinedBy
rdf:resource="http://lod.gesis.org/thesoz/ext/thesoz_e
xt.rdf"/>
    <rdfs:subClassOf
rdf:resource="http://www.w3.org/2004/02/skos/core#
Concept"/>
  </rdf:Description>
```

A concrete concept in TheSoz looks like:

```
<rdf:Description rdf:about="concept/10034303">
    <skos:inScheme
rdf:resource="http://lod.gesis.org/thesoz/"/>
    <prv:containedBy
rdf:resource="http://lod.gesis.org/thesoz/"/>
    <prv:createdBy
rdf:resource="http://lod.gesis.org/thesoz/Creation"/>
  </rdf:Description>
```

The concepts in TheSoz are encoded by numbers. We like this approach, consisting in not adding any natural language expression to the identification of knowledge objects outside of the annotation properties like "label", "comment" or "definition", although this is practiced in a number of cases, and experiments have been done on considering this type of language data used in RDF identifiers for cross-lingual applications (Fu et al., 2012). Those language independent concept IDs can be associated to a plurality of labels containing natural language expres-

[3] Online visualizations and access are available at http://lod.gesis.org/thesoz/

sions, being either multilingual variations of a preferred label, or different usages of a term in one language.

The interesting point for us is the fact that TheSoz introduces explicitly the knowledge objet "term" at the same level as "concept":

```
<rdf:Description rdf:about="concept/10034303">
    <skosxl:prefLabel
        rdf:resource="term/10034303"/>
```

Here a term is being encoded as a prefLabel (could also be an altLabel) of a concept. A "term" is also represented by a number, which is the value of the property "prefLabel". We find this solution more flexible than the one proposed by (Reymonet et al., 2009) where concepts and terms are both encoded as OWL descriptors, and where meta-classes are needed for describing the relations between domain concepts and terms, leading to an OWL full system. But important is the fact that both approaches are describing terms as full knowledge objects. An added-value of having terms as knowledge objects is the fact that relations between terms (and not only between concepts) can be formally described.

We would additionally advocate for separating the domain modeling (using RDF and OWL, also the rdfs:label property for supporting human consumption of the model) from the terminological description (modeled with RDF, SKOS and SKOS-XL) and cross-linking both by the use of a mixture of RDF and SKOS properties.

The approach by (Reymonet et al., 2009) includes basic linguistic information in the classes introducing the terms. We would also like to depart from this and have linguistic information encoded as independent but related knowledge objects. And this is something, which is also not realized in TheSoz. There the content associated to the term objects are just displayed as strings within the annotation property "literalForm":

```
<rdf:Description rdf:about="term/10034303"
    <skosxl:literalForm
xml:lang="de">Abbrecher</skosxl:literalForm>
</rdf:Description>
```

This line specifies that the knowledge object "term/10034303", which is the prefLabel of "concept/10034303", is linguistically realized with the word "Abbrecher" (*drop-out*). Another "term" knowledge object is associated to the concept as its altLabel:

```
<rdf:Description rdf:about="concept/10034303">
    <skosxl:altLabel rdf:resource="term/10034307"/>
</rdf:Description>
```

with literalForm:

```
<rdf:Description rdf:about="term/10034307">
    <skosxl:literalForm
xml:lang="de">Studienabbrecher</skosxl:literalForm
>
</rdf:Description>
```

"Studienabbrecher" (*university drop-out*) is thus one of the term variant for the object concept/10034303.

What now about the strings displayed as the values of the property "literalForm"? As mentioned above, Reymonet et al. (2009) include basic linguistic information in the ontological elements used for representing the terms. Here we opt rather for a system that encodes linguistic information as an independent knowledge object, as this is described by the *lemon* model (McCrae et al., 2012). In this model, the semantics of the language data included in labels of ontologies is described by an explicit reference made to a class or a property of the domain ontology under consideration.

Differently to (McCrae et al., 2012), we link the linguistic description encoded in *lemon* not to an ontology element but to a term in a corresponding knowledge organization system (realized as SKOS conceptScheme). The following (simplified) example shows how our *lemon* representation of the linguistic information associated to the decomposed literalForm "university drop-out" refers to the term "10034307", as its semantic "anchor":

```
:university_drop-out [lemon:writtenRep "univer-
sity drop-out"@en]
lemon:sense [lemon:reference
        :thesoz:term/10034307];
lemon:decomposition ( :university_comp
:drop-out_comp ) ;
lemon:phraseRoot [lemon:constituent :NP ;
lemon:edge [lemon:constituent :NP ;
lemon:edge [lemon:constituent :NN ;
lemon:leaf university_comp ] ;
lemon:edge [lemon:constituent :NN ;
lemon:leaf drop-out_comp ] ];
]
```

In this, lexical data refer only indirectly to classes or properties of ontologies, mediated by term objects in a terminological concept scheme.

A good example of how a (multilingual) lexicon encoded as a knowledge object could look like is given by the recent RDF version of Wiktionary in the context of DBpedia[4], which is still in a very early development phase. In this lexicon, *lemon* references and other relations to lexicon external knowledge objects are fully supported.

3 Actual Work

We are currently working in reducing redundancy of information in TheSoz and in the *lemon* model. We note that in both case tokens in a "literalForm" or a *lemon* entry can be repeated in the concept scheme or in the ontology lexicon when they are referred by different concepts (*lemon*) or terms (TheSoz). We are adding for this a specific representation for every word used in labels, and replace the components of the labels by pointers to the uniquely represented word (or lemma combined with morphological information), which could be in fact a RDF representation of a Wiktionary entry, situating thus the lexicon in the Linked Open Data framework. In doing so, we have a much more compact representation of the lexicon that is used in knowledge sources. This work consists in fact of a SKOS and *lemon* formalization of some ideas formerly described in (Declerck & Lendvai, 2010).

We expect from our work to generate a harmonized lexical basis for a wide range of Knowledge-Driven natural language processing applications.

Acknowledgments

The work presented in this paper has been supported by the TrendMiner project, co-funded by the European Commission with Grant No. 287863.
The author is thanking the reviewers for their very helpful comments, which led to substantial changes brought to the final version of the paper.

References

Buitelaar, P., Cimiano, P. Haase, P., Sintek, M. 2009. Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference*. Springer Berlin/Heidelberg, pp. 111—125.

Declerck, T,. Buitelaar, P. 2012. Ontologies as a Source for the Automatic Generation of Grammars for Information Extraction Systems. In: *Proceedings of the SWAIE 2012 Workshop: Semantic Web and Information Extractio.n*

Declerck, T., Gromann, D. 2012. Towards the Generation of Semantically Enriched Multilingual Components of Ontology Labels. In: *Proceedings of the 3rd Multilingual Semantic Web Workshop*.

Declerck, T., Lendvai, P. 2010. Towards a standardized linguistic annotation of the textual content of labels in Knowledge Representation Systems. In: *Proceedings of the seventh international conference on Language Resources and Evaluation*, Valetta, Malta, ELRA.

Ell, B., Vrandecic, D., Simperl, E. 2011. Labels in the Web of Data. In Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A. (eds.): *Proceedings of the 10th international conference on the semantic web* - Volume Part I (ISWC'11), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, pp.162_176.

Fu, B., Brennan, R., O'Sullivan, D. 2012. A Configurable Translation-Based Cross-Lingual Ontology Mapping System to Adjust Mapping Outcomes. Journal of Web Semantics, Vol. 15, pp.15_36.

Gromann, D., Declerck, T. 2013. Cross-Lingual Correcting and Completive Patterns for Multilingual Ontology Labels. In Buitelaar, P. and Cimiano, P. (eds) *Multilingual Semantic Web*, Springer-Verlag (to appear)

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T. 2012. *Interchanging lexical resources on the SemanticWeb.* Journal of Language Resources and Evaluation, pp.1_19.

Reymonet, A., Thomas, J., Aussenac-Gilles, N. 2009. Ontology based information retrieval: an application to automotive diagnosis. In *Proceedings of International Workshop on Principles of Diagnosis.*

Silberztein, Max. 2003. *NooJ manual*. Available at the WEB site http://www.nooj4nlp.net (200 pages)

Wimalasuriya, D. C., Dou, D. 2012. *Ontology-based information extraction: an introduction and a survey of current approaches.* Journal of Information Science, Vol. 36, No. 3, pp.306-323.

Zapilko, B., Johann Schaible, Philipp Mayr, Brigitte Mathiak. 2012. *TheSoz. A SKOS Representation of the Thesaurus for the Social Sciences*. Semantic-Web Journal.

[4] See http://dbpedia.org/Wiktionary

Retour d'expérience sur la création d'une ressource termino-ontologique (RTO) juridique

Sylvie Szulman
Université Paris 13
Sorbonne Paris Cité
LIPN, CNRS, (UMR 7030),
Villetaneuse, France.
sylvie.szulman@lipn.univ-paris13.fr

Haifa Zargayouna
Université Paris 13
Sorbonne Paris Cité
LIPN, CNRS, (UMR 7030),
Villetaneuse, France.
haifa.zargayouna@lipn.univ-paris13.fr

E. Paul
Groupe Victoires-Editions
38, rue Croix-des-Petits-Champs
75038 PARIS cedex 01
e.paul@victoires-editions.fr

Abstract

Dans ce papier, nous relatons l'expérience pluridisciplinaire d'une construction collaborative d'une ressource termino-ontologique (RTO). Cette RTO est construite pour annoter des documents juridiques et administratifs et ainsi répondre à des questions relatives à l'activité juridique d'une commune. Nous mettons en lumière les enseignements que l'on peut tirer concernant le travail de terminologie juridique, la pluridisciplinarité et l'impact de l'application sur le contenu de la ressource.

This paper presents a multidisciplinary experience on collaborative building of termino-ontology resource (TOR). This resource is built in order to annotate law and administrative documents and hence allow answering questions about local legal activity. The paper also presents lessons learnt concerning the work on legal terminologies, multidisciplinary and the impact of the application on the content of the resource.

1 Introduction

Les ressources sémantiques sont au cœur de la recherche d'information (Vallet et al., 2005). Elles peuvent prendre plusieurs formes qui vont de la liste de termes à une ontologie formelle. Dans le cadre du projet LégiLocal qui vise à proposer des fonctionnalités d'accès à l'information juridique locale pour les citoyens, les élus et personnels de mairie, la ressource sémantique est une ressource termino-ontologique (RTO) constituée de terminologies reliées à une ontologie modulaire. Les terminologies doivent servir à annoter des documents pour pouvoir répondre à des questions relatives à l'activité juridique d'une commune. Dans le cadre du projet de recherche, cette activité juridique a été restreinte aux problèmes communaux induits par l'activité de randonnée pédestre. Ainsi un secrétaire de mairie peut rechercher des documents lui permettant de répondre à la question suivante : « un quad peut-il circuler sur le GR ® [1] traversant la commune ? ».

Dans ce papier nous relatons l'expérience pluridisciplinaire d'une construction collaborative de la RTO et des enseignements que l'on peut en tirer concernant le travail de terminologie juridique, la pluridisciplinarité et l'impact de l'application sur la ressource créée. La section 2 présente un bref état de l'art sur la construction de RTO, puis la section 3 explicite notre travail.

La section 4 met en lumière les aspects liés à la terminologie juridique ainsi que le problème de lien très fort entre terminologie et application et présente l'expérience menée.

2 Etat de l'art

L'état de l'art est riche autour des questions de construction d'une RTO. Dans (Charlet et al., 2012), les auteurs présentent une expérience de construction d'une RTO médicale (ONTOLURGENCES) pour la recherche d'information. Ils s'intéressent notamment aux questions de réutilisation et d'intégration de ressources existantes. Pour le domaine du droit des collectivités territoriales, et pour le droit en général, très peu de ressources existent en dehors de celles de haut niveau telle que LKIF-Core (Hoekstra et al., 2007) qui contient 3 modules consacrés au droit qui constituent une ontologie générique du droit. Dans (Aussenac-Gilles et al., 2002), les auteurs

1. sentier de Grande Randonnée.

montrent que la nature de l'application visée conditionne les différentes étapes de construction de produits terminologiques.

Nous nous sommes intéressés spécifiquement au processus de validation des termes extraits automatiquement. Cet objectif nous différencie des travaux qui s'intéressent à la validation de termes pour évaluer la qualité des outils d'acquisition (Zargayouna and Nazarenko, 2010) qui supposent l'existence d'une référence.

Les travaux qui mettent en place un cadre collaboratif sont de plus en plus nombreux (Siorpaes et al., 2008). Dans un premier temps, nous ne nous posons pas la question de spécifications de fonctionnalités d'un outil comme dans (He et al., 2009) mais plutôt des recommandations méthodologiques qui permettent d'expliciter et de tracer les choix effectués.

3 Ressource et méthode suivie

La ressource terminologique à construire doit décrire un domaine vaste qui a été découpé en plusieurs sous-domaines. Pour identifier ces sous-domaines donnant lieu à des modules, nous sommes parties de la tâche qui est d'aider les secrétaires de mairie (ou DGS [2]) à rédiger des actes conformes à la légalité et de rendre visible au citoyen une grande partie des actes publiés par les collectivités. Les modules identifiés sont détaillés dans (Ressad-Bouidghaghen et al., 2013).

Pour la construction de ces modules, des corpus ont été créés pour délimiter le périmètre. Certains modules sont construits en ré-utilisant des ressources existantes telles que l'ontologie de l'insee [3] ou l'annuaire de la DILA [4].

La construction de la RTO par modules permet une construction incrémentale par ajout de modules. Dans un premier temps nous avons sélectionné un sous-domaine relatif à la randonnée pédestre qui doit contenir les principales notions propres à la pratique de la randonnée du point de vue des collectivités territoriales. Nous détaillons dans la suite la constitution du volet terminologique du module Randonnée.

Le corpus est constitué d'un livre (Louarn, 2010). Ce corpus contient moins de 50 000 mots. Le contenu du livre est un guide présentant la responsabilité des organisateurs de randonnées, le droit des propriétaires des terrains traversés, des élus et des associations dans la pratique de la randonnée pédestre. Le discours est moins juridique que ne le sont les actes juridiques qui y sont expliqués. Ce choix de corpus a été guidé par l'adéquation de son discours à la langue juridique utilisée par les secrétaires de mairie. La taille et la nature du corpus ne permettent pas de détecter des régularités ou de faire des calculs statistiques.

Le corpus a été traité par l'outil d'extraction terminologique de notre partenaire (société Temis [5]). Un plugin eclipse a été créé et ajouté à la plateforme TERMINAE (Szulman, 2011) pour visualiser les résultats de l'extraction et permettre la constitution de la terminologie.

Le travail de validation a été découpé en trois phases : (i) une phase d'explicitation des choix de validation sur la base d'un ensemble restreint de candidats termes, (ii) une phase de validation par les terminologues et (iii) une phase de double validation avec l'expert juridique.

La première phase de validation a impliqué l'expert juridique et deux terminologues. Elle a porté sur 120 termes candidats choisis en fonction de leur distribution dans le corpus. Cet échantillon a permis de constater qu'un filtrage automatique ne peut pas être effectué sur cette base. En effet, 44% (resp. 35%) de termes jugés pertinents (resp. non pertinents) par l'extracteur ont été validés. Le but de cette première phase est d'arriver à un accord entre les personnes impliquées dans la tâche de validation. Les 120 termes ont été jugés par chaque intervenant (terminologue, expert juridique). Un guide de validation a été mis en place et enrichi au fur et à mesure des discussions et des choix. Ce guide a permis d'expliciter les choix et de garder trace des décisions via des commentaires normalisés : (i) pour exprimer le doute sur certains termes qui nécessite des discussions, (ii) argumenter les choix, (iii) proposer de nouveaux termes issus du terme à valider.

La liste des termes candidats a été ainsi divisée en plusieurs listes :

– la liste des termes validés : A chaque terme valide correspond une fiche terminologique décrite ci-après.

2. DGS : Directeur Général des Services.

3. http://rdf.insee.fr/geo/

4. DILA : Direction de l'Information Légale et Administrative http://www.dila.premier-ministre.gouv.fr/

5. http://www.societe.com/societe/temis-sa-432265585.html

– la liste des termes invalidés : soit des termes non pertinents pour le domaine considéré (exemples *découverte de paysage rural*), soit des termes comportant des ENs qui ne sont pas directement utilisables dans la terminologie (exemples : *CAA de Bordeaux*). Cependant certains candidats termes invalides ont permis de proposer de nouveaux termes. Comme exemple on peut citer le candidat terme *dépense d'entretien de chemin rural*, ce candidat terme a été jugé invalide mais le sous-groupe nominal *entretien de chemin rural* a été jugé pertinent.
– la liste noire des candidats termes : des termes qui doivent être systématiquement supprimés de toute extraction terminologique concernant le domaine juridique (*sens des dispositions*). Cette liste est notamment utile pour les nouvelles extractions dans le domaine juridique.

La fiche terminologique présente plusieurs rubriques décrivant le terme, comme l'ensemble des variantes lexicales. Pour chaque variante lexicale, un qualificatif décrit par la rubrique USE doit être défini. Ce qualificatif a trois entrées (AUTORISÉ (ALLOWED), INTERDIT (FORBIDDEN), RECOMMANDÉ (RECOMMANDED)). La langue juridique utilise beaucoup d'acronymes qui doivent être présents dans la fiche. Ainsi le terme *recours pour excès de pouvoir* est très souvent utilisé sous son acronyme *REP*. La forme interdite est présente dans la langue juridique. En effet certains termes comme *conclusions* n'ont de sens en droit que s'ils sont utilisés au pluriel. Or tout extracteur de terme produit des candidats termes lemmatisés au singulier. Il faut indiquer que le terme conclusion au singulier ne doit pas être utilisé, cette information est d'autant plus utile pour l'annotation où on a généralement recours à la lemmatisation.

Chaque fiche est désignée par un terme et contient entre autres les différentes variantes lexicales. Les fiches proviennent soit des termes candidats validés, soit de termes candidats jugés invalides mais dont une partie est considérée valide comme *ordre public* dans le candidat terme *cœur de l'ordre public*. Certains termes notés "termes ?" sont des termes étudiés pour lesquels la catégorie n'a pas été établie et qui ont fait l'objet de discussions.

termes	nombre
invalides	764
valides	886
nouveaux termes	112
termes ?	157

La deuxième phase de validation a permis de créer 998 fiches terminologiques par deux terminologues n'ayant pas de connaissance particulière en droit. Des points réguliers ont été effectués avec l'expert juriste. Ils ont servi à éclaircir des notions et à mettre à jour le guide de validation. La deuxième étape a duré 60 heures approximativement. La première phase avait duré 10 heures approximativement. De découpage a donc permis un gain de temps considérable en permettant de paralléliser le processus de validation. Pour conclure sur l'expérience menée, l'interaction entre expert métier et terminologues est impérative pour construire une RTO répondant aux objectifs qui lui sont assignés. L'expert ignore, souvent, ses implicites et donc le choix des terminologues lui permet d'en prendre conscience et de les gérer. De même, son modèle de raisonnement, non conscient souvent, oriente ses choix et sa validation.

4 Bilan

Nous avons mis en pratique des notions théoriques connues mais essentielles lors d'une approche pragmatique. Ainsi des choix doivent être effectués et sont guidées par l'application. La ressource que nous construisons est utilisée pour un accès sémantique à l'information. Deux types d'accès ont été définis : une recherche d'information par mots clés et une navigation via la structure taxonomique de l'ontologie. Ces deux types d'accès nécessitent des choix différents : (i) l'un très lié à l'annotation des documents et à l'exploitation de ces annotations par le moteur de recherche d'information, l'autre très lié à l'ontologie et donc aux choix de modélisation. Deux exemples pour illustrer cet impact :

– terme simple versus terme composé : Les termes simples comme *randonnée* sont utiles pour la structuration mais s'ils sont utilisés pour l'annotation, augmenteront le bruit. les termes composés comme *droit de la randonnée*, *randonnée pédestre* permettent une recherche plus précise.
– Les entités nommées : les entités nommées (Omrane et al., 2011) peuvent servir à

déterminer les termes du domaine et aider à créer une ontologie. Ainsi l'EN *R 160-24 du CU* dans la phrase *Le maire a donc le devoir (R 160-24 du CU) d'y signaler les dangers ...* décrit un devoir du maire et permet de modéliser les obligations du maire. C'est l'EN extraite par l'extracteur de terme qui donne accès à l'expression dans laquelle elle est présente qui décrit une propriété du domaine à modéliser. Malgré leur utilité, ces ENs ne doivent pas être présentes dans la terminologie, car elles sont trop nombreuses et peuvent changer. Elles serviront à la population de l'ontologie et seront considérées comme des instances.

Les deux types d'accès possibles nous ont contraintes à chaque fois d'expliciter le choix fait. La liste des termes produite contient à la fois les termes qui serviront à l'annotation et ceux qui sont utiles à la modélisation.

5 Conclusion

Cet article décrit un retour d'expérience de construction de la partie terminologique d'une RTO par plusieurs intervenants de profils divers. Nous avons expérimenté un processus de construction collaboratif et incrémental. L'objectif de la collaboration est double puisqu'il s'agit à la fois de s'assurer d'une représentation consensuelle et de pouvoir paralléliser certaines tâches. Nous avons mis en place un guide qui explicite les choix terminologiques effectués.

Sur le plan, scientifique, la mise en lumière de la finalité du produit terminologique pour les choix de modélisation montre la justesse du paradigme prôné par le groupe TIA.

L'objectif de la constitution de la RTO est de permettre un accès sémantique à l'information juridique. Pour atteindre cet objectif il faudrait garantir une bonne couverture de la ressource ce qui pose la question récurrente de passage à l'échelle. La mise en place d'une construction modulaire et incrémentale garantit, à notre sens, de produire des ressources de qualité et volumineuses. La phase de validation par des non experts permet de faciliter le travail de l'expert et de permettre de valider assez rapidement des listes volumineuses de candidats termes.

Remerciements

Ce travail à été partiellement financé par le projet LégiLocal(FUI 9).

References

Aussenac-Gilles, N., Condamines, A., and Szulman, S. (2002). Prise en compte de l'application dans la constitution de produits terminologiques. In Maitre, J. L., editor, *Information, Interaction, Intelligence : Actes des 2e Assises Nationales du GDR I3*, pages 289–303. Cépaduès Editions.

Charlet, J., Declerck, G., Dhombres, F., Gayet, P., Miroux, P., Vandenbussche, P.-Y., et al. (2012). Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. In *Proceedings of IC 2012*, pages 33–48.

He, S., Nachimuthu, S., Shakib, S., and Lau, L. (2009). Collaborative authoring of biomedical terminologies using a semantic Wiki. In *Proceedings of AMIA Symposium*, pages 234–238.

Hoekstra, R., Breuker, J., Bello, M. D., and Boer, E. (2007). The lkif core ontology of basic legal concepts. In Casanovas, P., Biasiotti, M. A., Francesconi, E., and Sagri, M. T., editors, *Proceedings of LOAIT 2007*, pages 43–63.

Louarn, P. L. (2010). *Le droit de la randonnée pédestre*. Victoires Editions.

Omrane, N., Nazarenko, A., and Szulman, S. (2011). Le poids des entités nommées dans le filtrage des termes d'un domaine. In *9th TIA*, pages 80–86, Paris, France.

Ressad-Bouidghaghen, O., Szulman, S., Zargayouna, H., and Paul, E. (2013). Construction collaborative d'une Ressource Termino-Ontologique (RTO) pour le droit des collectivités territoriales. In *PFIA2013*, IC 2013, Lille, France. FUI 09.

Siorpaes, K., Hepp, M., Klotz, A., and Hackl, M. (2008). myontology : Tapping the wisdom of crowds for building ontologies. Technical report, STI-Semantic Technology Institute Innsbruck, University of Innsbruck, Austriafus.

Szulman, S. (2011). Une nouvelle version de l'outil Terminae de construction de ressources termino-ontologiques. In *22èmes Journées francophones d'Ingénierie des Connaissances. (IC 2011)*, page démonstration, Chambéry, France.

Vallet, D., Fernández, M., and Castells, P. (2005). An ontology-based information retrieval model. In *ESWC*, pages 455–470. Springer.

Zargayouna, H. and Nazarenko, A. (2010). Evaluation of Textual Knowledge Acquisition Tools : a Challenging Task. In *Proceedings of LREC 2010*, pages 435–440, Valletta, Malte.

Benefits of Natural Language Techniques in Ontology Evaluation: the OOPS! Case

Mari Carmen Suárez-Figueroa
Ontology Engineering Group.
Universidad Politécnica de Madrid
mcsuarez@fi.upm.es

Mouna Kamel
Institut de Recherche en Informatique de Toulouse (IRIT) - CNRS
Mouna.Kamel@irit.fr

María Poveda-Villalón
Ontology Engineering Group.
Universidad Politécnica de Madrid
mpoveda@fi.upm.es

Abstract

Natural language techniques play an important role in Ontology Engineering. Developing ontologies in a manual fashion is a complex and time consuming process, which implies the participation of domain experts and ontology engineers to build and evaluate them. Natural language techniques traditionally help to (semi)-automatically build ontologies and to populate them. However, the general trends for evaluating ontologies are mainly expert reviewing, evaluating quality dimensions and criteria, and evaluating against existing ontologies and set of common errors. That is, the use of natural language techniques in ontology evaluation is not widely spread. Thus, in this paper we aim at the use of natural language techniques during the ontology evaluation process. In particular, we propose a first attempt towards a language-based enhancement of the pitfall detection process within the ontology evaluation tool OOPS!.

1 Introduction

Developing ontologies manually is a complex and time consuming process, which involves both ontology engineers and domain experts. Natural language (NL) techniques have been traditionally used for extracting knowledge from texts to build semantic resources. In fact knowledge acquisition from text plays an important role in Ontology Engineering. It is divided into several steps, according to the "ontology learning layer-cake" (Cimiano, 2006): (a) identifying and extracting terms, (b) eliciting concepts and relations linking concepts from these terms, (c) organizing concepts and relations into hierarchies, and (d) identifying axioms.

During the ontology building, a wide range of difficulties and handicaps can appear. These situations may have as consequence the inclusion of anomalies in the ontology. Thus, the ontology evaluation process plays a key role in ontology engineering developments. Currently, the general trends in ontology evaluation involve different approaches (e.g., the comparison of the ontology to a "gold standard" or the detection of common errors in the ontology). However, what seems to be less present in the ontology evaluation field is the intensive use of NL techniques. For example, some structural or naming errors in the ontology may be automatically pointed out with a linguistic analysis of concept labels. Thus, our intention in this paper is to aim at the use of NL techniques during the ontology evaluation process. In particular, we propose a first attempt of improving the pitfall detection methods implemented within OOPS! by means of NL techniques.

The remainder of this paper is structured as follows: Section 2 summarizes different NL techniques used in Ontology Engineering. Section 3 presents the relation between ontology evaluation and NL-based techniques. Section 4 briefly describes OOPS!. In Section 5 our proposal towards a language-based enhancement of the pitfall detection process within OOPS! is presented. Finally, Section 6 outlines some conclusions and future steps.

2 Natural Language Techniques in Ontology Engineering

NL techniques traditionally help on the (semi)-automatic building of ontologies and on the population of ontologies with instances.

Most of the approaches for building ontologies from text, known as ontology learning methods, usually implement lexico-syntactic patterns (Hearts, 1992; Montiel-Ponsoda and Aguado de Cea, 2010), clustering methods or machine learning algorithms (essentially unsupervised) (Poelmans et al., 2010), to exploit various linguistic clues. Some platforms exist and implement one or a combination of these methods using different NLP tools (term or relation extractors, parsers, etc.). Examples are Text2Onto (Cimiano and Völker, 2005), which discovers concepts and hyperonimic relations between concepts, thanks to lexico-syntactic patterns and associative rules automatically learned from examples and OntoLearn (Velardi et al., 2005), which uses Wordnet (Fellbaum, 1998) for identifying lexical relations.

Regarding the population of ontologies, tools like TEXCOMON (Zouaq and Nkambou, 2008) uses linguistic patterns for instance identification, using named entity recognition techniques.

Linguistic approaches have been also applied to ontology matching where Euzenat and Shvaiko (2007) distinguish between language-based methods and methods which are based on linguistic resources, whereas the more general class of terminological approaches also includes string-based methods. We can mention the work by Ritze et. al (2010) that shows how complex matching can benefit from NL techniques.

3 Ontology Evaluation and Natural Language Techniques

Ontology evaluation process, which checks the technical quality of an ontology against a frame of reference (Suárez-Figueroa, 2010), plays a key role in ontology engineering projects.

To help developers during the ontology evaluation process, there are different approaches (Sabou and Fernandez, 2012; Poveda-Villalón et al., 2012): (a) comparison of the ontology to a "gold standard", (b) detection of common errors from catalogues in the ontology, (c) use of dimensions and criteria for describing the quality and goodness of the ontology, (d) use of the ontology in an application and evaluation of the results, (e) comparison of the ontology with a source of data about the domain to be covered, and (f) evaluation by experts who check the ontology against the requirements.

In addition, ontology evaluation can be supported by NL techniques in several ways (Gangemi et. al, 2005):

- When the ontology directly supports information retrieval or text mining applications and thus concerns objects mentioned in texts.
- When a corpus of documents is available, NLP can be used to identify mentions of instances (i.e. occurrences in text) of classes and relations which are mentioned in the text. A corpus-based evaluation of the ontology can reveal important properties of the ontology that might not be discovered otherwise.
- When (semi)-automatic population of the ontology is performed, NLP can help in the identification of new senses of already known instances, for example because the instance is polysemous and/or ambiguous (e.g., "Washington" is a person and a location).

However, ontology evaluation approaches could take more advantage of NL techniques. In this sense, we propose here a first attempt towards a NL-based upgrade of OOPS!.

4 OOPS!: OntOlogy Pitfall Scanner!

OOPS![1] (Poveda-Villalón et al., 2012) is a web-based tool, independent of any ontology development environment, for detecting potential pitfalls that could lead to modelling errors. Currently, OOPS! provides mechanisms to automatically detect as many pitfalls as possible, thus it helps developers in the diagnosis activity, which is part of the ontology validation process.

OOPS! takes as input an ontology to be evaluated and a pitfall catalogue in order to produce a list of evaluation results. The current version of the catalogue[2] consists on 35 pitfalls. Some examples are creating synonyms as classes, defining wrong inverse relationships, missing annotations, missing domain or range in properties, or defining wrong equivalent classes. Up to now, OOPS! detects semi-automatically a subset of 21 pitfalls related to the following dimensions: human understanding, logical consistency, modelling issues, ontology language specification and real world representation.

[1] http://oeg-upm.net/oops/
[2] http://www.oeg-upm.net/oops/catalogue.jsp

5 Towards a Language-based Enhancement of OOPS!

In this section we propose a first attempt towards a language-based enhancement of the pitfall detection process within the ontology evaluation tool OOPS!. To do this, we have reviewed the current catalogue of pitfalls in order to determine (a) which pitfalls, already implemented, could be detected in a better way by means of applying linguistic techniques and (b) which ones, not detected yet by OOPS!, could be implemented based on linguistic aspects.

Regarding the proposals for enhancing pitfalls already detected by OOPS!, we can mention the following ones:

• *P2. Creating synonyms as classes*: several classes whose identifiers are synonyms are created and defined as equivalent. Its detection could be improved by using linguistic resources such as WordNet and EuroWordNet, particularly by looking for the synonymy information of the class name.

• *P3. Creating the relationship "is" instead of using "rdfs:subClassOf", "rdf:type" or "owl:sameAs"*: the "is" relationship is created in the ontology instead of using OWL primitives for representing the subclass relationship ("subclassOf"), the membership to a class ("instanceOf"), or the equality between instances ("sameAs"). The detection could be enriched by creating specific language-dependent lexico-syntactic patterns to discover the use of 'is' and by using named entity recognition tools for characterizing the "instanceOf" relation.

• *P5. Defining wrong inverse relationships*: two relationships are defined as inverse relations when they are not necessarily. As first attempt, the implementation of this pitfall could be improved by creating specific lexico-syntactic patterns for direct/inverse relationship name structure.

• *P7. Merging different concepts in the same class*: a class is created whose identifier is referring to two or more different concepts (e.g., "StyleAndPeriod", or "ProductOrService"). As first attempt, its detection could be enhanced by creating specific language-dependent lexico-syntactic patterns and regular expressions to discover the use of 'and' or 'or' in the concept name.

• *P12. Missing equivalent properties*: when an ontology is imported into another, developers normally miss the definition of equivalent properties in those cases of duplicated relations and attributes (e.g., "hasMember" and "has-Member" in two different ontologies). The detection could be enriched by (a) using linguistic resources such as WordNet and EuroWordNet, specifically by looking for the synonymy information of the property name and (b) creating specific language-dependent lexico-syntactic patterns.

• *P13. Missing inverse relationships*: this pitfall appears when a relationship (except for the symmetric ones) has not an inverse relationship defined within the ontology. As first attempt, its implementation could be improved by creating specific lexico-syntactic patterns for direct/inverse relationship name structure (e.g., isSoldIn-sells; hasAuthor-isAuthorOf; hasParent-isParentOf).

• *P21. Using a miscellaneous class*: to create in a hierarchy a class that contains the instances that do not belong to the sibling classes instead of classifying such instances as instances of the class in the upper level of the hierarchy. This class is normally named "Other" or "Miscellaneous". As first attempt, its detection could be improved by creating a set of lexico-syntactic patterns that represent different ways of naming concepts that are usually miscellaneous entities.

With respect to those pitfalls not detected yet by OOPS!, we can propose the following ideas for their implementation based on NL aspects:

• *P1. Creating polysemous elements*: an ontology element whose name has different meanings is included in the ontology to represent more than one conceptual idea. As first approach, its detection could be implemented by (a) using linguistic resources such as WordNet and EuroWordNet, specifically by analyzing the different synsets in which the element name appears and (b) by analysing labels of neighbourhood concepts for disambiguation.

• *P9. Missing basic information*: information that is required and/or useful is not included in the ontology. As first approach and in certain situations, this pitfall could be implemented by using linguistic resources such as WordNet and EuroWordNet, specifically by analyzing the antonym information of the relationships name.

• *P30. Missing equivalent classes*: when an ontology is imported into another, classes with

the same conceptual meaning that are duplicated in both ontologies should be defined as equivalent classes. As first step, this pitfall could be detected by using linguistic resources such as WordNet and EuroWordnet, specifically by looking for the synonymy information of the class name.

• *P31. Defining wrong equivalent classes*: two classes are defined as equivalent when they are not necessarily. As first step, this pitfall could be implemented by using linguistic resources such as WordNet and EuroWordNet, specifically by looking for the hyperonym information of the class name.

6 Conclusions and Future Work

In this paper, we have presented the first efforts towards a NL-based enhancement of the pitfall detection process within the ontology evaluation tool OOPS!. We have reviewed the 35 pitfalls in the OOPS! catalogue and analyzed which pitfall detections could be linguistically improved and which pitfalls could be implemented based on NL as first attempt. In summary, we have proposed the improvement of 7 pitfall detection processes and the automation of 4 pitfalls not detected yet by OOPS!. Thus, we have planned to enhance OOPS! with the NL techniques presented in this paper.

7 ACKNOWLEDGMENTS

This work has been supported by the Spanish project BabelData (TIN2010-17550).

References

Philipp Cimiano. 2006. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer, November.

Philipp Cimiano and Johanna Völker. 2005. Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. in A. Montoyo, R. Munoz, E. Metais (eds), Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), vol. 3513 of LNCS, Springer, Heidelberg, p. 227-238.

Jérôme Euzenat and Pavel Shvaiko. 2007. Ontology Matching. Springer.

Christiane Fellbaum (ed.). 1998. Wordnet. An Electronic Lexical Database. The MIT Press, Cambridge Massacuhsets.

Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, and Jens Lehmann. 2005. Ontology evaluation and validation: An integrated formal model for the quality diagnostic task. Technical report, Laboratory of Applied Ontologies. CNR, Rome, Italy.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In Procs. of the 14th International Conference on Computational Linguistics COLING1992, p. 539–545.

Elena Montiel-Ponsoda and Guadalupe Aguado de Cea. 2010. Using natural language patterns for the development of ontologies. In V. Bhatia, P. Sánchez-Hernandez & P. Pérez Paredes, Eds., Researching specialized languages, p. 211–230. John Benjamins Pub.

Jonas Poelmans, Paul Elzinga, Stijn Viaene, and Guido Dedene. 2010. Formal concept analysis in knowledge discovery: a survey. In Proceedings of the 18th international conference on Conceptual structures: from information to intelligence, ICCS'10, p. 139–153, Berlin, Heidelberg: Springer-Verlag.

María Poveda-Villalón, Mari Carmen Suárez-Figueroa, and Asunción Gómez-Pérez. 2012. Validating ontologies with OOPS!, in Knowledge Engineering and Knowledge Management, Vol. 7603, Lecture Notes in Computer Science (Springer-Verlag, Berlin, 2012), pp. 267–281.

Dominique Ritze, Johanna Völker, Christian Meilicke, and Ondřej Šváb-Zamazal. 2010. Linguistic analysis for complex ontology matching. Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China.

Marta Sabou and Miriam Fernandez. 2012. Ontology (Network) Evaluation. Ontology Engineering in a Networked World. Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (Editors). Pp. 193-212, Springer. ISBN 978-3-642-24793-4.

Mari Carmen Suárez-Figueroa. 2010. PhD Thesis: NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse. Spain. Universidad Politécnica de Madrid.

Paola Velardi, Paolo Fabriani and Michele Lissikof. 2005. Using text processing techniques to automatically enrich a domain ontology. In Proceedings of ACM- FOIS., Maine, publication of ACM, 270 – 284.

Amal Zouaq and Roger Nkambou. 2008. Building Domain Ontologies from Text for Educational Purpos-es. IEEE Transactions on Learning Technologies 1(1), pp. 49 –62.

Session : Acquiring Semantic Relations in Linguistic Resources

Hybrid acquisition of semantic relations based on context normalization in distributional analysis

Amandine Périnet[1,2]
[1] Lingua et Machina
c/o INRIA Rocquencourt
BP 105, 78153 Le Chesnay Cedex
ap@lingua-et-machina.com

Thierry Hamon[2]
[2] LIM&BIO (EA3969)
Université Paris 13, Sorbonne Paris Cité
74, rue Marcel Cachin, 93017 Bobigny
thierry.hamon@univ-paris13.fr

Abstract

Semantic relations between terms are important and usefull information for many applications that exploit specialized texts. In this paper we address the limits of semantic relation acquisition methods on such texts. Among these methods, distributional analysis is statistical and usually used with big amounts of data. But with low frequency words, improvements are still needed. To overcome this limit, we propose a hybrid method combining several approaches. We especially focus on the integration of three methods that acquire hyperonymy relations and morpho-syntactic variants in distributional contexts. We experiment the hybrid method on a corpus of nutrition, and evaluate the relations in terms of precision. The best hybrid model to acquire semantic relations appears to be the generalization of contexts with hyperonymy relations, for both nouns and terms as targets.

1 Introduction

Whatever the domain, specialized texts are characterized by terms and relations between terms. Identifying these relations is crucial in many applications in Natural Language Processing (NLP), such as information retrieval, question-answering systems, information extraction in search engines or specialized automatic translation. For instance, a semantic relation that links the terms *sucre (sugar)* and *saccharose* will allow to increase recall in a retrieval information system.

Those relations may be provided by terminologies, but usually those resources are not tuned to the targeted texts (Bourigault and Slodzian, 1999). Relations may also be automatically acquired from specialized corpora, through different strategies. We can take into account morphological (Grabar and Zweigenbaum, 2000), syntactic (Jacquemin, 1997) or semantic information (Jacquemin, 1999), define lexico-syntactic patterns through observation in corpora (Hearst, 1992; Morin, 1999; Auger and Barriere, 2008), use machine learning techniques (Snow et al., 2005) or distributional analysis (Habert and Zweigenbaum, 2002), etc. All the methods show various limits. Regarding the quality of the results, they can either get a low recall (methods are too restrictive) or a low precision (ambiguities or polysemy are not well identified). Furthemore, approaches usually aim at acquiring only one relation type (for eg., hyperonymy).

Let's take the example of two methods :

- Lexico-syntactic patterns allow to get a good precision, but are limited by their (very) low recall (Embarek and Ferret, 2008), because of the quite restricted contexts they use.

- On the contrary, distributional analysis (DA) is more flexible and allows to put many terms in relation, with a great diversity of relation types (Morlane-Hondère and Fabre, 2012), but without returning a type of relation. Indeed, DA do not seem to offer any obvious way to distinguish between syntagmatic (collocations, noun-verb relations) and paradigmatic relations (synonymy, hyperonymy) (Fabre and Bourigault, 2006).

Furthemore, methods based on DA are generally used with big amounts of data and tend to be less efficient with low frequency words (Caraballo,

1999). Results obtained with general language are promising, but improvement is still required with specialized texts, even if good results have already been achieved (Habert and Zweigenbaum, 2002).

As mentionned above, one of DA's limit is low frequency words. Indeed, for those words, similarity is computed from very little information (i.e. the one in contexts), that leads to generate poorer quality groupings of terms (Caraballo, 1999). We assume that this information could be increased with semantic information as the one contained in an existing resource or acquired by a relation acquisition method, as for example, using hyperonymy relations acquired with patterns. Following this idea, we intend to define a hybrid method that switches words in DA contexts for their hierarchical parent or morphosyntactic variant. This method normalizes contexts (Henneron et al., 2005), to increase their frequency.

We first present the related work, then our hybrid method and we finally describe the different experiments we led. The results we get are then evaluated in terms of precision.

2 Related work

This work uses a DA method, based on the harrissian hypothesis that states that words appearing in a similar context tend to be semantically close (Harris, 1954). The DA principle has been automated in the 90's, and concepts and procedures used in distributional computations have been well defined (Sahlgren, 2006; Turney and Pantel, 2010; Baroni and Lenci, 2010). However, this area of research still represents some current issues concerning the building, the evaluation and the use of distributional resources[1]. We focus here on the building of distributional resources.

In that respect, during the past few years, research has shifted from using DA methods for modelling the semantics of words to tuning them for the semantics of larger units such as phrases or entire sentences (Hermann et al., 2013). Most approaches tackle the problem through vector composition. Mitchell and Lapata (2008) use linear algebraic vector operations, testing both additive and multiplicative models, and a combination of these models. Grefenstette and Sadrzadeh (2011) apply unsupervised learning of matrices for relational words to their arguments, in order to compute the meaning of intransitive and transitive sentences. Baroni and Zamparelli (2010) use matrices to model meaning, but only for adjective-noun phrases, whereas Grefenstette and Sadrzadeh (2011)'s work also applies to sentences containing combinations of adjectives, nouns, verbes and adverbs. Recently, the framework proposed by Grefenstette et al. (2013) combines both approaches.

An important issue in DA improvement focuses on distributional contexts, and more precisely on weighting contexts. Broda et al. (2009) consider that what matters is not the feature's exact frequency. They do not use these frequencies as simple weights but rank contexts and take into account this rank in DA. Influence on contexts may also be done by embedding additional semantic information. With a method based on bootstrapping, Zhitomirsky-Geffet and Dagan (2009) modify the weights of the elements in contexts relying on the semantic neighbours found with a distributional similarity measure. Based on this work, Ferret (2013) faces the problem of low frequency words by using a set of positive and negative examples selected in an unsupervised way from an original distributional thesaurus to train a supervised classifier. This classifier is then applied for reranking the semantic neighbours of the thesaurus selection. With the same purpose of solving the problem of data sparseness, other methods are based on dimensionality reduction, as Latent Semantic Analysis (LSA) in (Padó and Lapata, 2007), or on a bayesian approach of DA (Kazama et al., 2010).

Above work exploits great collection of texts of general language. However, few works are also interested in applying DA to specialized domains where text collections are generally smaller and frequencies lower (Habert and Zweigenbaum, 2002; Embarek and Ferret, 2008). As presented previously, Ferret (2013) attempts to exploit machine learning approaches to face the problem of low frequency of words and contexts. In our work, we propose an approach that exploits relations acquired with linguistic approaches in order to normalize contexts and increase their frequency. As we work with specialized texts, our approach dif-

[1] See for instance, the recent workshop at ACL 2013 https://sites.google.com/site/cvscworkshop/ or at the TALN 2013 http://www.taln2013.org/ateliers/appel-atelier-semantique-distributionnelle/

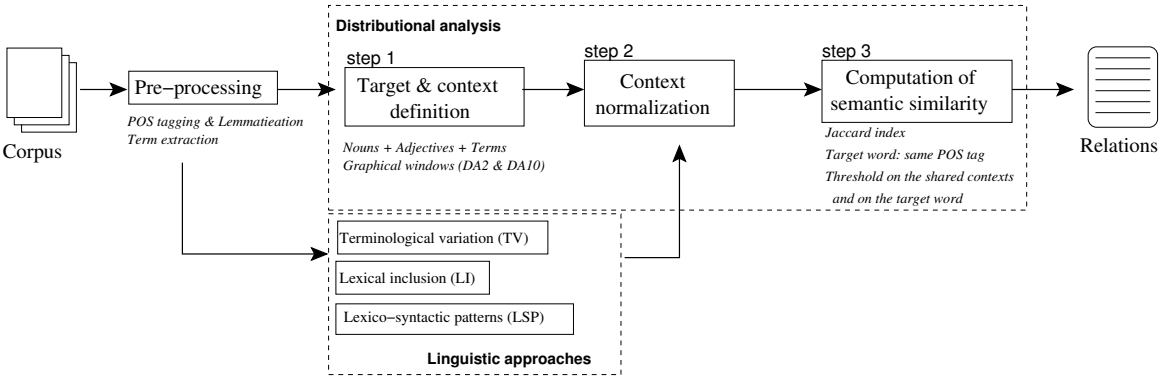


Figure 1: Processing steps

fers in considering nouns, adjectives and both simple and complex terms.

3 Hybrid method

The contexts in which occurs a target word have associated frequencies which may be used to form probability estimates. The goal of our hybrid method is to influence those distributional context frequencies by normalizing contexts. Indeed, normalization tends to decrease diversity in contexts in order to increase contexts' frequency. Our hybrid method follows the scheme presented in figure 1.

Target and context definition During Step 1, we define target words and contexts. Through the literature, syntactic analysis is mainly used to get dependency relations. But as it is time-consuming and heavy, we choose to use instead graphical windows within a sentence and around the target word. As we work on specialized texts, we also identify terms with the term extractor YATEA (Aubin and Hamon, 2006).

We define the following parameters:

- Target words: words are in relation when they have the same POS tag; restricted to adjectives, nouns and terms.
- Distributional contexts: contexts are made of words that co-occur in a graphical window. In contexts, we don't take into account non-content words (determiners, conjunctions, adverbs, etc.) and keep only adjectives, nouns, verbs and terms.
- Fixed window size: we tested two different sizes described in section 4.
- Word form: for both contexts and target words we use the lemmas.

Linguistic approaches During the normalization process described below, we use three existing linguistic approaches: two methods that aim at acquiring hyperonymy relations and one that allows to get morphosyntactic variants.

- Lexico-syntactic Patterns (LSP): we use the patterns defined by (Morin and Jacquemin, 2004):

  1. {some | several etc.} NP : LIST.
  2. {other}? NP such as LIST.

  where NP is a noun phrase and LIST a list of noun phrases.

- Lexical Inclusion (LI): uses the syntactic analysis of the terms. Based on the hypothesis that if a term (ex: *épice (spice)*) is lexically included in another (ex: *épice aromatique (aromatic spice)*), there is a hyperonymy relation between the two terms generally (Bodenreider et al., 2001).

- Terminological Variation (TV): uses rules that define a morpho-syntactic transformation. This transformation may be an insertion, as the insertion of the adjective *aromatic* in *épice asiatique (asian spice) - épice aromatique asiatique (asian aromatic spice)* (Jacquemin, 1996).

Context normalization Once targets and contexts are defined comes the core of the hybrid method with context normalization. During

Step 2, we normalize contexts with the relations acquired by the three linguistic approaches we mentionned.

The relations are integrated in contexts in the following way: a word in context is replaced by its hyperonym or its morphosyntactic variant. We define two rules :

- If the word in context matches with only one hyperonym, context is replaced by this hyperonym. For example, if LSP give the relation *matière grasse (fat)/beurre (butter)*, *beurre (butter)* is replaced by *matière grasse (fat)*.
- If the context matches with several hyperonyms or variants, we take the hyperonym's or variant's frequency into account, and choose the one that is the most frequently in relation with the word in context. For example, if LSP give the following supposed hyperonyms: *matière grasse (fat)*, *pâte feuilletée (falky pastry)*, *béchamel*, *casserole (sauce pan)*, the one that enters the most frequently in relation with *beurre (butter)* is selected and used to replace this word in context.

We normalize contexts with each method separetly and sequentially: the first normalization is processed on all contexts before the second normalization starts, and so on.

Computation of semantic similarity When contexts have been normalized, similarity between two target words of the same POS tag is computed. As we decrease diversity in contexts during the normalization step, we choose among the existing measures (Weeds et al., 2004) a measure that favors words appearing in similar contexts compared to words appearing in many different contexts.

The Jaccard Index (Grefenstette, 1994) normalizes the number of contexts shared by two words by the total number of contexts of those two words.

$$sim{-}JACCARD_{mn} = \frac{|ctxt(w_m) \bigcap ctxt(w_n)|}{|ctxt(w_m) \bigcup ctxt(w_n)|}$$

Parameter: threshold We filter the relations according to three parameters, two of them applied on the contexts and the third one on the target.

- Number of shared contexts: number of lemmatized contexts. For example, if two words share *crème (cream)*, *battre (shake)*, *poivre (pepper)*, *sel (salt)*, *crèmes (creams)*, *battant (shaking)*, the number of shared contexts is 4.
- Frequency of the shared contexts: number of occurrences of the same lemma when shared in the context position of two target words. In the previous example, frequencies are *crème (cream)-2*, *battre (shake)-2*, *sel (salt)-1* and *poivre (pepper)-1*.
- Frequency of the target words: number of occurrences of the lemma in the target position.

For each parameter, a threshold is automatically computed, according to the corpus. It corresponds to the mean of the values taken by each parameter on the whole corpus.

4 Experiments

In order to evaluate the contribution and influence of relations acquired by the three methods, we define several sets of experiments and evaluate the relations acquired on existing resources.

4.1 Corpus

We use the merging of the two corpora provided by DEFT 2013 French challenge[2]: the training corpus (2,388,731 words) and the test corpus (1,539,927 words). They are both French corpora and contain cooking recipes. Each text of the corpus is made of a title, ingredients and the body of the recipe, and we use all the information.

We pre-process the corpus within the Ogmios platform (Hamon et al., 2007). We perform morphosyntactic tagging and lemmatization with Tree Tagger (Schmid, 1994), and use the term extractor YaTeA(Aubin and Hamon, 2006).

4.2 Parameters and models of hybridization

In these different sets we vary two main kinds of parameters (cf. table 1): window size and models of hybridization.

We test two window sizes. With a large one of 20 words around the target (10 before, 10 after, henceforth W10) we may take into account the highest number of possible relations, because the average size of a sentence in French is 20 words

[2]http://deft.limsi.fr/2013/

Window size	4 (W2) and 20 (W10) words around the target
Hybridization	none: DAonly one method: DA/LSP, DA/LI, DA/TV two method combination: DA/LI+LSP, DA/LSP+LI, DA/TV+LSP, DA/LSP+TV three method combination: DA/LI+LSP+TV, DA/LSP+LI+TV, DA/TV+LSP+LI

Table 1: Parameters

and we restricted the relation acquisition to the sentence level. But such a large window may face a lack of specificity and get too much noise. We also test a window of 4 words (2 before, 2 after, henceforth W2). Such a size applied after removing the function words is comparable to a 8 word window applied to the original texts (Rapp, 2003).

We test different models of hybridization. We first use DA on its own, without normalizing the contexts (DAonly). This set is a reference to which compare the hybridization sets. As for the models of hybridization, we first separately evaluate the contribution of each method (LSP, LI, TV) in distributional context, and then different types of combinations of the methods integrated in DA. Within these combinations, we first exploit two methods together and then three. Our goal is to evaluate the impact of the order of the methods and the contribution of each method.

4.3 Comparison with existing resources

In order to evaluate the quality of the acquired relations, we compare our relations with three different resources: Agrovoc[3], of 75,222 relations [AGRO], and UMLS[4].

With the UMLS resource, we build two different resources: one more general [UMLS] of 2,325,006 relations, and a more specific one restricted to terms belonging to the *Food* concept (semantic type T168) [UMLS/Food] of 1,843 relations.

We only use the relations for the nouns and terms of our corpus, because adjectives were not represented in the resources. In that respect, we evaluate our work with 1,551 ([AGRO]), 1,800 ([UMLS]) and 871 ([UMLS/Food]) relations.

We use those three resources because of availability. The comparison with UMLS/Food and Agrovoc is justified by the presence of relations between food terms in both resources. But in cooking recipes, we may find other types of relations, as the relation between a food term and a term belonging to another semantic class. The comparison with the whole UMLS may allow to detect other relations than ingredient relations. Even if we can not expect an important overlap between these resources and the corpus, the comparison of our results to the relations issued from these resources gives an indication of the contribution of each proposed hybridization model.

We compute precision for each target term: semantic neighbours (acquired by our method) found in the resource by the semantic neighbours acquired by our method. For each target term, we sorted the semantic neighbours we obtained according to their similarity measure, and apply four thresholds: precision after examining 1 (P@1), 5 (P@5), 10 (P@10) and 100 (P@100) neighbours.

5 Results and discussion

We proceed to the analysis and discussion of the results we obtain with our hybrid method. Regarding the relations provided by the terminologies, we present here the results obtained for nouns and terms only.

We evaluate precison after examining four groups of neighbours. The best results are obtained with P@1, and decrease when we consider more neighbours: the more neighbours we consider, the lower precision is. For instance, for nouns-W10, precision decreases from 0.089 for P@1 to 0.009 for P@100, when compared with Agrovoc. We make similar observations on all the sets of results. Best results in first position means that the values of the measures rank quite correctly the proposed relations, and therefore that the choice of the measure was a good choice.

The table 2 presents the results for P@1, given the two window sizes (W10 and W2). We describe here only those results. The relations produced by DA (DAonly) are considered as our baseline. The low precision of our results was expected and

[3]http://aims.fao.org/standards/agrovoc/about

[4]http://www.nlm.nih.gov/research/umls/

Resources	Context definition and window size	DAonly	DA/TV	DA/LI	DA/LSP	DA/TV+LSP	DA/LI+LSP	DA/LSP+TV	DA/LSP+LI	DA/TV+LSP+LI	DA/LI+LSP+TV	DA/LSP+LI+TV
Agrovoc	Noun-W2	0.024	0.024	0.024	**0.073**	**0.072**	0.067	**0.073**	0.039	0.039	0.067	0.039
	Noun-W10	0.089	0.089	0.071	**0.109**	**0.109**	**0.111**	**0.109**	0.071	0.056	**0.111**	0.071
	Term-W2	0.023	0.024	0.000	**0.034**	**0.034**	0.000	**0.034**	0.000	0.000	0.000	0.000
	Term-W10	0.010	0.010	**0.031**	0.000	0.000	0.000	0.000	**0.047**	**0.047**	0.000	**0.047**
UMLS	Noun-W2	0.098	0.098	**0.139**	0.074	0.074	0.034	0.074	0.051	0.051	0.034	0.051
	Noun-W10	0.088	0.088	0.086	0.077	0.077	0.038	0.077	0.086	0.081	0.038	0.086
	Term-W2	0.094	**0.100**	0.000	**0.115**	**0.120**	0.000	**0.115**	0.000	0.000	0.000	0.000
	Term-W10	0.037	**0.042**	0.000	**0.042**	**0.043**	0.000	**0.042**	0.000	0.000	0.000	0.000
UMLS/food	Noun-W2	0.059	0.059	**0.094**	**0.083**	**0.080**	0.034	**0.083**	0.054	0.054	0.034	0.054
	Noun-W10	0.075	0.075	**0.095**	0.074	0.074	0.038	0.074	**0.095**	**0.102**	0.038	**0.095**
	Term-W2	0.070	0.071	0.000	**0.097**	**0.097**	0.000	**0.097**	0.000	0.000	0.000	0.000
	Term-W10	0.026	0.026	0.000	**0.032**	**0.031**	0.000	**0.032**	0.000	0.000	0.000	0.000

Table 2: Precision of the results against each resource after examining the first neighbour (P@1)

can be explained by the fact that even if the resources are relevant for our corpus, they are not fully adapted. However, the comparison of the precision values gives important information on the usefulness of the hybridization models.

Results are better for nouns (between 0.056 and 0.111 for nouns-W10 and between 0.024 and 0.073 for nouns-W2, with Agrovoc) than for terms (between 0 and 0.047 for terms-W10, and between 0 and 0.34 for terms-W2, with Agrovoc). This is not surprising because terms do not match easily with other terms in resources. This can be due to two main factors: terms are less frequent and it is difficult to match terms from the terminological resources in the corpus. As for the window size, we observe that generally W10 gives good results for nouns and W2 is better with terms. But when we look more in details, we observe that the quality of the results depends on the resource used for comparison. For nouns, with Agrovoc and UMLS/food, W10 gives the best results, but when compared with UMLS results are better with W2. The difference is similar with terms, but in this case results are better with W2 when compared with UMLS and UMLS/food, and better with W10 when compared with Agrovoc.

Linguistic approaches Considering the three methods individually, TV seems to have no influence on the computation of semantic similarity; the results obtained with DAonly and DA/TV are identical, except for terms W2 and W10 when compared with UMLS. Also, in all the hybrid sets, exploiting TV in the distributional contexts doesn't influence the results, except for DA/LI+LSP+TV with nouns-W10 when compared with Agrovoc and UMLS/food, and DA/TV+LSP with term-W2 when compared with UMLS. This may be because of the small number of relations used and our current way of DA hybridization with TV.

On the contrary, LSP is the method that most influences the results: most of the time they give better results than DA. The best hybridization model for terms is the normalization with LSP, whereas for nouns the combination of LI and LSP is the best choice. The order of the methods also matters, but results also differ according to the resource; DA/LSP+LI (and DA/LSP+LI+TV) give better results when compared with UMLS/food, and DA/LI+LSP (and DA/LI+LSP+TV) give better results when compared with Agrovoc. What emerge from these results is that generalization with hyperonyms is the best configuration, for both terms and nouns, and that the quality of the hyperonymy relation is important as well. Lexical inclusion used after patterns does not seem to bring new relations but allow to rule out noisy relations. By noisy relations, we mean relations not found in the resource. But these relations may be interesting and may be domain relations.

Resources and relation types The relations found by our method in UMLS/Food are co-

hyponyms (eg: *ail (garlic)/oignon*), those found in Agrovoc are both hyperonyms (eg: *épice (spice)/poivre (pepper)*) and meronyms (eg: *miel (honey)/sucre (sugar)*). Relations found in the whole UMLS are the same as those found in UMLS/food. The identification of terms allow to find more relations, between simple terms and complex terms. For instance, in UMLS/food, our method found the co-hyponyms *poivre blanc (white pepper)/poivre noir (black pepper)* and *miel (honey)/fruit* that are not identified by taking into account nouns only.

6 Conclusion

In this work, we present our hybrid method based on normalization of distributional contexts. Our method aims at acquiring semantic relations from specialized texts, and is adapted to low frequency words. We normalize contexts with relations acquired by three linguistic approaches; two methods of hyperonymy relation acquisition and a method of morpho-syntactic variant acquisition. We focus on relations between nouns and terms. We tested our method on a French corpus composed of cooking recipes, varying one parameter in our DA method, the window size, and testing different models of normalization. Normalization obtains the best results when realized with hyperonyms and also depends on the quality of the hyperonymy relations. In our method, the hyperonym used for normalization is the one with the highest frequency. Even if precision values presented in this work are currently low and results differ according to the resource used for evaluation, it emerges that the best parameters are for nouns a W10 with LSP, and for terms a W2 with LSP and LI. This set of parameters is to be used for classical types of relations. But other types may be acquired with our DA++ method, especially domain specific relations. In order to have a better knowledge of the influence of each hybridization model, quality of the results has to be analyzed more deeply by manually checking with the validation of a subset of relations, and with a study of relations that are in common or not between the various results sets. For future work, we plan to investigate other strategies of normalization by assigning a weight to the relations proposed by the linguistic methods, or taking into account the level in the hierarchy. In that latter approach, the choice of the hyperonym used for the normalization could be guided by a distance (Resnik, 1995; Leacock and Chodorow, 1998). Relations used for normalization can also be issued from terminological resources. Furthemore, we will intend to combine the methods before normalization and exploit other similarity measures.

References

Sophie Aubin and Thierry Hamon. 2006. Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, number 4139 in LNAI, pages 380–387.

Alain Auger and Caroline Barriere. 2008. Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*, 14(1):1–19.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP 2010*, pages 1183–1193, Stroudsburg, PA, USA.

Olivier Bodenreider, Anita Burgun, and Thomas Rindflesch. 2001. Lexically-suggested hyponymic relations among medical terms and their representation in the umls. In *Proceedings of TIA 2001*, pages 11–21, Nancy, France.

Didier Bourigault and Monique Slodzian. 1999. Pour une terminologie textuelle. *Terminologies Nouvelles*, 19:29–32.

Bartosz Broda, Maciej Piasecki, and Stan Szpakowicz. 2009. Rank-based transformation in measuring semantic relatedness. In *Canadian Conference on AI*, volume 5549, pages 187–190.

Sharon A. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *ACL*, pages 120–126.

Mehdi Embarek and Olivier Ferret. 2008. Learning patterns for building resources about semantic relations in the medical domain. In *Proceedings of LREC 2008*, Marrakech, Morocco. ELRA.

Cécile Fabre and Didier Bourigault. 2006. Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In *TALN 2006*, pages 121–129, Leuven.

Olivier Ferret. 2013. Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. In *TALN 2013*, pages 48–61, Les Sables d'Olonne, France.

Natalia Grabar and Pierre Zweigenbaum. 2000. Automatic acquisition of domain-specific morphological resources from thesauri. In *RIAO*, pages 765–784.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP 2011*, pages 1394–1404.

Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. 2013. Multi-step regression learning for compositional distributional semantics. *Proceedings of IWCS 2013*.

Gregory Grefenstette. 1994. Corpus-derived first, second and third-order word affinities. In *Sixth Euralex International Congress*, pages 279–290.

Benoit Habert and Pierre Zweigenbaum, 2002. *Contextual Acquisition of Information Categories: what has been done and what can be done automatically?*, pages 203–231. Nevin (Bruce), Amsterdam.

Thierry Hamon, Adeline Nazarenko, Thierry Poibeau, Sophie Aubin, and Julien Derivière. 2007. A robust linguistic platform for efficient and domain specific web content analysis. In *Proceedings of RIAO*, Pittsburgh, USA.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *International Conference on Computational Linguistics*, pages 539–545, Nantes, France.

Gérard Henneron, Rosalba Palermiti, and Yolla Politi. 2005. *L'organisation des connaissances, approches conceptuelles*. Harmattan.

Karl Moritz Hermann, Edward Grefenstette, and Phil Blunsom. 2013. "not not bad" is not "bad": A distributional account of negation. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 74–82, Sofia, Bulgaria.

Christian Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In *CoRR*, pages 425–438.

Christian Jacquemin. 1997. Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus. Mémoire d'HDR en informatique, Université de Nantes.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of ACL 1999*, pages 341–348, University of Maryland.

Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa. 2010. A bayesian method for robust estimation of distributional similarities. In *In proceedings of ACL 2010*, pages 247–256, Stroudsburg, PA, USA.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In *Proceedings of MIT Press*, pages 265–283, Cambridge, Massachusetts.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL-08: HLT*, pages 236–244, Columbus, Ohio.

Emmanuel Morin and Christian Jacquemin. 2004. Automatic Acquisition and Expansion of Hypernym Links. *Computers and the Humanities*, 38(4):363–396.

Emmanuel Morin. 1999. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse de doctorat, Institut de recherche en informatique de Nantes.

François Morlane-Hondère and Cécile Fabre. 2012. Étude des manifestations de la relation de méronymie dans une ressource distributionnelle. In *TALN'2012*, pages 169–182, Grenoble, France.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Comput. Linguist.*, 33(2):161–199.

Reinhard Rapp. 2003. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of MT Summit'2003*, pages 315–322.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI'1995*, pages 448–453, San Francisco, CA, USA.

Magnus Sahlgren. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University, Stockholm, Sweden.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *New Methods in Language Processing*, pages 44–49, Manchester, UK.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of NIPS 17*, pages 1297–1304.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING'2004*, Stroudsburg, PA, USA.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3):435–461.

Filtrage terminologique par le lexique transdisciplinaire scientifique : une expérimentation en sciences humaines

Evelyne Jacquey
Atilf UMR 7118 – CNRS/Université de Lorraine
evelyne.jacquey@atilf.fr

Agnès Tutin
Lidilem EA 609 – Université Grenoble Alpes
agnes.tutin@u-grenoble3.fr

Laurence Kister
Atilf UMR 7118 – CNRS/Université de Lorraine
laurence.kister@univ-lorraine.fr

Marie-Paule Jacques
Lidilem EA 609 – Université Grenoble Alpes
marie-paule.jacques@ujf-grenoble.fr

Sylvain Hatier
Lidilem EA 609 – Université Grenoble Alpes
sylvain.hatier@u-grenoble3.fr

Sandrine Ollinger
Atilf UMR 7118 – CNRS/Université de Lorraine
sandrine.ollinger@atilf.fr

Résumé

Nous examinons les interactions syntaxiques entre termes et lexique transdisciplinaire avec un objectif de reconnaissance automatique des termes en texte intégral dans le discours scientifique des sciences du langage. Nous réalisons une expérience sur un corpus d'articles scientifiques (155 157 occurrences - 42 articles de Scientext) pour déterminer dans quelle mesure et sous quelles conditions les unités du lexique transdisciplinaire (ULT) sont délimiteurs de termes (T). Après une annotation terminologique validée manuellement (ACABIT et TERMOSTAT), un lexique transdisciplinaire est projeté dans les textes. L'analyse quantitative et qualitative des résultats permet de conclure que les ULT sont majoritairement délimiteurs de termes. Les résultats montrent que : (1) le statut d'ULT doit être raffiné (unité phraséologique transdisciplinaire ou langue générale ayant un rôle de délimiteur), (2) la contribution du lexique transdisciplinaire doit être corrélée à une classification sémantique des ULT pour définir les différents rôles joués, (3) certaines ULT sont en intersection avec le domaine de spécialité d'où un biais réductible par l'intégration d'autres disciplines des sciences humaines (archéologie, psychologie, philosophie, etc.).

1 Introduction

Cet article décrit une expérimentation visant à faciliter la reconnaissance automatique des termes (T) en texte intégral. Elle contribue à un objectif final d'indexation de textes via la reconnaissance automatique de termes en textes intégraux[1]. Nous cherchons à affiner cette reconnaissance en nous focalisant sur l'environnement textuel des candidats termes, dans une philosophie similaire à celle de (Bachimont et *al.* 2005 ; Bourigault et *al.* 2001 ; Bourigault et *al.* 2004). Nous nous inscrivons dans le champ de la terminologie textuelle (Bourigault et Slodzian 1999) qui appréhende les termes dans leur fonctionnement textuel.
L'hypothèse que nous testons est celle d'une interaction privilégiée entre les unités de lexique transdisciplinaire scientifique (ULT) et les termes du vocabulaire de spécialité d'une discipline. Plus précisément, à la suite de (Kister et Jacquey 2012), nous faisons l'hypothèse que les co-occurrences portées par une relation syntaxique entre les unités du lexique transdisciplinaire (ULT) et les termes peuvent être des indices du statut terminologique des termes. Par exemple, dans le domaine des sciences du langage, les termes *diglossie* et *locution* apparaissent dans les textes avec *concept* et *analyser* : *le concept de diglossie* et *nous analysons les locutions*.

[1] TermITH : ANR-12-CORD-0029 CONTINT. ATILF, INIST, LIDILEM, LINA, INRIA NGE et Saclay : http://www.atilf.fr/ressources/termith/ (page consultée le 17/09/2013).

L'expérience de (Kister et Jacquey 2012), qui portait sur un petit corpus contrastif scientifique *vs* vulgarisation, a montré l'intérêt de mener une expérience à plus grande échelle : extension du corpus, automatisation, structuration sémantique du lexique transdisciplinaire (toutes les ULT ne sont pas équivalentes du point de vue de l'hypothèse examinée).

Notre expérience, centrée sur le discours scientifique, automatise deux étapes clés : (1) la projection du lexique transdisciplinaire, (2) la détection et la qualification des relations syntaxiques T-ULT. Cette étape d'automatisation permet ainsi d'augmenter la taille du corpus de travail.

Enfin, nous effectuons une analyse quantitative et qualitative des relations détectées. L'analyse qualitative est réalisée en fonction d'une première structuration sémantique du lexique transdisciplinaire. Des pistes d'exploration sont proposées pour affiner les procédures dans la perspective d'une exploitation automatique du lexique transdisciplinaire.

2 Le lexique transdisciplinaire : délimiteur de termes

Le lexique transdisciplinaire des écrits scientifiques est un lexique particulier qui ne renvoie pas aux objets scientifiques des domaines de spécialité, mais plutôt aux discours sur les objets et les procédures scientifiques (Tutin 2007). Il est mis en œuvre dans la description et la présentation de l'activité scientifique et est ainsi partagé par la communauté scientifique. Il s'agit donc d'un lexique de *genre* plus que d'une terminologie propre à un domaine. Il est en grande partie partagé par de nombreuses disciplines, même si des différences se font jour entre familles de disciplines.

Ce lexique concerne ainsi des unités lexicales qui renvoient aux procédures scientifiques (*étudier, analyser, recenser*), aux opinions (*de notre point de vue, nous pensons*), à l'évaluation (*valide, intéressant, pertinent*), aux artefacts scientifiques (*approche, hypothèse, modèle*), aux observables (*données, résultats*). Il intègre à la fois des mots simples et de très nombreuses séquences lexicalisées aux fonctions diversifiées (Tutin, à paraître).

Pour les mots simples, à la suite du Vocabulaire Général d'Orientation Scientifique (Phal 1971), plusieurs listes ont été proposées pour l'anglais (Coxhead 2000 ; Bolshakova 2008 ; Paquot 2010) et pour le français (Drouin 2007) et (Tutin 2007). Ces listes ont été constituées sur la base de critères statistiques (distribution dans plusieurs disciplines, fréquences, spécificité). Dans une perspective d'indexation de textes, nous envisageons d'utiliser le lexique constitué par fusion de ceux de (Drouin 2007) et (Tutin 2007) de deux manières :

- comme lexique d'exclusion pour l'extraction des termes simples et complexes, en particulier en exploitant les co-occurrences définies par une fenêtre de mots ou la co-occurrence syntaxique. Par exemple, le mot *sujet* pourra être considéré comme un terme en co-occurrence avec *phrase* et *objet direct* mais ne le sera pas dans le cas d'une co-occurrence avec *article, aborder* et *analyse.*
- comme lexique d'introduction des termes (l'hypothèse que nous examinons). Nous pensons qu'un sous-ensemble de ce lexique est particulièrement susceptible d'introduire des termes : par exemple, *le* [*concept*]$_{\text{ULT}}$ *de [diachronie]*$_T$, *nous* [*analysons*]$_{ULT}$ la *[diachronie]*$_T$. Nous supposons que ces unités lexicales, propres au genre des écrits scientifiques, sont souvent introductrices de termes parce qu'elles sont associées à des procédures scientifiques ou des commentaires sur les concepts du domaine.

En approfondissant et en cherchant à évaluer cette dernière hypothèse, nous espérons contribuer à l'extraction et à la reconnaissance automatique des termes dans les textes, sous l'angle de la délimitation des unités terminologiques d'un domaine par le biais du lexique transdisciplinaire.

3 Méthodologie

Comme brièvement décrit dans l'introduction, plusieurs phases segmentent les traitements réalisés sur les textes (*cf.* Tableau 1).

Phase 1	Projection du lexique transdisciplinaire sous forme lemmatisée dans les articles annotés en termes
Phase 2	Détection des co-occurrences entre ULT et termes à l'intérieur des phrases segmentées par l'étiqueteur morpho-syntaxique (TreeTagger)
Phase 3	Analyse syntaxique des relations de co-occurrences et qualification des types de relations à l'œuvre[2] (XIP[3])
Phase 4	Filtrage et analyse des résultats

Tableau 1 : Les différentes étapes de l'expérience

3.1 Corpus de travail, ressources lexicales et terminologiques

Les articles de l'expérimentation sont extraits du corpus Scientext[4] qui est composé de documents appartenant tous au domaine des sciences du langage (42 textes répartis en 22% d'articles et 78% de communications - 155 157 occurrences).
La ressource terminologique de référence est constituée du vocabulaire Francis[5] (6 100 entrées) et des termes de Thesaulangue[6] (1 200 entrées).
Le lexique transdisciplinaire utilisé est le résultat d'une fusion des lexiques de Tutin (2007) et Drouin (2007), restreint aux entrées nominales et verbales afin de réduire le bruit, ce qui entraîne l'exclusion des adjectifs. Il comporte 390 noms (*cas*, *étude*, *travail*, *type*) et 321 verbes (*agir, considérer, correspondre)*, soit 711 entrées.

3.2 Annotation terminologique

L'annotation terminologique est réalisée en deux étapes principales.
Les extracteurs ACABIT (Daille 1996) et TERMOSTAT (Drouin 2003a et b) fournissent des listes de candidats termes en appliquant sur les articles donnés en entrée, un jeu de règles linguistiques et de critères statistiques.
Après avoir été projetés dans les articles sources, les candidats termes sont sélectionnés manuellement pour ne retenir que les occurrences terminologiques valides, qu'il s'agisse d'unités simples ou complexes. Pour déterminer la validité des candidats, nous avons eu recours à des experts linguistes qui ont pu, si besoin, utiliser les deux ressources terminologiques de référence, Thesaulangue et le vocabulaire Francis (*cf.* section 3.1).
Parmi les 43 324 occurrences de candidats, nous avons conservé 11 772 occurrences jugées valides (21%). A l'issue de cette sélection, les candidats validés sont considérés comme des occurrences de termes.

3.3 Projection du lexique transdisciplinaire

La projection du lexique transdisciplinaire est réalisée par un module distinct qui prend en entrée les articles enrichis en termes et le lexique transdisciplinaire sélectionné. A l'issue d'un étiquetage morpho-syntaxique (TreeTagger ré-entraîné dans le cadre du projet PERCEO[7]) et d'une projection du lexique transdisciplinaire, l'ensemble des co-occurrences d'une même phrase est reporté en tenant compte de la combinatoire possible. Une même phrase qui contient deux termes (T_1 et T_2) et 2 ULT (ULT_1 et ULT_2) différents est reportée 4 fois en mettant en valeur à chaque fois un couple (T_i, ULT_j) différent. A l'issue de ce traitement, nous obtenons 53 281 relations de co-occurrence impliquant 1 273 termes différents : *corpus* 1 417 relations, *mots*[8] 1 156, *mot* 1 068, *verbe* 1 067, *sens* 814, etc.
Ensuite, les phrases contenant des co-occurrences T-ULT sont analysées syntaxiquement (à l'aide de XIP) afin de déterminer si la relation de co-occurrence est une relation syntaxique de dépendance et de préciser la nature de la dépendance quand il y en a une.

[2] Aït-Mokhtar *et al.* (2002)
[3] Xerox Incremental Parser (XIP)
[4] ANR Scientext piloté par le Lidilem – Grenoble 3 : http://scientext.msh-alpes.fr/scientext-site (page consultée le 17/09/2013).
[5] Francis : base de données de références bibliographiques en sciences humaines et sociales constituée par l'Inist.
[6] Thesaulangue : thesaurus en sciences du langage conçu et maintenu par le centre de documentation de l'Atilf – UMR 7118.
[7] http://www.cnrtl.fr/corpus/perceo/ (page consultée le 17/09/22013)
[8] Pour l'enrichissement en candidats termes proposés par ACABIT et TERMOSTAT, nous avons choisi les formes plein-texte afin d'éviter de niveler des distinctions terminologiques pouvant être importantes.

3.4 Détection et qualification des relations syntaxiques entre termes et ULT

Comme indiqué précédemment, les phrases traitées sont analysées en dépendance avec XIP de façon à repérer automatiquement les relations syntaxiques de dépendance entre un terme et une ULT.
Sur les 53 281 co-occurrences T-ULT du corpus, 4 565 relations de dépendance directes ont été repérées. Ces relations syntaxiques de dépendance sont classées par ULT, par nature de la relation de dépendance et par orientation de cette relation (ULT rectrice ou régie). Nous avons ainsi pu identifier les associations lexico-syntaxiques les plus fréquentes ($relation_i$, ULT_j).

3.5 Relations de dépendances

L'analyse quantitative et qualitative des résultats obtenus porte sur les seules relations directes et plus particulièrement sur trois patrons de relations : NMOD[9], OBJ, SUBJ et leurs alternances passives DEEPSUBJ et DEEPOBJ. Ces trois types de relations ont été choisis, d'une part, pour leur productivité (3 406 sur 4 565) et, d'autre part, pour leur représentativité. La relation de type NMOD s'établit principalement au sein du groupe nominal complexe. Les relations de type OBJ et SUBJ apparaissent dans le domaine verbal. De cette manière, nous espérons couvrir tout ou partie des cas les plus intéressants d'interaction entre termes et ULT.
Nous présentons ci-dessous des exemples de relations entre un *[terme]$_T$* régi et une *[unité du lexique transdisciplinaire]$_{ULT}$* rectrice :

- *Par conséquent, nous [analyserons]$_{ULT}$ uniquement les [textes]$_T$ des bilingues dans cet article.*

  OBJ ([textes]$_T$, [analyserons]$_{ULT}$)

- *En conséquence, il semble qu'on ne puisse pas [considérer]$_{ULT}$ le tunisien comme une [langue à satellites]$_T$.*

  OBJ ([langue à satellites]$_T$, [considérer]$_{ULT}$)

- *On s'intéresse en particulier, dans cet article, aux [travaux]$_{ULT}$ sur la [morphologie]$_T$ des adjectifs.*

  NMOD([morphologie]$_T$, [travaux]$_{ULT}$)

- *Le [bilinguisme]$_T$ est [défini]$_{ULT}$ comme la possession d'une compétence de locuteur natif dans les deux langues.*

  DEEPOBJ ([bilinguisme]$_T$, [défini]$_{ULT}$)

- *Certains mécanismes peuvent être [contrôlés]$_{ULT}$ par les [locuteurs]$_T$.*

  DEEPSUBJ ([locuteurs]$_T$, [contrôlées]$_{ULT}$)

3.6 Filtrage manuel des résultats

Nous avons écarté les erreurs d'analyse syntaxique parmi les 3 406 relations étudiées ce qui nous conduit à rejeter 768 analyses erronées : erreur de recteur ou de régi, erreur de relation.
Puis, nous avons classé les 2 638 couples restants (T_i, ULT_j) : les cas pour lesquels l'ULT joue un rôle de délimiteur de terme et ceux pour lesquels l'ULT ne joue pas ce rôle[10]. Chaque couple analysé est ainsi placé dans l'une des trois catégories suivantes :

- Catégorie 1 : l'hypothèse est vérifiée, car l'ULT entretient une relation syntaxique avec le terme
- Catégorie 2 : l'hypothèse n'est pas vérifiée, car l'ULT a une valeur terminologique ou de langue courante évidente
- Catégorie 3 : les cas douteux.

4. Analyse quantitative

La répartition obtenue pour les 2 638 relations analysées est présentée dans le tableau 2.

[9] NMOD : relation entre un nom et ses modifieurs.
OBJ : relation entre un verbe et un complément d'objet direct ou l'attribut du sujet.
SUBJ : relation entre un prédicat et un sujet.
DEEPSUBJ : relation entretenue par un prédicat et un complément d'agent dans une tournure passive.
DEEPOBJ : relation qui s'établit entre un prédicat et son sujet dans une tournure passive.

[10] La valeur terminologique des candidats termes co-occurrents n'est plus à mesurer car l'annotation terminologique a déjà rempli cette tâche. (*cf.* section 3.2)

Relation	Conformité hypothèse	Non-conformité Hypothèse	Cas douteux	Total	Total / type de relation	Proportion de couples conformes à l'hypothèse
NMOD	877	142	32	1 051	1 422	75%
~NMOD	186	179	6	371		
OBJ	378	132	6	516	624	73%
DEEPOBJ	76	30	2	108		
SUBJ	435	2	142	579	592	74%
DEEPSUBJ	4	5	4	13		
Total	1 734	490	415	2 638		

Tableau 2 : Taux de conformité des relations

Pour certaines occurrences, la configuration est très claire : l'ULT putative a pour fonction d'exposer des procédures et des outils de l'activité scientifique considérée.

Dans le cas de notre étude, les quatre textes [constituant]$_{ULT}$ notre [corpus]$_{T}$ de travail ont été automatiquement annotés .

Le même verbe peut avoir une valeur relevant de la langue générale comme *constituant* dans l'exemple suivant :

Les enfants marchent : chaîne marquée à trois éléments : déterminant + nom [constituant]$_{\sout{ULT}}$ le [groupe nominal]$_{T}$ sujet, verbe marcher au présent.

Dans certains contextes, l'ambiguïté est importante au point de rendre une décision spontanée difficile. Pour ces cas, on s'interroge au sujet de l'emploi terminologique ou langue générale de l'ULT.

Il est important de noter que les phrases ci-dessus ont été [produites]$_{ULT}$ par des [enfants francophones]$_{T}$ […]

Dans ce cas, *produire* a un sens ambigu entre l'acception de langue générale plus proche du verbe support *faire* et le sens terminologique de *production langagière* par opposition à *compréhension*.

Un dernier type de difficultés tient à notre méthode qui isole les phrases et n'offre pas toujours un contexte suffisant pour comprendre le sens de l'occurrence. C'est le cas lorsqu'on est en présence d'anaphores :

La première [concerne]$_{ULT}$ le désintérêt affiché pour les tâches scolaires d'[écriture]$_{T}$, précédemment cité (cf. section 4.1)

Dans cet exemple, il est difficile de retrouver l'antécédent de *la première*.

5. Analyse qualitative

A côté de l'analyse quantitative, nous avons souhaité observer plus finement les ULT ou les classes d'ULT les plus susceptibles d'introduire des termes. Nous faisons, en effet, l'hypothèse que certaines ULT qui ont une fonction métalinguistique (*concept, terme*) ou introduisent des procédures scientifiques *(étudier, analyser*) sont plus propices à l'introduction de termes. Pour cette tâche, nous avons effectué un typage sémantique des ULT les plus productives en observant également les ULT projetées n'entrant pas dans ces relations. Nous nous sommes pour cela basés sur un typage effectué par Tutin (2007) étendu pour cette expérience.

5.1 Les ULT nominales introductrices de termes

Les noms transdisciplinaires ont été caractérisés dans des grandes classes sémantiques, construites largement à partir de critères distributionnels et inspirées de l'approche de Flaux et Van de Velde (2000). Certaines classes sont assez génériques comme les noms de processus (*choix*), les noms humains *(personne, individu*), les noms quantitatifs *(nombre de, ensemble de*) alors que d'autres sont spécifiques de la langue scientifique comme les noms de processus scientifique (*étude, description, recherche*), les noms d'observables scientifiques avec les objets étudiés par l'activité scientifique (*données, paramètres)*, les noms d'artefact scientifique avec les objets construits par la réflexion scientifique (*approche, méthode, analyse*). Bien entendu, certains noms peuvent relever de plusieurs catégories. La catégorisation a été effectuée en observant l'ensemble des contextes de façon systématique pour les noms les plus productifs de la liste des 390 noms transdisciplinaires.

Classe sémantique	ULT rectrice de terme	Freq[11]	Nb ULT	% cas positifs
processus	changement emploi utilisation choix usage production	92	6	97%
processus scientifique artefact scientifique	analyse construction description recherche étude	101	5	95%
artefact scientifique	structure représentation schéma modèle	44	4	89%
nom quantitatif	ensemble groupe nombre	45	3	96%
nom classifieur	type	38	1	100%
caractérisation	présence	11	1	100%

Tableau 3 - ULT nominales introductrices de termes les plus productives (nom recteur de terme)

Dans la cinquième colonne du tableau 3, on peut constater un très net accroissement de la proportion de cas satisfaisants ce qui confirme l'hypothèse selon laquelle certaines classes de noms spécifiques de la langue scientifique sont clairement des introducteurs privilégiés de termes. C'est en particulier le cas des noms de processus scientifiques : artefacts scientifiques (*analyse, construction, description, recherche, étude*) et artefacts scientifiques purs (*structure, représentation, schéma, modèle***)**. On observe aussi un ensemble de noms a priori moins spécifiques :

- quantitatifs (*ensemble de, groupe de)* ou qualitatifs (*type de*), analysables comme les têtes de déterminants complexes (Flaux et Van de Velde 2000) et probablement analysables comme tels par l'analyseur syntaxique
- de processus généraux, renvoyant à des noms d'action assez abstraits, non spécifiques de l'activité scientifique (*emploi, usage, production*)
- de caractérisation (*présence*) renvoyant à une propriété.

Contrairement à nos attentes, les noms métalinguistiques (*terme, notion, concept, mot, nom*) ne sont pas les meilleurs introducteurs de termes bien que les noms *terme* et *mot* figurent parmi les ULT les plus fréquentes du corpus Scientext dans son ensemble. Dans notre échantillon, seuls *notion* et *concept* sont utilisés à cette fin, *mot, nom* et *terme* sont surtout – ce qui est évidemment étroitement lié à la discipline des textes – utilisés comme des termes. Par ailleurs, on observe que quelques classes de noms transdisciplinaires sont peu employées comme introducteurs de termes :

- les noms de supports écrits de la recherche (*article, chapitre, figure*)
- les noms d'observables (*résultats, données*)
- les noms d'acteur scientifique (*auteur, chercheur*)
- les noms de relation (*cause, conséquence*)
- les noms temporels (*année, moment ou durée*).

En résumé, si l'hypothèse est fortement vérifiée pour certaines classes sémantiques de noms (en particulier pour les processus scientifiques et artefacts scientifiques et pour les processus dans leur ensemble), ce n'est pas le cas d'autres catégories bien représentées dans notre genre, principalement pour deux raisons :

- un sous-ensemble de ces classes renvoie à des noms concrets (*article, figure*) ou humains (*auteur*) un peu moins susceptibles d'entrer dans des structures de type ULT prep T (ils sont plus catégorématiques que syncatégorématiques)
- les classes des relations ou des noms temporels, bien que syncatégorématiques, relient des éléments très abstraits (*événements, faits*) rarement représentés par des termes.

La catégorisation sémantique des noms du lexique transdisciplinaire apparaît donc indispensable pour effectuer un filtrage plus efficace.

[11] Les fréquences indiquées ont été obtenues en additionnant les fréquences absolues respectives de chaque ULT citée dans la colonne précédente et présente dans le corpus de travail.

5.2 Les ULT verbales introductrices de termes

Nous avons également observé le fonctionnement des verbes introducteurs de termes apparaissant avec les relations SUBJ/DEEPSUBJ et OBJ/DEEPOBJ. La liste de verbes utilisés, peu filtrée, a été caractérisée à l'aide de classes de quasi-synonymes pour les éléments les plus productifs (*cf.* Tableau 4).

Classe synonymique	ULT rectrice de terme	Fréq	Nb ULT	% cas positifs
être	être_SUBJ être_OBJ apparaître_SUBJ constituer_SUBJ	223	4	94%
avoir	avoir_SUBJ avoir_OBJ présenter_SUB présenter_OBJ prendre_OBJ	82	5	94%
modaux	pouvoir_SUBJ permetttre_SUBJ sembler_SUBJ devoir_SUBJ	63	4	84%
identification	identifier_OBJ distinguer_OBJ trouver_OBJ	29	3	91%
étude	classer_OBJ considérer_OBJ	14	2	100%
description	caractériser_SUBJ décrire_OBJ	12	2	93%
faire	faire_SUBJ	19	1	90%
utiliser	utiliser_OBJ	10	1	56%
concevoir	considérer_DEEPOBJ	9	1	100%
but	consister_SUBJ	6	1	100%
porter_sur	concerner_OBJ	6	1	86%
nommer	appeler_OBJ	6	1	100%

Tableau 4 : ULT verbales les plus productives comme introductrices de termes (nom recteur de T)

Comme dans le cas des noms, certaines classes de verbes apparaissent plus clairement introductrices de termes. La cinquième colonne du tableau 4 montre une nette augmentation de la proportion de cas pour lesquels l'hypothèse est vérifiée, excepté pour la classe *utiliser*. Cependant, la tâche de caractérisation est rendue plus difficile par une polysémie plus grande que pour les noms.

Les verbes sémantiquement quasi-vides (*être, avoir*) sont largement représentés, ce qui apparaît facilement explicable par les formules d'identification, souvent utilisées pour l'introduction de termes comme :

Leur seule entrée dans le langage [est]$_{ULT}$ donc le [langage oral]$_{T}$, tel que perçu…

Viennent ensuite les modaux (*devoir, sembler*), largement représentés, eux aussi, dont nous pensons qu'ils peuvent être considérés comme des auxiliaires modaux dans la tâche d'analyse syntaxique.

A côté de ces verbes *à tout faire*, fréquents dans la langue générale, on relève plusieurs classes synonymiques productives comme celle de l'identification (*identifier, distinguer, trouver*), de l'étude (*étudier, classer, considérer*[12]), de la description (*décrire, caractériser*).

Comme pour les noms, certaines classes synonymiques paraissent peu représentées : celles du *point de vue* ou des *liens logiques* (*cause, conséquence*).

L'examen des classes sémantiques introductrices de termes apparaît comme une piste prometteuse, certains champs paraissent clairement privilégiés comme introducteurs de termes (*processus et artefacts scientifiques, processus de description et d'étude, formules identificatoires*). Cela mérite un approfondissement sur un ensemble de données plus vaste, travail que nous souhaitons entreprendre en extrayant semi-automatiquement la combinatoire syntaxique et sémantique du lexique transdisciplinaire à partir de corpus arborés (sous-catégorisation syntaxique, co-occurrents lexico-syntaxiques) de façon à accélérer et faciliter le processus de catégorisation.

6. Conclusion et perspectives

L'expérience décrite dans cet article examine, sur un corpus de 42 articles scientifiques en sciences du langage, la manière et la proportion selon lesquelles les unités d'un lexique transdisciplinaire scientifique peuvent jouer un rôle de délimiteurs de termes. L'évaluation quantitative de cette hypothèse montre que celle-ci est vérifiée pour environ 74% des cas traités (*cf.* Tableau 2). Nous analysons manuellement 2 638 couples (T_i, ULT_j) parmi 3 406 couples qualifiés par une relation syntaxique de dépendance (NMOD, OBJ, SUBJ et leurs variantes. Sur le plan quantitatif, l'analyse manuelle des 2 638 cas pour lesquels il n'y a aucune erreur d'analyse syntaxique montre que l'hypothèse est majoritairement valide mais insuffisamment régulière pour utiliser en l'état les ULT pour filtrer automatiquement les candidats termes proposés par les extracteurs automatiques de termes, ACABIT et TERMOSTAT.

L'analyse qualitative des résultats les plus fréquents fait émerger une première piste

[12] Au sens de *prendre comme objet d'étude.*

d'amélioration en mettant en valeur plusieurs classes d'unités nominales et verbales pour lesquelles on constate un net accroissement de la proportion de cas satisfaisants (*cf.* Tableau 3 et Tableau 4, section 5).

Au-delà de l'expérience décrite, plusieurs pistes nous semblent intéressantes pour améliorer les résultats.

Tout d'abord, l'expérience menée dans le domaine des sciences du langage engendre un biais disciplinaire important : par exemple l'ambiguïté constatée pour le verbe *produire* mentionnée ci-dessus. Ce biais pourra être réduit par une extension de l'expérimentation à d'autres disciplines scientifiques et la prise en compte des spécificités de chacune d'elles avant de projeter le lexique transdisciplinaire.

Ensuite, l'extension de l'expérience ne sera que plus productive si on opère un affinage du typage sémantique du lexique et en particulier de la combinatoire des ULT.

De plus, il faut envisager de diversifier le repérage des relations syntaxiques. L'idée est de ne plus limiter l'expérience à des relations binaires mais de mettre en place une reconnaissance de patrons identificatoires (Jacques 2011) de la forme *un T [est]*$_{ULT}$ *un T qui […]*, par exemple.

Enfin, il sera nécessaire d'analyser les contextes où les termes apparaissent sans pour autant être introduits par des ULT.

Références

Aït-Mokhtar, S., Chanod, J-P. et Roux, C. (2002). Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(2-3), 121-144.

Bachimont B., Baneyx A, Malaisé V, Charlet J et Zweigenbaum P. (2005). Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles. *TIA*, Rouen.

Bolshakova E. (2008). Common Scientific Lexicon for Automatic Discourse Analysis of Scientific and Technical Texts. *Informational Theories and Applications*, v.15, 189-195.

Bourigault D, Aussenac-Gilles N et Charlet J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. M. Slodzian (ed*), Revue d'Intelligence Artificielle, Numéro spécial sur les techniques informatiques de structuration de terminologies*, Hermès, 18(1), 87-110.

Bourigault D, Jacquemin C et L'Homme MC. (2001). *Recent Advances in Computational Terminology*. Amsterdam-Philadelphie : John Benjamins

Bourigault, D. et Slodzian, M. (1999). Pour une terminologie textuelle. Terminologies Nouvelles, No 19, Bruxelles, Revue coéditée par l'Agence de la Francophonie et Communauté française de Belgique, 29-32.

Coxhead, A. (2000). A New Academic Word List. In : *TESOL Quarterly*, 34 (2), 213 - 238.

Daille B. (1996). ACABIT : une maquette d'aide à la construction automatique de banques terminologiques. in Clas A, Thoiron P et Béjoint H, (eds.), *Lexicomatique et Dictionnairique*, FMA, Beyrouth, 123-136.

Drouin P. (2003a). Term extraction using on-technical corpora as a point of leverage. *Terminology*, 9(1), 99-117

Drouin P. (2003b). Acquisition des termes simples fondée sur les pivots lexicaux spécialisés. *TIA*, Strasbourg.

Drouin P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue Française de linguistique appliquée*, 12(2), 45-64.

Flaux N. et Van de Velde D. (2000). *Les noms en français, esquisse de classement*, Paris, Ophrys.

Jacques, MP. (2011). Nous appelons « X cet Y » : X est-il un terme émergent ? *TIA*, Paris

Kister L., Jacquey E. (2012). Relations syntaxiques entre lexiques terminologique et transdisciplinaire : analyse en texte intégral. *Actes du Congrès Mondial de Linguistique Française*, Lyon, 909 – 919.

Paquot M. (2010). *Academic vocabulary in learner writing : From extraction to analysis.* Continuum.

Phal, A. (1971). Vocabulaire général d'orientation scientifique (V.G.O.S.) - Part du lexique commun dans l'expression scientifique. Paris, Didier.

Tutin A. (2007). Lexique et écrits scientifiques. *Revue française de linguistique appliquée*, XII(2).

Tutin A. (à paraître). La phraséologie transdisciplinaire des écrits scientifiques : des collocations aux routines sémantico-rhétoriques. *Les écrits scientifiques : du lexique au discours. Autour de Scientext.* Presses Universitaires de Rennes.

Enrichissement d'une ontologie de domaine par extension des relations taxonomiques à partir d'un corpus spécialisé

Olena OROBINSKA
Université de Lyon
ERIC Lyon 2
69676 Bron, France
Olena.Orobinska@univ-lyon2.fr

Jean-Hugues CHAUCHAT
Université de Lyon
ERIC Lyon 2
69676 Bron, France
Jean-Hugues.Chauchat@univ-lyon2.fr

Natalya CHARONOVA
KhNTU KhPI National University
Kharkov
61002, Ukraine
nvsharonova@mail.ru

Abstract

We propose a simple method to construct an ontology from a corpus and a first glossary. This method automatically discovers useful terminological patterns for a domain, and the role of the expert in the field is limited to the validation of the proposed candidate terms. We apply it to the field of radiation protection.

1 Introduction

On cherche souvent à construire l'ontologie d'un nouveau domaine d'activité ; il est donc utile de disposer d'un procédé rapide et efficace. Dans cet article, nous proposons une méthode simple pour construire une ontologie à partir d'un corpus et d'un premier glossaire ; cette méthode découvre automatiquement les patrons terminologiques utiles pour un domaine. Cela permet de limiter le rôle de l'expert du domaine à la validation des candidats-termes proposés ; nous l'appliquons au domaine de la Radioprotection. Le cadre général a été présenté en langue russe dans (Orobinska, 2012).

Les auteurs partagent l'opinion de Simperl & al (2009) : l'ingénierie des ontologies a atteint sa maturité en utilisant la diversité des méthodologies, approches et techniques. Mais ce processus reste très laborieux et coûteux et il faut chercher des procédés rapides et efficaces, capables de dégager automatiquement des éléments de connaissances dans les ressources textuelles.

Globalement, une ontologie est une structure sémantique qui encode les concepts, les relations et les axiomes définissant le modèle d'un domaine donné (De Nicola & al. 2009; Gruber 1995 ; Neches & al. 1991). A la différence de la lexicologie, où les notions de « concept » et de « terme » sont bien distinctes[1], ici, pour la construction d'ontologies, ces deux mots seront synonymes ; un « concept » est un élément constructif d'une ontologie. Les termes correctement détectés dans le corpus vont être implantés dans notre ontologie, soit comme des classes ou sous-classes, soit comme de nouveaux individus des classes ou sous-classes.

La section 2 introduit l'état de l'art sur les techniques d'extraction de termes et de relations à partir de ressources textuelles ; nous présentons notre approche d'extraction des nouveaux candidats-termes dans la section 3 ; la section 4 détaille l'implémentation de notre méthode d'enrichissement de la taxonomie pour chaque concept d'une « ontologie plate » ; la section 5 présente les résultats des expérimentations et la section 6 souligne des problèmes à résoudre et introduit quelques perspectives.

2 Etat de l'art et travaux liés

Pour enrichir une ontologie à partir de textes, on distingue deux catégories de méthodes : les méthodes purement statistiques et celles qui se basent sur le Traitement Automatique de la Langue. L'approche linguistique permet de formuler des règles d'inférences ou de trouver les heuristiques. La fusion de ces approches a engendré des méthodes hybrides. L'exploitation des connaissances linguistiques est devenue possible grâce à la mise au point et à la diffusion d'outils et ressources spécialisés : analyseurs grammaticaux et dictionnaires électroniques. Pour les dictionnaires, la communauté de Text-Mining se sert surtout d'outils comme WordNet, mais beaucoup d'autres ressources utiles sont

[1] ISO 704 (1987) définit les concepts comme « des constructions mentales qui servent à classer les objets individuels du monde extérieur ou intérieur à l'aide d'une abstraction plus ou moins arbitraire », tandis que les termes verbalisent des concepts.

Olena Orobinska,Jean-Hugues Chauchat,Natalya Charonova

accessibles sur la Toile[2]. A l'avenir, l'ingénierie des ontologies utilisera plus largement l'ensemble de ces outils.

Depuis la première application de patrons grammaticaux pour extraire les mots liés par les relations taxonomiques de grand corpus par Hearst (1992), un grand nombre de chercheurs travaille dans cette direction. L'utilisation de patrons syntaxiques permet d'obtenir des « unités de connaissances » pertinentes mais leur élaboration reste laborieuse.

De nombreuses expérimentations ont été effectuées ; nous citons ceux qui ont particulièrement inspiré nos propres expérimentations. Une longue liste de patrons pour extraire des compétences dans le domaine du management est proposée par Buitelaar & Eigner (2008) ; cependant, ces patrons semblent définis a priori par des experts linguistes et non issus automatiquement du corpus.

Des résultats intéressants sur l'instanciation semi-automatique d'ontologie et sur l'extraction de relations non taxonomiques sont présentés dans (Makki & al. 2009) pour le domaine de « risk management ». Mustière & al. (2011) proposent des techniques automatiques de traitement du langage et d'alignement d'ontologies.

N. Aussenac-Gilles (2006) et ses collègues travaillent depuis 20 ans sur la modélisation terminologique à base de textes. Leurs résultats sont rassemblés dans la plate-forme TERMINAE[3]. Depuis 2009, cette équipe développe DAFOE, une plateforme multi modèles et multi méthodes pour construire des ontologies à partir de corpus (Charlet & al, 2009).

La méthode se basant sur test statistique qui vise les items spécifiques au corpus technique est proposée par P. Douin (2003) et elle est réalisée dans le projet TermoStat[4].

L'analyse des concepts formels (Cimiano & al, 2005) permet de dégager à la fois les concepts et relations qui les lient. Sánchez (2010) présente les recherches menées sur l'utilisation des ressources web où les textes sont plus structurés et contiennent déjà des métadonnées qui facilitent l'extraction de connaissances.

Par ailleurs, en langue russe, Zolotova G. (2011) a produit un travail sur la généralisation des structures syntaxiques minimales qui permettent de distinguer le sens de phrases.

Nous développons nos méthodes pour l'élaboration, l'installation et l'enrichissement de l'ontologie de la radioprotection à la demande du groupement de recherche et production « Métrologie » de Kharkov, Ukraine. Cette organisation, membre de l'AIEA, souhaite disposer d'une base de connaissances en anglais, français et russe. En collaboration avec ces spécialistes, nous avons construit un modèle du domaine (Fig.1).

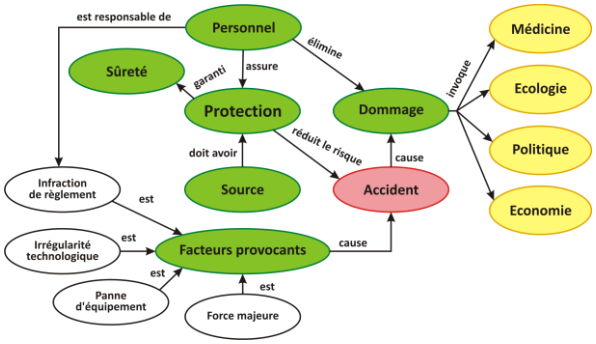


Fig .1. Modèle du domaine de radioprotection pour l'installation de l'ontologie du domaine

3 Notre approche pour l'enrichissement de l'ontologie de radioprotection

3.1 Bases théoriques

Nous posons comme hypothèse que les instances des concepts et des relations qui existent dans l'ensemble des textes publiés dans le domaine portent la sémantique du domaine à travers les règles linguistiques qui les réunissent. En disposant d'une liste de concepts génériques et en découvrant empiriquement les règles de formation de leurs attributs à partir d'un corpus, on peut élargir l'ontologie sous-jacente par des termes nuancés, formant la taxonomie (Fig. 2).

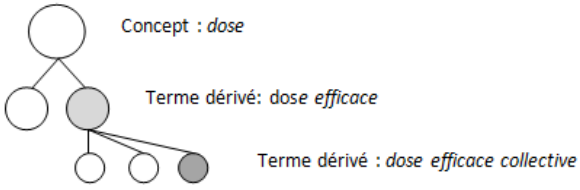


Fig. 2. Installation de hiérarchie de concepts à partir de « patrons directs »

Exemple :

- *dose efficace* : somme des doses équivalentes dans tous les tissus et organes spécifiés du corps ;
- *dose efficace collective :* somme de toutes les doses efficaces individuelles pendant la pé-

[2] http://www.mkbergman.com/sweet-tools-simple-list/ et, pour le français, http://portail-du-fle.info/

[3] http://lipn.univ-paris13.fr/terminae/index.php/Main_Page

[4] http://olst.ling.umontreal.ca/~drouinp/termostat_web/

riode de temps ou pendant l'opération considérée.

Définition : une ontologie de noyau est une structure $O := (C, \leq_C, A)$ qui se compose de : (i) l'ensemble *C* des identifiants de concepts ; (ii) la relation de subsomption entre les concepts qui est transitive, réflexive et antisymétrique (ordre partiel) ; (iii) l'ensemble *A* des attributs des concepts. De telles structures sont nommées « hiérarchies des concepts » ou « taxonomies » (Cimiano, 2005 ; Ganter, 1999).

Nous allons montrer comment enrichir une ontologie de noyau par l'extension de termes génériques en utilisant des *patrons terminologiques*. La combinaison de mots créant de nouveaux termes suit des règles syntaxiques que nous allons découvrir dans le corpus. Les termes sont formés par une combinaison syntaxique hiérarchisée de mots, appelés *syntagmes terminologiques* ou *synapsies* (Cabré, 1998). L'application de ces syntagmes terminologiques permet l'extraction de termes à partir de leur forme syntaxique.

3.2 Formation des patrons terminologiques et installation de taxinomies partielles

Nos patrons terminologiques sont formés en deux étapes à partir des fréquences de structures syntaxiques dans le corpus, celui-ci étant préalablement étiqueté, tagué, par les balises correspondantes aux « parties de discours », puis à partir de l'analyse syntaxique des termes du glossaire de domaine.

Les fragments de phrases qui correspondent aux patrons sont extraits du corpus, puis validés. Par construction, tous les fragments extraits contiennent des termes dérivés de concepts de l'ontologie sous-jacente. Autrement dit, chaque terme nouveau contient un des concepts en tant que radical.

Après validation, les termes dérivés qui ont la même racine forment la taxonomie partielle (§ 5). Ils sont rajoutés dans l'ontologie en tant que sous-classes de concepts correspondants.

La chaîne de traitements d'extraction de termes, présentée Fig. 3, a été programmée en Java.

Dans cette chaîne, le module *Prétraitement* effectue la conversion des fichiers PDF en format textuel puis le nettoyage des textes obtenus en éliminant les fragments contenant des caractères autres que des lettres, chiffres ou signes de ponctuation ; le module *Etiquetage* produit l'étiquetetage des textes par TreeTagger ; le module *Formation des patrons* recense tous les patrons terminologiques ; le module *Extraction* permet de récupérer les fragments qui correspondent aux patrons et il forme les taxonomies pour chaque racine. Ensuite, les résultats sont présentés aux experts du domaine et rajoutés à l'ontologie initiale s'ils sont validés.

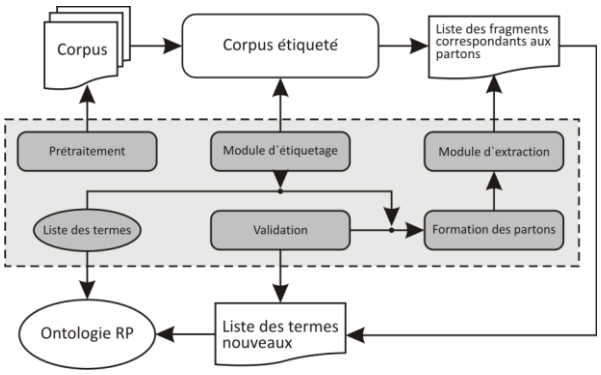


Fig. 3. Schéma du système d'enrichissement de l'ontologie du domaine par la terminologie dérivée de la liste des termes génériques (concepts)

4 Implémentation de la méthode

4.1 Installation du corpus

Nous avons constitué un corpus de textes français abordant la problématique du domaine à partir de documents officiels accessibles en ligne, notamment les normes de sûreté et les rapports de l'IAEA et de la Commission internationale de protection radiologique, et à partir d'articles de revues spécialisées publiés depuis 1999 ; ce corpus (texte brut) contient 1 500 000 mots.

La plupart des textes étant initialement en format PDF, on a dû les convertir dans un format (texte ou HTML) utilisable pour le traitement.

4.2 Etiquetage des textes

A l'étape suivante les textes ont été étiquetés avec TreeTagger (Schmid, 1994) de façon à obtenir des fichiers où les textes sont transformés sous la forme « …pos/lemme pos/lemme pos/lemme…» où :

- pos – correspond à une certaine *partie du discours* reconnue par TreeTagger. Dans cet article, nous utilisons le mot « *balise* » pour désigner une telle *partie du discours* ;
- lemme – présente la forme normale d'un mot : l'infinitif pour les verbes ; le masculin-singulier pour les noms et adjectifs, etc.

Par exemple, pour le texte initial :
«*…détriment causé au personnel médical est pris en compte de façon subsidiaire*.»,
on obtient la structure étiquetée :
« NOM/*détriment* PPP/*causer* PRP/*au* NOM/*personnel* ADJ/*médical* VER/*être* PPP/*prendre* PRP/*en* NOM/*compte* PRP/*de* NOM/*façon* ADJ/*subsidiaire* SENT/. »

TreeTagger distingue 33 différentes formes grammaticales pour le français, ce qui est trop détaillé pour nos objectifs ; pour générer les patrons grammaticaux nous avons réduit à 17 le nombre de ces formes (Tab. 1).

Notons que nous avons conservé des balises telles que les articles, les propositions et les pronoms, qui sont souvent négligées par d'autres auteurs.

4.3 Installation de la liste des patrons terminologiques

Un patron terminologique est une chaîne de balises que l'analyseur automatique (TreeTagger) est capable de distinguer. Par exemple, « NOM/PRP/NOM » est le patron pour tout fragment qui se compose d'un nom suivi par une préposition suivie par un nom.

Balise	Descripteur grammatical
ABR	abréviation
ADJ	adjectif
ADV	adverbe
ART	articles
DET	pronom possessif
KON	conjonction
NAM	nom propre
NOM	nom
NUM	nombre ou tout chiffre
PPP	participe passé
PPR	participe présent
PRO	tous les autres pronoms
PRP	toutes les prépositions (liés aux articles inclus)
PUN	toute ponctuation autre que point
SENT	point
SYM	symbole
VER	toutes les formes des verbes

Tab.1. Notre liste réduite de 17 balises de parties de discours obtenues à partir de TreeTagger

Nos patrons terminologiques sont des N-grams des balises grammaticales qui ont remplacés les mots dans le corpus ; nous avons utilisé des grams de taille (N) variant de 2 à 6. Nous avons extrait du corpus tous les fragments de phrases correspondants à ces N-grams.

La sélection des patrons potentiellement pertinents a été faite à partir des termes génériques (= concepts) de l'ontologie sous-jacente. La liste de ces termes génériques, contenant 38 noms, a été installée après consultation d'un expert du domaine ; ces termes ont été utilisés pour créer les concepts d'ontologie sous-jacente. Ci-après nous utiliserons le mot « concept » pour ces termes-là.

On a retenu les patrons pour lesquels au moins 70% des phrases correspondantes contiennent un tel concept ; ce seuil a été choisi expérimentalement. Les scores de cette partie de l'expérimentation sont présentés dans la table 2.

Nous avons enrichi notre liste de patrons à l'aide du glossaire proposé en 2007 par la Commission Internationale de Protection Radiologique. Ce glossaire compte 162 termes. La structure syntaxique des certains termes est assez évoluée ce qui permet d'élargir la liste des patrons.

Les termes contiennent en moyenne 2,8 mots (entre 1 et 10 mots et le plus souvent entre 2 et 5) ; la table 4 en donne des exemples.

N	Patron	% de concepts dans le patron
1	*NOM+ADJ*	*100%*
2	*NOM+PRP+NOM*	*100%*
3	*NOM+PRP+NOM+ADJ*	*92%*
4	*NOM+PRP+NOM+PRP+NOM*	*92%*
5	*NOM+ADJ+PRP+NOM*	*92%*
6	*NOM+ADJ+ADJ*	*84%*
7	*NOM+PPP*	*84%*
8	*NOM+PRP+NOM+PPP*	*76%*
9	*NOM+PRP+NOM+PRP+ART+NOM*	*74%*
10	*NOM+ADJ+PPP*	*71%*
11	*NOM+PRP+ART+NOM+PRP+NOM*	*71%*

Tab.2. Score de validations des patrons terminologiques formés automatiquement

Nous avons ajouté 7 nouveaux patrons terminologiques (Tab. 3) issus de l'analyse du glossaire ; leur longueur peut éventuellement dépasser 6.

N	Patron	% des concepts dans le patron
1	*NOM+PRP+NOM+PRP+NOM+ADJ*	58%
2	*NOM+ADJ+ADJ+PRP+NOM*	*53%*
3	*NOM+ADJ+PRP+NOM+PRP+ART+NOM où : NOM+ADJ+ PRP+NOM+PRP+NOM*	*39%*
4	*NOM+PRP+NOM+PRP+ART+NOM + ADJ*	39%
5	*NOM+PPP+ADJ*	*37%*
6	*NOM+PRP+NOM+PRP+ART+NOM + PRP+NOM*	*37%*
7	*NOM+PRP+NOM+PRP+NOM +NAM*	18%

Tab.3. Patrons terminologiques formés à partir du glossaire

N	Patron	Equivalent d'une phrase
1	*NOM+ADJ*	*zone urbaine*
2	*NOM+ADJ+ADJ*	*accident nucléaire grave*
3	*NOM+ADJ+ADJ+PRP+NOM*	*dose efficace professionnelle par an*
4	*NOM+ADJ+PPP*	*brûlure radiologique étendue*
5	*NOM+ADJ+PRP+NOM*	*effets stochastiques des rayonnements*
6	*NOM+ADJ+PRP+NOM+PRP +ART+NOM où : NOM+ADJ+ PRP+NOM+PRP+NOM*	*effet tardif du rayonnement dans le tissu*
7	*NOM+PPP*	*zone surveillée*
8	*NOM+PPP+ADJ*	*dose absorbée individuelle*
9	*NOM+PRP+ART+NOM+PRP +NOM*	*sûreté sur la gestion du déchet*
10	*NOM+PRP+NOM*	*réacteur à eau*
11	*NOM+PRP+NOM+ADJ*	*entreposage du déchet radioactif*
12	*NOM+PRP+NOM+PPP*	*débit de dose absorbée*
13	*NOM+PRP+NOM+PRP+ART +NOM*	*action de prévention de la pollution*
14	*NOM+PRP+NOM+PRP+ART +NOM+ADJ*	*coefficient de risque pour l'effet nocif*
15	*NOM+PRP+NOM+PRP+ART +NOM+PRP+NOM*	*seuil de dose pour le risque de mortalité*
16	*NOM+PRP+NOM+PRP+NO M*	*durée de vie du réacteur*
17	*NOM+PRP+NOM+PRP+NO M+ADJ*	*site de stockage des éléments combustibles*
18	*NOM+PRP+NOM+PRP+NO M+NAM*	*fonctionnement du réacteur de type BWR*

Tab.4. Liste finale des patrons terminologiques et exemples de termes correspondant

La liste finale, avec des exemples des termes correspondant à chaque patron terminologique, est présentée dans la table 4.

5 Résultats des expérimentations

Une intervention humaine est nécessaire pour la validation définitive des « candidats-termes » détectés dans le corpus à partir des patrons terminologiques. Les scores d'évaluation de ces candidats-termes sont donnés dans le tableau 5 où on lit les informations suivantes : la colonne **Patron** contient la liste des patrons terminologiques; la colonne **Total** contient le nombre des fragments détectés dans le corpus, qui correspondent aux patrons; la colonne **Taux des fragments contenant un concept** fournit le nombre de fragments où entre au moins un des concepts de l'ontologie de départ; la colonne **Taux des termes dérivés** donne le nombre de termes du domaine parmi les fragments précédents.

Par exemple, ligne 1 du tableau 5: dans le corpus, on rencontre 14 759 expressions différentes contenant le patron « *NOM + ADJECTIF* » ; parmi elles, 1 296 contiennent un mot de la liste initiale, soit 8,8 % (0,88=1296/14 759). Parmi ces 1 296, 558 expressions sont acceptées par l'expert comme des termes qui compléteront l'ontologie, soit 43 % (0,43 = 558/1 296).

Notons que la pertinence d'un patron terminologique dépend de sa taille (nombre de *parties du discours* formant le patron terminologique) ; les patrons longs sont en moyenne plus pertinents.

Nous allons utiliser les définitions suivantes :

- *taxonomie partielle* : taxonomie de chaque concept de l'ontologie, formée par ses descendants (termes dérivées), ainsi chaque concept est le « sommet » de sa taxonomie partielle ;

- *racine du patron terminologique* : la structure minimale linguistique à laquelle un terme du domaine peut correspondre. Nous distinguons trois types des racines : NOM+ADJ, NOM+PPP et NOM+PRP+NOM ; l'ensemble des termes correspondants à une racine forment le niveau I de la taxonomie d'un concept ;

- *terme-descendant*, ou *descendant* : le terme dérivé, formé à la base d'une racine; les descendants forment les niveaux II et III de chaque taxinomie partielle.

Les patrons terminologiques réunis autour de chaque racine permettent de former les taxonomies à trois niveaux (sans compter les niveaux des concepts en tant que tels). Ceci est illustré dans les figures 4-a, 4-b, 4-c.

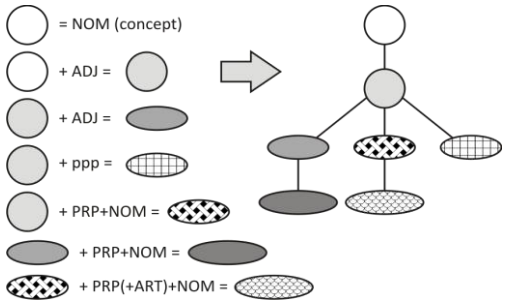


Fig. 4-a. Taxonomie avec la «racine» NOM+ADJ

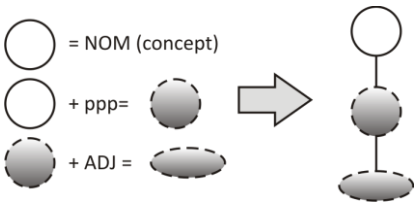


Fig. 4-b. Taxonomie avec la «racine» NOM+PP

N	Patron	Nb total d'expressions différentes	Taux de fragments contenant un concept		Taux de termes dérivés acceptés par l'expert du domaine	
1	*NOM+ADJ*	14759	1296	8,8 %	558	43 %
2	*NOM+ADJ+ADJ*	1998	356	17,8 %	160	45 %
3	*NOM+ADJ+ADJ+PRP+NOM*	179	34	19,0 %	7	21 %
4	*NOM+ADJ+PPP*	1434	299	20,8 %	30	10 %
5	*NOM+ADJ+PRP+NOM*	3556	425	11,9 %	204	48 %
6	*NOM+ADJ+PRP+NOM+PRP+ART+NOM où : NOM+ADJ+ PRP+NOM+PRP+NOM*	119	35	29,4 %	25	71 %
7	*NOM+PPP*	5294	582	11,0 %	58	10 %
8	*NOM+PPP+ADJ*	187	46	24,6 %	14	30 %
9	*NOM+PRP+ART+NOM+PRP+NOM*	1898	151	8,0 %	30	20 %
10	*NOM+PRP+NOM*	16201	1134	7,0%	590	52 %
11	*NOM+PRP+NOM+ADJ*	5344	160	3,0 %	80	50 %
12	*NOM+PRP+NOM+PPP*	1850	166	9,0 %	17	10 %
13	*NOM+PRP+NOM+PRP+ART+NOM*	1670	200	12,0 %	81	40 %
14	*NOM+PRP+NOM+PRP+ART+NOM+ADJ*	219	56	25,6 %	39	70 %
15	*NOM+PRP+NOM+PRP+ART+NOM+PRP +NOM*	192	53	27,6 %	27	50 %
16	*NOM+PRP+NOM+PRP+NOM*	4382	219	5,0 %	88	40 %
17	*NOM+PRP+NOM+PRP+NOM+ADJ*	868	95	10,9 %	57	60 %
18	*NOM+PRP+NOM+PRP+NOM+NAM*	75	14	18,7 %	4	30 %
	Total	***60 225***				

Tab.5. Scores des patrons formé à partir de la première liste des termes

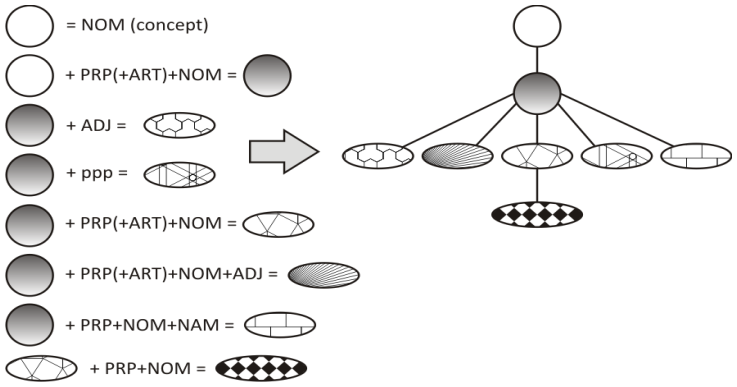


Fig. 4-c Taxonomie avec la «racine» NOM+PRP+NOM (ou NOM+PRP+ART+NOM)

Les résultats obtenus après installation des taxonomies partielles sont présentés dans le tableau 6.

Type de racine / Niveau de taxonomie	*NOM +ADJ*	*NOM +PPP*	*NOM +PRP (+ART) +NOM*
I	*558*	*58*	*620 (*)*
I + II	*238*	*13*	*380*
I + II +III	*7*	-	*4*

Tab. 6. Recapitulatif d'inductions des taxonmies partielles à partir de trois racines

(*) Parmi ces 620 termes, 590 termes correspondent au patron NOM+PRP+ NOM et 30 au patron NOM+PRP +ART + NOM

Des exemples de taxonomies, obtenus pour chaque type de racine, sont présentés dans le tableau 7.

Niveau de taxonomie	*I*	*I + II*	*I + II +III*
NOM+ADJ	*effet néfaste*	*effet néfaste d'exposition*	*effet néfaste d'exposition pour la santé*
NOM+PPP	*dose absorbée*	*dose absorbée individuelle*	-
NOM+PRP (+ART) + NOM	*action de prévention*	*action de prévention de la pollution*	*action de prévention de la pollution par le rejet radioactif*

Tab. 7. Exemples de taxonomies de termes derivés

Notons deux faits :

1) les ensembles de termes retenus aux différents niveaux ne sont pas forcément emboîtés ; exemple « *exposition continue* » (NOM+ADJ) n'est pas retenu au niveau I, car non spécifique du domaine, mais « *exposition continue au rayonnement* » (NOM+ADJ+PRP+NOM) est retenu au niveau II pour enrichir l'ontologie ;

2) le nombre des termes directement descendant du niveau I vers les niveaux II et III est inférieur au nombre des termes correspondant à chaque patron car chaque terme du niveau supérieur n'a pas nécessairement de terme-fils.

6 Conclusion

Nous avons montré qu'à partir d'un corpus du domaine, on peut enrichir une ontologie « plate » (constituée d'une première liste de mots) en recherchant les patrons terminologiques des phrases qui contiennent ces mots dans le corpus, puis, inversement, les termes plus ou moins complexes contenus dans les phrases construites sur ces patrons.

Nous avons construit un grand corpus original du domaine de la radioprotection (environ 1 500 000 mots) et nous y avons appliqué notre méthode.

Nous avons produit des « candidats-termes » qui peuvent être soumis à un expert pour validation.

La terminologie d'un domaine de haute-technologie, tel la radioprotection, ne se limite pas à une liste de mots isolés, et la formation des termes complexes suit des règles syntaxiques propres à chaque langue et à chaque domaine. Nous avons montré que la probabilité qu'une structure syntaxique fréquente dans le corpus spécialisé corresponde à un terme intéressant est assez élevée, jusqu'à 70%. Cela nous a permis de travailler sur les patrons terminologiques et de proposer une méthode de détection de ces patrons à partir des fréquences de chaque structure syntaxique dans le corpus.

La liste des patrons a été enrichie à partir d'un glossaire du domaine.

Parmi les difficultés rencontrées, citons la constitution du corpus et la transformation de textes du format PDF vers le format TXT ; nous avons créé un module qui réalise cela automatiquement. La faible capacité de TreeTagger à distinguer certaines formes grammaticales du français, par exemple adjectif et participe passé, crée du bruit qui perturbe toute la chaîne de traitement ; un perfectionnement des analyseurs linguistiques améliorerait nos résultats finaux

Dans le futur, nous envisageons d'étudier une généralisation des règles de construction des patrons terminologiques à partir d'analyse syntaxique.

References

Aussenac-Gilles Nathalie. 2006. Méthodes ascendantes pour l'ingénierie des connaissances : Synthèse des travaux. Institut de recherches en informatique de Toulouse, p.226.

Buitelaar Paul, Eigner Thomas. 2008. Topic Extraction from Scientific Literature for Competency Management. In Proceedings of the 3rd Expert Finder Workshop on Personal Identification and Collaborations: Knowledge Mediation and Extraction, 27th October 2008, Karlsruhe, Germany.

Cabré, Maria Teresa. 1998. La terminologie. Théorie, méthode et applications, Les Presses de l'Université d'Ottawa.

Charlet J. et al. Apport des outils de TAL à la construction d'ontologies : propositions au sein de la plateforme DaFOE, IC 2009.

Cimiano Philipp, Hotho Andreas and Staab Steffen. 2005. Learning concept hierarchies from text corpora using formal concept analysis. J. Artif. Int. Res. 24(1):305-339.

De Nicola Antonio, Missikoff Michele and Navigli Roberto. 2009. A software engineering approach to ontology building. *Inf. Syst.* 34(2):258-275.

Drouin, Patrick. 2003. "Term extraction using non-technical corpora as a point of leverage", In Terminology, vol. 9, no 1, p. 99-117.

Ganter Bernhard and Wille Rudolf. 1999. Formal concept analysis – mathematical foundations. Springer, ISBN 978-3-540-62771-5.

Gruber Thomas R. 1995. Towards principles for the design of ontologies used for knowledge sharing Int. J. Hum.-Comput. Stud. 43(5-6):907-928.

Hearst Marti A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceeding of the Fourteenth International Conference on Computational Linguistics, Nantes, France.

Makki Jawad, Alquier Anne-Marie and Violaine Prince. 2009. Ontology Population via NLP Techniques in Risk Management. International Journal of Humanities & Social Sciences;2009, Vol. 3 Issue 3, p. 212.

Mustière Sébastien, Abadie Nathalie, Aussenac-Gilles Nathalie, Bessagnet Marie-Noëlle, Kamel Mouna, Kergosien Eric, Reynaud Chantal, Safar Brigitte, Sallaberry Christian. 2011. Analyses linguistiques et techniques d'alignement pour créer et enrichir une ontologie topographique. Revue Internationale de géomantique, 21(2) :155-179.

Neches Robert, Fikes Richard, Finin Tim, Gruber Tom, Patil Ramesh, Senator Ted, and Swartout William R. 1991. Enabling technology for knowledge sharing. *AI Mag.* 12(3):36-56.

Orobinska Olena A. 2012. Automatic Method of Domain Ontology Construction based on Corpora Characteristics POS-Analysis. In Proceedings of the XV International Conference IMS-2012 , St-Petersburg.

Sánchez David. 2010. A methodology to learn ontological attributes from the Web. *Data Knowl. Eng.* 69, 6 (June 2010), 573-597.

Schmid Helmut. 1994. Probabilistic Part-of-Speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, pages 44–49.

Simperl Elena, Mochol Malgorzata, Bürger Tobias and Igor O. Popov. 2009. Achieving Maturity: The State of Practice in Ontology Engineering in 2009. In *Proceedings of the Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009 on On the Move to Meaningful Internet Systems: Part II* (OTM '09), Robert Meersman, Tharam Dillon, and Pilar Herrero (Eds.). Springer-Verlag, Berlin, Heidelberg, 983-991.

Zolotova G. Syntactic Dictionary. The repertoire of elementary units of Russian syntax. Едиториал УРСС, 2011, P .356.

Une typologie multi-dimensionnelle des structures énumératives pour l'identification des relations termino-ontologiques

Jean-Philippe Fauconnier[1] Mouna Kamel[1] Bernard Rothenburger[1]

[1] Institut de Recherche en Informatique de Toulouse (IRIT)
Université Paul Sabatier, 118 Route de Narbonne, 31060 Toulouse Cedex 5
{prénom}.{nom}@irit.fr

Résumé

Ce travail s'inscrit dans le cadre de la construction de ressources termino-ontologiques. Il vise à améliorer l'extraction des relations sémantiques en exploitant les structures énumératives contenues dans les textes. Nous proposons ici une typologie multi-dimensionnelle de ces structures énumératives, selon les axes visuel, rhétorique, intentionnel et sémantique. Cette typologie intervient dans le cadre d'une campagne d'annotation outillée par LARAt (Logiciel d'Acquisition de Relations par l'Annotation de textes), pour l'identification de relations par apprentissage supervisé.

1 Introduction

La structure énumérative (dorénavant appelée SE) est une structure textuelle ayant la propriété d'exprimer des connaissances hiérarchiques au travers de différents composants. Elle présente, au sein d'un même objet textuel, un thème énumératif, dit *énumérathème*, justifiant la réunion de plusieurs éléments en fonction d'une identité de statut (Ho-Dac et al., 2010). Sur le plan sémantique elle forme un tout. Sur le plan de la mise en forme, elle peut être exprimée selon différents modes, allant d'une forme linéaire discursive à une forme visuelle usant de dispositifs typo-dispositionnels. Ces propriétés autorisent son apparition dans tout type de texte, lui permettant par là même de rendre compte de connaissances de nature différente.

Elle a ainsi fait l'objet de nombreuses études au cours desquelles différentes typologies ont pu être proposées. Les SE linéaires ont été essentiellement analysées dans le cadre de l'analyse du discours. Elles ont d'abord donné lieu à des typologies comme celle de (Vergez-Couret et al., 2008) où les SE à un temps ont été opposées aux SE à deux temps, ou encore comme celle de (Ho-Dac et al., 2010) où les SE ont été classifiées selon leur niveau de granularité (SE dont les items sont des titres, SE en tant que listes formatées, SE multi-paragraphiques sans marque visuelle, SE intra-paragraphiques). Les SE usant de dispositifs typo-dispositionnels, dites verticales, ont quant à elles été notamment analysées dans le cadre de la génération de texte. Hovy et Arens (1991) distinguent les listes d'items (ensemble de composants de même niveau), des listes énumérées (pour lesquelles l'ordre des composants est pris en compte), alors que Luc (2001) propose une typologie qui oppose les SE parallèles aux SE non parallèles. Cette dernière typologie est basée sur la composition du modèle rhétorique de la RST[1] (Mann and Thompson, 1988) et du MAT[2] de Virbel (1989).

À notre connaissance, les SE n'ont pas été exploitées pour l'extraction de relations sémantiques à partir de textes. Or ces SE sont très fréquentes dans les textes scientifiques ou encyclopédiques qui sont justement appropriés pour la construction de ressources sémantiques. Les méthodes classiques d'extraction des relations sont le plus souvent limitées à l'identification de relations binaires intra-phrastiques, après analyse du texte rédigé par des patrons lexico-

[1] RST : Rhetorical Structure Theory
[2] MAT : Modèle d'Architecture Textuelle

syntaxiques (Hearst, 1992; Montiel-Ponsoda and de Cea, 2011; Aussenac-Gilles and Jacques, 2008), des techniques de clusterisation ou des algorithmes d'apprentissage automatique (essentiellement non supervisé) (Buitelaar et al., 2005; Poelmans et al., 2010). L'exploitation des SE apparait alors comme un moyen d'élargir les méthodes classiques d'extraction de relations pour la construction ou l'enrichissement de ressources sémantiques telles que les ontologies, les Ressources Termino-Ontologiques (RTO), les thesaurus, etc.

Cet article propose une typologie multi-dimensionnelle qui permettra de cibler puis d'exploiter automatiquement les SE porteuses de relations termino-ontologiques. Cette typologie caractérise les SE selon les axes visuel et rhétorique à l'instar de (Luc, 2001), mais également selon les axes intentionnel et sémantique. C'est cette typologie que nous présentons en section 3, après avoir rappelé en section 2 quelques définitions et propriétés des SE. Vu l'inadéquation des outils classiques d'extraction de relations pour ce genre de structure textuelle, nous envisageons une approche alternative, à base d'apprentissage supervisé, nécessitant une campagne d'annotation basée sur cette typologie. La section 4 montre comment cette typologie intervient dans le cadre du processus d'annotation, et décrit sommairement l'outil d'annotation développé pour ces besoins. Nous concluons et présentons nos perspectives en section 5.

2 SE : définitions et propriétés

Comme indiqué précédemment, l'acte d'énumération consiste à énoncer les éléments successifs d'un même champ conceptuel, ces éléments entretenant un lien hiérarchique direct ou indirect avec un concept classifieur. La forme générale d'une SE est alors caractérisée par la présence d'une *amorce* (phrase contenant l'énumérathème et introduisant l'énumération), d'une *énumération* composée d'au moins deux *items* (appartenant au même champ conceptuel), et éventuellement d'une *clôture* (ou conclusion).

D'un point de vue visuel, la SE a la propriété de pouvoir être formulée de diverses façons. Elle peut être énoncée discursivement en dehors de toute MFM, au sein de la même phrase ou à travers plusieurs phrases n'appartenant pas nécessairement au même paragraphe. Elle peut également être mise en évidence par l'usage de marqueurs typographiques et/ou dispositionnels, marqueurs qui pallient alors les marqueurs lexicaux. Ces marqueurs sont de l'ordre de la métalangue (Harris, 1976; Porhiel, 2007) et permettent alors d'organiser des segments de texte successifs non forcément contigus.

Différentes définitions de la SE existent, dont celle de Pascual pour qui "énumérer, c'est conférer une égalité d'importance à un ensemble d'objets, et ensuite c'est ordonner ces objets selon des critères variés" (Pascual, 1991). Ces objets sont considérés comme visuellement et fonctionnellement équivalents. On parle alors de SE parallèles.

D'un point de vue rhétorique, l'analyse des SE montre qu'il existe des relations de discours entre les différents composants. La définition de Pascual citée ci-dessus correspond au cas où ces relations montrent une égalité d'importance entre les items. Or des études de corpus ont montré que les SE ne présentent pas toutes cette équivalence visuelle et fonctionnelle entre items (Luc, 2001).

Dans un souci de généralisation, nous préférons la définition proposée par (Virbel, 1999) qui nous semble mieux prendre en compte à la fois les phénomènes architecturaux du texte et l'intention de l'auteur : "l'acte textuel consiste à transposer textuellement la coénumérabilité des entités recensées par la coénumarabilité des segments linguistiques qui les décrivent, ceux-ci devenant par le fait les entités constitutives de l'énumération (les items)."

D'un point de vue intentionnel, à l'image des textes qui peuvent être de différents types (narratifs, procéduraux, descriptifs, etc.), les SE reflètent l'intention de l'auteur. Nous proposons de reprendre cette typologie des textes pour caractériser l'intention de l'auteur lorsqu'il rédige une SE.

Enfin, d'un point de vue sémantique, les SE peuvent exprimer des connaissances de nature différente. Ces connaissances peuvent décrire de

façon consensuelle ou conjoncturelle le monde réel ou imaginaire, la langue, les émotions, les sentiments, les opinions, etc.

3 Typologie de la SE

La typologie que nous proposons est basée sur les différentes propriétés décrites ci-dessus. Elle s'appuie sur les dimensions visuelle, rhétorique, intentionnelle et sémantique, l'objectif étant à terme de repérer et d'exploiter les SE paradigmatiques bénéficiant de mise en forme et véhiculant des connaissances propices à la construction de ressources sémantiques.

Les différentes caractéristiques observées au sein de chacune des dimensions sont illustrées par des exemples extraits du corpus de Virbel (1999) et d'un corpus composé de pages Wikipédia, ce deuxième corpus ayant été élaboré dans le but d'enrichir l'ontologie OntoTopo construite lors du projet GEONTO[3] (Kamel and Rothenburger, 2011).

3.1 Typologie selon l'axe visuel

Les types définis dans cet axe ont pour but d'aider au repérage des SE. Nous distinguons la **SE horizontale** qui peut bénéficier ou non de mise en forme typographique, de la **SE verticale** qui bénéficie de mise en forme typographique et dispositionnelle.

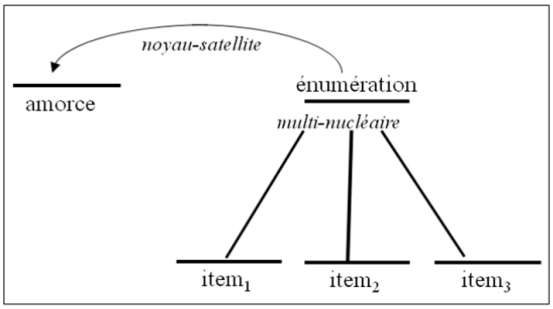
La **SE horizontale** s'inscrit dans la linéarité du texte et ne fait pas usage du "dispositionnel". Elle est caractérisée soit par des MIL[4] comme "premièrement", "deuxièmement", "d'abord", "ensuite", etc. qui permettent d'introduire les items (fig. 3.a), soit par des marqueurs lexicaux comme "tels que", "comme", etc. qui permettent d'introduire l'énumération (fig. 3.b). Mais elle peut aussi faire usage de marqueurs typographiques pour délimiter l'énumération, comme les parenthèses dans (fig. 3.c).

La **SE verticale** présente des discontinuités par rapport à la linéarité du texte. Des marqueurs typo-dispositionnels sont alors utilisés pour organiser, subdiviser et hiérarchiser les différents composants de la SE, comme le montre (fig. 3.d). Les items apparaissent en retrait par rapport à l'amorce, les items sont introduits par des puces, des tirets, etc.

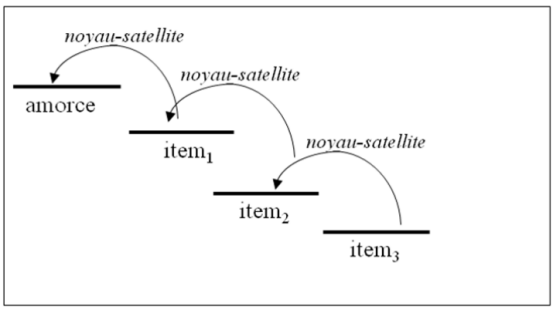
SE verticales et horizontales peuvent être combinées et imbriquées au sein d'une même SE. C'est le cas lorsqu'un item décrit lui-même une SE, avec ou sans mise en forme typo-dispositionnelle (fig. 3.e).

3.2 Typologie selon l'axe rhétorique

À ce niveau nous prenons en compte la nature des relations du discours qui relient les différents composants de la SE. Les relations entre items peuvent être de type noyau-satellite ou multi-nucléaire, selon la RST (Mann and Thompson, 1988). Une relation noyau-satellite relie une unité du discours plus saillante à une unité du discours qui supporte l'information d'arrière-plan, alors qu'une relation multi-nucléaire relie des unités du discours de même importance. Les SE, dont les items montrent une égalité d'importance, suscitent pour nous un intérêt particulier, car leur traduction en structures hiérarchiques est assez immédiate.



(a) structure rhétorique de la SE paradigmatique



(b) structure rhétorique de la SE syntagmatique

Figure 1: Représentations rhétoriques des SE paradigmatique et syntagmatique selon la RST.

Nous distinguons alors les **SE paradigmatiques**, les **SE syntagmatiques**, les **SE hybrides** et les **SE bivalentes**, reprenant ainsi en partie la terminologie utilisée par Luc (2001).

[3] ANR-07-MDCO-005, http://geonto.lri.fr/

[4] MIL : Marqueurs d'Intégration Linéaire

La **SE paradigmatique** est composée d'items indépendants dans un contexte donné. Elle porte alors une relation rhétorique multi-nucléaire entre les items successifs, chacun des items étant lié à l'amorce par une même relation de type noyau-satellite (fig. 1.a). Les exemples (a), (b), (c), entre autres, de la fig. 3 sont des cas de SE paradigmatiques. À l'opposé, la **SE syntagmatique** est composée d'items qui n'ont pas la même importance, et qui ne sont donc pas indépendants. La SE syntagmatique porte alors une relation rhétorique noyau-satellite entre items successifs (fig. 1.b). Le cas (fig. 3.f) en est un exemple.

Lorsqu'une SE porte une relation rhétorique noyau-satellite entre au moins deux items et une relation rhétorique multi-nucléaire entre au moins deux items, elle est qualifiée d'**hybride**. Enfin, les caractères paradigmatique et syntagmatique peuvent coexister au sein de la même SE, et dans ce cas la SE est dite **bivalente** (fig. 3.g).

3.3 Typologie selon l'axe intentionnel

À ce niveau nous prenons en compte l'intention de communication de l'auteur. Nous avons repris la typologie des textes pour l'adapter aux SE, en différenciant les **SE descriptives**, les **SE narratives**, les **SE prescriptives**, les **SE procédurales**, les **SE explicatives**, et les **SE argumentatives**. Ces types se sont révélés être les plus fréquents dans nos corpus. L'objectif est de caractériser les types de SE propices à la construction de RTO, pour ensuite proposer un modèle de représentation des connaissances adapté.

La **SE descriptive** décrit une entité qui peut être un objet du monde animé ou pas, artificiel ou naturel (fig. 3.a, fig. 3.b, fig. 3.c), alors que la **SE narrative** articule une succession d'actions ou d'événements, réels ou imaginaires (fig. 3.j). Les notions de conseil, d'indication, d'injonction peuvent être intégrées à ces types de SE. Dans ce cas la SE est dite **prescriptive** (fig. 3.i). De plus, lorsque ces conseils, indications, injonctions sont énoncés selon une volonté d'ordonnancer (comme dans les modes d'emploi, les notices explicatives, les guides d'utilisation, les manuels, les recettes de cuisine, etc.), pour atteindre un but donné, la SE est dite **procédurale** (fig. 3.h).

Enfin, la **SE explicative** répond en général à un questionnement de type "comment ?", "pourquoi?", "dans quelles circonstances?" etc. (fig. 3.f). Si des arguments sont avancés dans le but de défendre une opinion, dans le but de convaincre, la SE est dite **argumentative** (fig. 3.k).

En ce qui concerne cet axe, une même SE pourra posséder plusieurs traits intentionnels. La hiérarchie présentée en (fig. 2) décrit les combinaisons de types intentionnels les plus fréquentes.

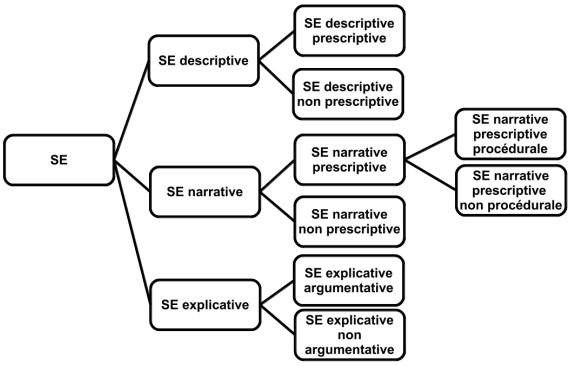


Figure 2: Combinaisons possibles des traits intentionnels au sein d'une même SE

Il existe cependant des SE pour lesquelles aucune des catégories de l'axe intentionnel précitées n'a pu être identifiée. Pour les catégoriser, nous avons défini le type **SE intentionnelle autre**.

3.4 Typologie selon l'axe sémantique

À ce niveau nous rendons compte de la dimension référentielle des SE, conformément à notre objectif de construction de ressources termino-ontologiques. Nous avons divisé les SE en trois catégories : **SE à visée ontologique** concerne des connaissances du monde (fig. 3.d et fig. 3.g), **SE métalinguistique** concerne la langue (fig. 3.l et fig. 3.m) et **SE sémantique autre** qui regroupe les SE qui ne sont ni à visée ontologique, ni métalinguistiques (fig. 3.o).

Une typologie des relations est associée aux types sémantiques "à visée ontologique" et "métalinguistique". Les relations *is-a* (fig. 3.a, fig 3.b, fig. 3.c), *part-of* (fig. 3.d, fig. 3.g), *instance-of* (fig. 3.n), *ontologique autre* (relation ontologique transverse ou d'actance) (fig. 3.i) sont associées aux SE à visée ontologique.

Les relations d'*hyperonymie*, de *méronymie*, d'*homonymie* (fig. 3.m), de *synonymie*, de *multilinguisme* (fig. 3.l), *lexicale autre* (relation lexicale moins fréquente décrivant la

langue, telle que la paronymie qui associe deux mots à la graphie/prononciation proches mais aux sens différents) sont associées aux SE métalinguistiques.

De façon orthogonale, les connaissances portées par la SE peuvent être contextualisées dans l'espace (fig. 3.j, fig. 3.n), dans le temps, ou dans tout autre dimension (fig. 3.m), à l'aide de circonstants. L'annotation de ces derniers permet d'envisager l'identification de relations autres que binaires. Nous distinguons les **SE contextuelles** des **SE non contextuelles**.

(a) Deux phénomènes sont responsables de l'augmentation substantielle du rayon de l'étoile (qui peut atteindre un rayon 1 000 fois supérieur à celui du Soleil). Premièrement, la fusion en couche de l'hydrogène. Et deuxièmement, la contraction du cœur d'hélium, libérant une importante quantité d'énergie gravitationnelle.

(b) Le dromadaire a été répertorié dans 35 pays, tels que l'Inde, la Turquie, le Kenya, le Pakistan, la corne de l'Afrique et bien d'autres encore.

(c) Les Grecs fabriquent généralement des meubles en bois (type érable, chêne, if, saule), mais aussi en pierre et en métal (bronze, fer, or, argent).

(d) Une chaussure se compose principalement :
- du semelage, partie qui protège la plante des pieds, plus ou moins relevée à l'arrière par le talon
- de la tige, partie supérieure qui enveloppe le pied

(e) Le bénéfice imposable est la différence entre les recettes et les charges de l'entreprise durant l'exercice comptable.
- Sont pris en compte pour les produits (recettes) :
  - les produits d'exploitation autrement dit le chiffre d'affaires de l'entreprise ;
  - les produits accessoires, c'est-à-dire les recettes.
- Sont pris en compte pour les charges (...) retenues pour leur coût hors taxe :
  - les frais généraux : salaire, loyer commercial, frais de bureau, etc. ;
  - les charges financières (agios, intérêts d'emprunt)

(f) Est considéré comme "lecture savante", du point de vue fonctionnel, une pratique de lecture répondant aux critères suivants :
- c'est une lecture "qualifiée",
- qui se développe sur le temps long de la recherche scientifique,
- dans un parcours forcément individualisé,
- où l'écriture se combine à la lecture, souvent dans une perspective de publications.

(g) Chaque nucléotide est constitué de trois éléments liés entre eux :
- un groupe phosphate lié à :
- un sucre, le désoxyribose, lui-même lié à :
- une base azotée.

(h) Préparation de la recette :

Lavez les asperges, épluchez-les de la pointe vers la base. Faites-les cuire dans une casserole d'eau bouillante avec les tablettes de bouillon pendant 25 à 30 minutes.

Égouttez-les et déposez-les précautionneusement sur du papier absorbant. Laissez-les refroidir.

Coupez-les en deux en réservant les pointes d'une longueur de 10 à 12 cm d'une part, les queues d'autre part.

(i) Selon ce décret, la BnF a pour mission :
- de collecter, cataloguer, conserver et enrichir dans tous les champs de la connaissance, le patrimoine national dont elle a la garde, en particulier le patrimoine de langue française ou relatif à la civilisation française.
- d'assurer l'accès du plus grand nombre aux collections, sous réserve des secrets protégés par la loi, dans des conditions conformes à la législation sur la propriété intellectuelle et compatibles avec la conservation de ces collections.

(j) Les Berbères ont mené une vive résistance parfois qualifiée de "farouche".
- Algérie : De nombreux soulèvements ont été menés pour contrer la colonisation française, l'émir Abd el-Kader qui faisait remonter ses origines à la tribu berbère des Banou Ifren (Zénètes) a lutté après avoir déclaré la guerre aux Français, il fut capturé puis fait prisonnier. En juillet 1857, (...)
- Maroc : Le mouvement de résistance s'est illustré lors de la guerre du Rif menée par Abdelkrim al-Khattabi, qui est une guerre coloniale qui opposa les tribus berbères du rif aux armées françaises et espagnoles, de 1921 à 1926. (...)
- Libye : La lutte contre la colonisation italienne est d'abord menée par Omar Al Mokhtar surnommé "Cheikh des militants " qui est un chef musulman libyen d'origine berbère qui organisa la lutte armée contre la colonisation italienne au début du XXe siècle. D'autres leaders nationalistes (...)

(k) Du point de vue de la tradition textuelle juive, la division en chapitres est non seulement une innovation étrangère sans aucun fondement dans la messora, mais elle est également fort critiquable car :
- la division en chapitres reflète souvent l'exégèse chrétienne de la Bible ;
- quand bien même ce ne serait pas le cas, elle est artificielle, divisant le Texte en des endroits jugés inappropriés pour des raisons littéraires ou autres.

(l) Munich [mynik] (München en allemand, Minga en bavarois) est, avec 1 443 122 habitants1, la troisième ville d'Allemagne par la population après Berlin et Hambourg.

(m) Une arête est un nom commun féminin qui peut désigner :
- l'arête, 'barbe de l'épi de graminées' (notion de botanique) ;
- l'arête, 'partie du squelette d'un poisson' (notion d'ichtyologie) ;
- l'arête, 'ligne d'intersection de deux plans' (notion de géométrie dans l'espace, d'architecture, etc.).

(n) Manoirs célèbres
- Le manoir d'Ango à Varengeville-sur-mer, près de Dieppe.
- Le manoir de Brion au Mont-Saint-Michel
- Le manoir d'Eyrignac à Salignac-Eyvigues en Périgord

(o) S sait que p si et seulement si
1. p est vrai ;
2. S croit que p ; et
3. la croyance de S dans p est justifiée.

Figure 3: Exemples de SE issus de pages Wikipedia ou du corpus de Virbel (1999)

4 Processus d'annotation

La typologie décrite ouvre la voie à une caractérisation plus fine des SE. Corollaire de cette possibilité, elle offre une latitude plus large pour la discrimination des classes lors d'un apprentissage supervisé pour l'identification des relations que portent les SE (Fauconnier et al., 2013).

Afin d'éprouver cette typologie de manière empirique, nous avons débuté une campagne d'annotation avec trois annotateurs. La tâche d'annotation elle-même se déroule en trois phases principales qui consistent à :

(1) délimiter les différents composants de la SE (amorce, items, clôture) lorsqu'elle bénéficie de mise en forme.

(2) annoter la SE selon les critères rhétoriques, intentionnels et sémantiques définis ci-dessus. Chaque SE se voit affecter un type rhétorique, un ou plusieurs types intentionnels, un type sémantique. Lorsque la SE est paradigmatique, à visée ontologique ou métalinguistique, un type de relation est associé au type sémantique (associée ou non à un contexte).

(3) délimiter, lorsque la SE est paradigmatique et à visée ontologique ou métalinguistique, les unités textuelles qui dénotent le concept présent dans l'amorce, le concept présent dans chacun des items, le circonstant (lorsqu'il existe) et la relation entre l'amorce et chacun des items.

Pour être menée à bien, cette tâche d'annotation nécessitait un outil adapté à la caractérisation multi-dimensionnelle des SE, cas moins courant en TAL où l'on privilégie habituellement des annotations simple label. De plus, il était aussi indispensable que cet outil supporte le caractère imbriqué et potentiellement récursif des SE. Par exemple, une SE peut contenir d'autres SE et elle-même être imbriquée au sein d'une structure discursive plus large (e.g : citation) ou être étalée sur plusieurs d'entre elles (e.g : un titre et plusieurs paragraphes). Enfin, cet outil devait être modulable pour être facilement adapté à d'autres types d'objets avec mise en forme (e.g : énoncés définitoires, démonstrations mathématiques, etc.) et plusieurs types de format d'entrée (e.g : HTML, PDF, etc.).

Les outils d'annotation tels que MMAX2 (Müller and Strube, 2006), MAE (Stubbs, 2011) ou encore Glozz (Widlöcher and Mathet, 2009) ne répondent pas ou partiellement à ces exigences. MMAX2 et MAE prennent du texte brut en entrée et ne gardent pas la mise en forme originelle des textes. Glozz, initialement conçu pour l'annotation de relations discursives, supporte la mise en forme du texte mais n'est, en l'état, pas adapté pour une annotation rapide et ergonomique d'objets multi-labels. En outre, la possibilité de faire évoluer le code source de Glozz n'est pas assurée (licence restrictive).

Pour toutes ces raisons, nous avons développé LARAt (Logiciel d'Acquisition de Relations par l'Annotation de textes[5]), prononcé /laʁa/. Cet outil Java se veut portable, et open-source. Dans son état actuel, LARAt prend en entrée des fichiers HTML ou XML respectant la norme TEI[6], les affiche en respectant leur mise en forme et permet aux annotateurs d'annoter des objets textuels imbriqués ou éclatés sur plusieurs niveaux textuels (e.g : titres et sous-titres).

Dans la tâche d'annotation des SE, deux types d'annotation sont produits (type 1 et type 2). Les annotations de type 1 concernent exclusivement le repérage en document des SE. Une fois délimitée, les SE sont caractérisées avec des annotations de type 2 qui reprennent les éléments décrits dans la typologie présentée. Ainsi, à chaque annotation de type 1 est associée une ou plusieurs annotations de type 2. Cette manière modulaire de gérer l'annotation facilite les post-traitements et l'emploi spécialisé de ces dernières (e.g : étude d'un phénomène particulier, recherche d'un cas précis pour exemplifier un emploi, etc.).

À terme, cet outil sera amené à supporter le PDF ainsi que le post-traitement des annotations (alignement, Kappa de Cohen et Fleiss pour l'accord inter-annotateurs).

À noter qu'un guide d'annotation accompagne cette campagne d'annotation. Sa rédaction se déroule de manière itérative en prenant en compte les retours des annotateurs et les cas ambigus qui posent question. Au terme de la campagne, le corpus annoté, le guide ainsi que LARAt seront distribués sous licence libre.

[5](en) *Layout Annotation for Relations Acquisition tool*
[6]Text Encoding Initiative

5 Conclusion et perspectives

L'analyse que nous avons menée sur les SE a permis de définir une typologie multi-dimensionnelle, permettant de tenir compte de propriétés de nature différente et parfois orthogonales. Le but théorique de ce travail a été d'élucider le phénomène complexe des SE quant à sa forme, sa structure ou sa fonction. D'un point de vue pratique, ce travail nous permet d'une part d'améliorer le repérage des SE dans les textes et, d'autre part d'identifier la ou les relations sémantiques qui relient les concepts contenus dans la SE. À cet égard, nous avons développé l'outil d'annotation LARAt qui permet de catégoriser les SE extraites de textes suivant les différents axes de notre typologie. Une première campagne d'annotation à l'aide de cet outil est en cours. La principale perspective de poursuite de ce travail est son extension à d'autres objets textuels ayant un impact sur la sémantique des textes tels que la titraille et les énoncés définitoires.

Références

N. Aussenac-Gilles and M.-P. Jacques. 2008. Designing and evaluating patterns for relation acquisition from texts with Caméléon. *Terminology*, 14:45–73.

P. Buitelaar, P. Cimiano, and B. Magnini. 2005. Learning taxonomic relations from heterogeneous sources of evidence. In P Buitelaar, P Cimiano, and B Magnini, editors, *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123, pages 59–73. IOS Press, Amsterdam.

J. Fauconnier, M. Kamel, B. Rothenburger, and N. Aussenac-Gilles. 2013. Apprentissage supervisé pour l'identification de relations sémantiques au sein de structures énumératives parallèles. In *Actes de la 20e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, pages 132–145.

Z. Harris. 1976. A theory of language structure. *American Philosophical Quarterly*, 13(4):237–255.

M. A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, volume 2, pages 539–545. Association for Computational Linguistics.

L.-M. Ho-Dac, M.-P. Péry-Woodley, and L. Tanguy. 2010. Anatomie des structures énumératives. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*.

E. H. Hovy and Y. Arens. 1991. Automatic Generation of Formatted Text. In *Proceedings of the 9th AAAI Conference (AAAI 1991)*, Anaheim, CA.

M. Kamel and B. Rothenburger. 2011. Elicitation de Structures Hiérarchiques à partir de Structures Enumératives pour la Construction d'Ontologie. In *Journées Francophones d'Ingénierie des Connaissances (IC 2011)*, pages 505–522, Annecy.

C. Luc. 2001. Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. In *Actes de la 8e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, pages 263–272.

W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

E. Montiel-Ponsoda and G. A. de Cea. 2011. Using natural language patterns for the development of ontologies. In V. Bhatia, P. Sánchez Hernández, and P. Pérez Paredes, editors, *Researching specialized languages*, volume 47, pages 211–230. John Benjamins.

C. Müller and M. Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

E. Pascual. 1991. *Représentation de l'architecture textuelle et génération de texte*. Ph.D. thesis, Université Paul Sabatier. Toulouse, France.

J. Poelmans, P.l Elzinga, S. Viaene, and G. Dedene. 2010. Formal concept analysis in knowledge discovery: a survey. In M. Croitoru, S. Ferré, and D. Lukose, editors, *Conceptual Structures: From Information to Intelligence*, volume 18, pages 139–153. Springer.

S. Porhiel. 2007. Les structures énumératives à deux temps. *Revue romane*, 42(1):103–135.

A. Stubbs. 2011. MAE and MAI: Lightweight Annotation and Adjudication Tools. In *2011 Proceedings of the Linguistic Annotation Workshop V, Association of Computational Linguistics*, Portland.

M. Vergez-Couret, L. Prévot, and M. Bras. 2008. Interleaved discourse, the case of two-step enumerative structures. In *Proceedings of Contraints In Discourse III*, pages 85–94, Potsdam.

J. Virbel. 1989. The contribution of linguistic knowledge to the interpretation of text structures. pages 161–180.

J. Virbel. 1999. Structures textuelles, planches fascicule 1 : Enumérations, Version 1,. Technical report, IRIT.

A. Widlöcher and Y. Mathet. 2009. La plateforme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*.

Peuplement d'une ontologie guidé par l'identification d'instances de propriété

Driss Sadoun[1,2] **Catherine Dubois**[3,4] **Yacine Ghamri-Doudane**[5] **Brigitte Grau**[1,3]

[1]LIMSI/CNRS, B.P. 133 91403 Orsay Cedex, France
prénom.nom@limsi.fr
[2]Université Paris-Sud, 91400 Orsay, France
[3]ENSIIE, 1 square de la résistance, 91000 Evry, France
prénom.nom@ensiie.fr
[4]CNAM-CEDRIC, 292 Rue St Martin FR-75141 Paris Cedex 03, France
[5]Laboratoire L3i, Université de La Rochelle, Av. Michel Crépeau, 17042, La Rochelle CEDEX 1, France.
prénom.nom@univ-mlv.fr

Résumé

Dans le but de formaliser des spécifications d'exigence écrites en langage naturel, nous avons choisi de modéliser les connaissances du domaine par une ontologie et de représenter formellement les spécifications par son peuplement. L'approche de peuplement est centrée sur l'identification d'instances de propriétés à partir des textes. Pour cela, des règles d'extraction sont acquises automatiquement à partir d'un corpus d'apprentissage, puis appliquées sur les textes pour l'identification de mentions d'instances de propriété représentées par des triplets. Ces règles exploitent les niveaux d'analyse lexicale, syntaxique et sémantique et sont engendrées à partir des chemins syntaxiques récurrents entre les termes pouvant dénoter des instances de concept ou de propriété. Nous montrons que l'identification d'instances de propriétés permet d'identifier de façon précise les instances de concepts énoncées de façon explicite ou implicite dans les textes.

1 Introduction

Tout processus de réalisation d'un système repose sur une phase de spécification des exigences. La vérification de la correction et la consistance de ces spécifications nécessitent l'utilisation de méthodes formelles qui peuvent être appliquées uniquement sur des spécifications formelles.

Or, en pratique, les spécifications d'exigences sont le plus souvent rédigées en langage naturel (LN) (Mich *et al.*, 2004) et sont donc non formelles. Leur vérification nécessite alors de les transformer en spécifications formelles. Se pose ainsi naturellement la question de l'automatisation du passage entre spécifications LN et spécifications formelles. Cette problématique n'est pas récente et a été abordée par différentes approches (Fougères et Trigano, 1997; Ilic, 2007; Ilieva et Boley, 2008; Kof, 2010; Bajwa *et al.*, 2012; Guissé *et al.*,). L'ensemble des approches pointe la difficulté d'une transformation directe et la nécessité de passer par un modèle intermédiaire palliant l'écart entre spécifications LN et spécifications formelles. Les modèles intermédiaires proposés dans la littérature sont en général semi-formels, comme UML ou SBVR.

Dans le cadre du projet ENVIE VERTE [1] dans lequel se place le travail décrit dans cet aricle, nous avons choisi de modéliser les connaissances du domaine par une ontologie en OWL-DL, une version décidable d'OWL2 et de représenter formellement les spécifications par son peuplement.

Une ontologie modélise les concepts et leurs propriétés, définissant ainsi le vocabulaire conceptuel d'un domaine. Un concept est la description d'un ensemble d'individus (d'instances) ayant une sémantique et des propriétés communes. Une instance de concept est une concrétisation d'un concept. Par exemple, dans notre modélisation du domaine *kitchen* correspond à une instance du concept *Location* [2]. Une propriété est définie entre deux concepts, i.e. une relation, ou entre un concept et un type de donnée, i.e un attribut. Une instance de propriété relie donc deux instances de concepts, par exemple *Occured-in(movement,kitchen)*, ou une instance de concept et une valeur d'attribut, par exemple *Has-value*(temperature,25). Une ontologie offre un cadre formel permettant d'associer

1. financé par DIGITEO, projet DIM LSC 2010.

2. Les exemples sont en anglais, car les textes que nous analysons sont rédigés dans cette langue

une sémantique aux termes issus des textes. L'association des instances de concepts et propriétés à leurs formulations dans les textes est régie à l'aide d'une ontologie lexicale en SKOS (Simple Knowledge Organization System) contenant la terminologie liée au vocabulaire conceptuel.

Peupler une ontologie consiste à y ajouter de nouvelles instances sans en changer la structure conceptuelle (Petasis *et al.*, 2011). Ces nouvelles instances sont associées aux concepts et propriétés reconnus dans les textes. L'identification peut être centrée sur la reconnaissance d'instances de concepts (Thongkrau et Lalitrojwong, 2012), ou sur la reconnaissance d'instances de relation (Nakamura-Delloye et Stern, 2011).

Dans cet article, nous proposons de guider le peuplement de l'ontologie par l'identification de triplets de termes dans les textes correspondant à des instances de propriétés. Nous distinguons deux types de triplets : complets et partiels. Les triplets complets contiennent la mention d'une instance de propriété ainsi que les mentions des deux instances de concepts qu'elle lie. En revanche, les triplets partiels ont pour vocation de reconnaître des propriétés pour lesquelles une des deux instances n'est pas explicitement mentionnée. Guider le peuplement de l'ontologie par identification d'instances de propriétés permet de résoudre des cas d'ambiguïté de termes et d'informations implicites pour lesquels les concepts associés sont trouvés par inférence dans l'ontologie à partir des instances de propriétés. De la sorte, l'identification d'instances de concepts ne repose pas seulement sur la reconnaissance de leurs mentions dans les textes mais aussi sur les propriétés conceptuelles auxquelles elles sont associées.

Les triplets correspondant aux instances de propriétés sont extraits à l'aide de règles acquises par amorçage à partir d'un corpus d'apprentissage et d'une terminologie de départ. Ces règles correspondent à des formes lexico-syntaxiques récurrentes. Nous montrons que cette approche permet d'identifier de manière fiable des instances de propriétés, et dans un deuxième temps d'inférer les instances de concepts qui leurs sont associées. De plus, l'approche que nous proposons peut s'adapter aisément à d'autres domaines d'application dès lors que l'on peut le décrire par une ontologie.

2 Travaux connexes

Le peuplement d'ontologie consiste à identifier et à classer les instances extraites des textes. Se pose le problème de la reconnaissance de mentions d'instances de concepts et de propriétés dans les textes. L'hypothèse généralement faite est que les paires d'entités apparaissant dans un même contexte peuvent être considérées comme des instances de la même relation. La définition du contexte peut être restreinte par la présence d'un verbe et la reconnaissance de son entourage (Lin et Pantel, 2001; Makki *et al.*, 2008). D'autres travaux se fondent sur la classification entre couples d'entités connues pour être liées par une relation sémantique (Hasegawa *et al.*, 2004; Nakamura-Delloye et Stern, 2011; Thongkrau et Lalitrojwong, 2012). La majorité des approches exploitent des connaissances lexicales et syntaxiques pour la définition d'un contexte représentant une relation sémantique. Cependant, dans (IJntema *et al.*, 2012), les auteurs avancent que les patrons lexico-sémantiques sont plus à même de capturer dans les textes le contexte sémantique. Alors qu'ils proposent un langage d'écriture manuelle de règles, nous proposons d'acquérir ce type de règle automatiquement, à partir des trois niveaux de connaissances lexicale, syntaxique et sémantique.

L'ensemble des approches identifient les instances de concept au niveau du texte. La méthode que nous proposons tire partie des connaissances sémantiques pour inférer au sein de l'ontologie l'appartenance d'une instance à un concept. De plus, nous proposons d'aller plus loin que la classification d'instances en identifiant dans les textes les mentions dénotant un même individu.

3 Notations

Dans la suite de l'article nous emploierons les notations suivantes. Les noms de concept et de propriété sont en *italique* et commencent par une majuscule. Les noms d'instances de concept sont en *italique* et commencent par une minuscule. Un concept sera noté C_i et les instances de concept i_{C_i}. Les propriétés sont définies sur un domaine et une image [3]. Les propriétés entre concepts sont notées $P_k(C_i, C_j)$ et les instances de propriété

3. Le domaine contient un concept ou un ensemble de concepts et l'image contient soit un concept ou un ensemble de concepts ou un littéral (entier, chaîne de caractères, ...)

sont notées $P_k(i_{C_i}, i_{C_j})$. Dans les textes, un triplet contenant la mention d'une instance de propriété, t_P, entre deux instances de concepts est notée (t_P, t_{C_i}, t_{C_j}) avec t_{C_i} et t_{C_j} les termes qui dénotent respectivement les instances du concept C_i et du concept C_j.

4 Acquisition des règles d'extraction

Dans les textes, les instances de concepts et de propriétés sont dénotées par des termes. Guider l'identification d'instances de concepts à partir des propriétés qui les définissent requiert de reconnaître des triplets (t_P, t_{C_i}, t_{C_j}). Aussi, chaque règle d'extraction a comme objectif de reconnaître la mention d'une propriété de l'ontologie et d'extraire du texte les triplets de termes qui la dénotent. Deux types de triplets sont à reconnaître :

- triplet complet : les mentions d'instances des domaine et image de la propriété sont explicites dans le texte, soit (t_P, t_{C_i}, t_{C_j}) ;
- triplet partiel : l'une des mentions d'instances des domaine ou image de la propriété n'est pas explicite. Elle correspondra à une inconnue dans le triplet, soit $(t_P, ?i, t_{C_j})$ ou $(t_P, t_{C_i}, ?i)$.

Par exemple à partir de la phrase : *when a person moves into the kitchen, switch on the light.*, on peut identifier le triplet $(Occured\text{-}in, move, kitchen)$ qui dénote une instance de la propriété $Occured\text{-}in$ liant une instance du concept $Phenomenon$ à une insatnce du concept $Location$. Néanmoins dans cette même phrase, l'agent qui doit allumer la lumière n'est pas explicité. Le triplet à identifier dans ce cas est $(Turn\text{-}on$[4]$, ?A, light)$ qui dénote une instance de la propriété $Turn\text{-}on$ liant une instance non mentionnée du concept *Actuator* à une instance du concept *Physical-process*.

4.1 Méthode d'acquisition des règles d'extraction

L'acquisition des règles d'extraction se fait automatiquement à partir d'un corpus d'apprentissage et d'une termino-ontologie amorce. Elles sont acquises à partir des chemins syntaxiques les plus fréquents entre des paires de termes. Ces termes sont issus de classes sémantiques connues pour être liées dans l'ontologie. L'extraction des deux types de triplets nécessite l'acquisition de deux types de règles d'extraction.

Dans le cas de triplets partiels, les règles sont acquises à partir des chemins les plus fréquents entre les termes qui dénotent les instances d'une propriété et les instances de son domaine ou de son image. Pour les triplets complets, les chemins sont constitués des deux chemins partiels entre paires de termes dénotant les instances de domaine et propriété et celles dénotant propriété et image.

L'algorithme 1 décrit le processus d'acquisition des règles. Chaque phrase du corpus est analysée et son arbre de dépendances syntaxiques est engendré. Puis trois fonctions sont appelées pour l'extraction de chemins syntaxiques entre termes. Elles permettent d'identifier des chemins liant trois types d'ensemble de termes : les triplets de termes d'une propriété, de son domaine et de son image, les paires de terme d'une propriété et de son image et les paires de terme d'une propriété et de son domaine. Les chemins extraits sont comparés en fonction de leurs dépendances et des formes lemmatisées des termes. Les chemins identiques les plus fréquents pour chaque type de paires sont retournés. Enfin, les règles d'extraction sont générées à partir des caractéristiques des chemins retournés (cf. section 4.3).

4.2 Chemin syntaxique

L'analyse syntaxique des phrases permet d'en extraire des arbres de dépendances syntaxiques cf. figure 1. Chaque nœud est étiqueté par un terme et sa catégorie morpho-syntaxique (nom, verbe, etc). Les nœuds sont reliés deux à deux par des dépendances syntaxiques qui constituent des liens orientés. Un chemin syntaxique est composé des dépendances syntaxiques liant deux termes dans un arbre de dépendances. La figure 1 représente l'arbre de dépendances de la phrase *"When a person moves into the kitchen, switch on the light."* [5]

Par exemple, le chemin syntaxique entre le terme *moves* dénotant une instance du concept *Phenomenon* et le terme *kitchen* dénotant une instance du concept *Location* est $prep(moves, into)$-$pobj(into, kitchen)$ (chemin en gras, figure 1).

4.3 Génération des règles d'extraction

Les règles d'extraction ont pour rôle d'identifier des instances de propriétés. Ainsi, elles

4. Turn−on est la dénomination préférée de swich on

5. produit à l'aide de DependenSee.jar http ://chaoticity.com/dependensee-a-dependency-parse-visualisation-tool/

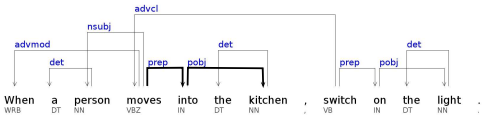


FIGURE 1: Arbre de dépendances syntaxiques

modélisent le contexte sémantique dans lequel les termes désignant des instances de concepts apparaissent. La génération de règles d'extraction est automatique. Elle exploite les caractéristiques issues des n chemins syntaxiques identiques les plus fréquents. Ces caractéristiques sont de trois types :

- dépendances syntaxiques ;
- catégorie termino-ontologique (sémantique) ;
- catégorie morpho-syntaxique.

Données: corpus, ontologie, skos
$Paths \leftarrow \emptyset$;
Variables : Ensemble de termes T_D, T_P, T_I;
pour *chaque phrase S du corpus* **faire**
  $Tr \leftarrow getDependencyTree(S)$;
  $C \leftarrow getConcepts(ontologie)$;
  for *each c in C* **do**
    $T_D \leftarrow getTerms(c, skos)$;
    $P \leftarrow getPropertiesOf(c, ontologie)$;
    for *each p in P* **do**
      $T_P \leftarrow getTerms(p, skos)$;
      $Img \leftarrow getImagesOf(p, ontologie)$;
      for *each img in Img* **do**
        $T_I \leftarrow getTerms(img, skos)$;
        // - instances de propriété complète -
        $Paths \leftarrow Paths \cup$ extrPaths(Tr,T_P,T_D, T_I);
        //- instances de propriété partielle -
        //domaine implicite
        $Paths \leftarrow Paths \cup$ extrPaths(Tr,T_P,T_I);
        //image implicite
        $Paths \leftarrow Paths \cup$ extrPaths(Tr,T_P,T_D);
      end
    end
  end
fin
$Chemins \leftarrow getMostFrequent(Paths)$;
pour *chaque ch de $Chemins$* **faire**
  $createCorrespondingRule(ch)$;
fin

Algorithm 1: Acquisition des règles d'extraction

Exemple : l'un des chemins récurrents entre les deux ensembles de termes des concepts *Phenomenon* et *Location* qui sont dans l'ordre le domaine et l'image de la propriété *Occured-in* est $prep(t_P, t_O) \wedge pobj(t_O, t_L)$, avec :

- les dépendances syntaxiques du chemin
  - $prep(t_P, t_O) - pobj(t_O, t_L)$
- les catégories morpho-syntaxiques
  - t_P est un verbe ;
  - t_L est un nom ;
  - t_O est une préposition ;
- les catégories termino-ontologiques
  - t_P dénote le concept *Phenomenon* ;
  - t_O dénote la propriété *Occured-in* ;
  - t_L dénote le concept *Location* ;

Les dépendances syntaxiques sont transformées en prédicats, les catégories morpho-syntaxiques et termino-ontologiques sont transformées en contraintes. Cela permet d'engendrer la règle d'extraction de la propriété $Occured\text{-}in(Phenomenon, Location)$ avec $T_{Location}, T_{Occured-in}, T_{Phenomenon}$ trois ensembles de termes issus de l'ontologie lexicale :

$prep(t_P, t_O) \wedge pobj(t_O, t_L) \wedge isPrep(t_O) \wedge isVerb(t_P) \wedge isNoun(t_L) \wedge T_{Location}(t_L) \wedge T_{Occured-in}(t_O) \wedge T_{Phenomenon}(t_P) \rightarrow Occured\text{-}in(t_P, t_L)$

5 Expansion de la terminologie

La terminologie est représentée en SKOS. Ce formalisme permet de définir pour chaque terme sa formulation préférée ainsi que sa liste de formulations synonymes.

Afin d'augmenter la terminologie amorce, nous appliquons les règles acquises pour extraire de nouveaux termes sur le corpus d'apprentissage. Chacune des règles engendre trois applications. Chaque application a comme objectif l'extraction de termes dénotant un ensemble sémantique représenté dans la règle. Soit R une règle d'extraction et $T_{Domaine}, T_{Image}, T_{Prop}$ trois ensembles de termes. R peut alors être considérée comme une fonction définie comme suit :

$R : T_{Domaine} \times T_{Prop} \rightarrow T_{Image}$
$R : T_{Image} \times T_{Prop} \rightarrow T_{Domaine}$
$R : T_{Domaine} \times T_{Image} \rightarrow T_{Prop}$

La pertinence d'un terme t extrait par des règles R_i est calculée selon la formule suivante :

$Pertinence(t) = (\sum_{i=1}^{n} freq(t, R_i)) * n$

avec $freq$ sa fréquence, et n le nombre de règles du même type à partir desquelles il est extrait. Cette formule permet de favoriser les termes extraits par plusieurs règles.

6 Peuplement de l'ontologie

Suite à l'identification des instances de propriété, vient la phase de classification des instances de concepts. À ce stade, les classes sémantiques des instances liées par les propriétés ne sont pas encore connues. La classification de ces instances est fondée sur le raisonnement permis par OWL sur leurs propriétés. De cette façon, les ambiguïtés liées aux termes sont résolues et les mentions d'instances implicites dans les textes sont déduites à partir des propriétés possédées par les instances de concept. Classer les instances dans l'ontologie ne suffit pas pour représenter de manière cohérente les connaissances issues des textes. Pour cela, il faut aussi être en mesure d'identifier de manière univoque chaque individu.

6.1 Classification des instances

La classification des individus consiste à les associer aux concepts qui les dénotent. Nous présentons dans cette section comment différents mécanismes d'inférence de OWL peuvent être exploités à cette fin.

6.1.1 Domaine et Image des propriétés

Lors du raisonnement sur les instances de l'ontologie, chaque instance qui participe à une propriété est associée à la classe sémantique du domaine ou de l'image de la propriété [6] de la manière suivante : Soit P_1 une propriété avec comme domaine D_1 et comme image I_2 et soit i_1 et i_2 deux instances de concepts. Alors si $P_1(i_1, i_2)$ on déduit : $i_1 \in D_1$ et $i_2 \in I_2$

6.1.2 Condition Nécessaire et Suffisante

OWL permet de définir des équivalences entre les concepts et certaines de leurs propriétés, formant des conditions nécessaires et suffisantes pour inférer le concept auquel appartient une instance à partir de ses propriétés. Par exemple l'axiome $C_1 \equiv P_2.C_2 \sqcap P_3.C_3$ définit une équivalence entre le concept C_1 et ses deux propriétés P_2 et P_3 ayant pour images dans l'ordre C_2 et C_3. Cet axiome permet l'inférence suivante : Soit $i_{C_2} \in C_2$ et $i_{C_3} \in C_3$ alors $P_2(i, i_{C_2}) \wedge P_3(i, i_{C_3}) \rightarrow i \in C_1$

6. Par défaut la classe sémantique est *Thing*

6.2 Identification des instances

Dans les ontologies en OWL, les instances identiques sont identifiées à l'aide de la propriété *SameAs*. L'inférence de *SameAs* exploite les propriétés qui, à l'instar des clés primaires que l'on trouve en base de données, permettent d'identifier de manière unique les individus. Les propriétés représentant une contrainte d'unicité peuvent être définies de deux manières.

6.2.1 Propriété fonctionnelle et inverse fonctionnelle

Une *propriété fonctionnelle* associe à chaque individu du domaine, un seul individu de l'image. Cela permet l'inférence suivante : Soit P_n une propriété fonctionnelle, i_i, i_j et i_k trois individus,

$P_n(i_i, i_j) \wedge P_n(i_i, i_k) \rightarrow SameAs(i_j, i_k)$

La *propriété inverse fonctionnelle* a, pour chaque individu de l'image, un seul individu de domaine possible. Cela permet l'inférence suivante : Soit P_n une propriété inverse fonctionnelle, i_i, i_j et i_k trois individus,

$P_n(i_j, i_i) \wedge P_n(i_k, i_i) \rightarrow SameAs(i_j, i_k)$

6.2.2 Contraintes en SWRL

Dans certains cas, plus d'une propriété est nécessaire pour représenter une contrainte d'unicité, par exemple une personne est identifiée par ses nom, prénom et date de naissance. Ce type de contrainte doit définir les propriétés que deux individus doivent nécessairement partager pour être inférés comme semblables. Ce type de contraintes peut être défini à l'aide de règles SWRL. Par exemple, la règle ci-dessous énonce que les deux individus i_x et i_y sont un même individu s'ils possèdent des valeurs similaires à travers les propriétés P_1 et P_2,

$P_1(i_x, i_{C_1}) \wedge P_1(i_y, i_{C_1}) \wedge P_2(i_x, i_{C_2}) \wedge P_2(i_y, i_{C_2}) \rightarrow SameAs(i_x, i_y)$

7 Expérimentation

Dans cette section, nous présentons le domaine d'application, les ressources utilisées, ainsi que les évaluations d'acquisition de termes et d'extraction d'instances de propriétés. Enfin, nous donnons un exemple de création, de classification et d'identification d'instances de concept à partir d'une phrase extraite de notre corpus de test.

7.1 Environnement intelligent

Un environnement intelligent est un ensemble d'objets communicants (capteurs, actionneurs et processus de contrôle), dont le comportement général est décrit ci-dessous :

- Un capteur détecte l'occurrence d'un type phénomène ou mesure un type phénomène dans une zone restreinte.
- Un capteur détecte ou mesure un type de phénomène s'il est localisé dans sa zone de capture.
- Un actionneur est connecté à un appareil (processus physique) de l'environnement qu'il peut actionner.
- La détection d'un phénomène peut conduire à l'activation d'un ou de plusieurs actionneurs et actionner une ou plusieurs actions (turn on, turn off, decrease ou increase) sur les processus physique qu'ils contrôlent.
- Un actionneur, pour être activé par un capteur, doit partager son type et être localisé dans sa zone de contrôle.

Afin de modéliser ce domaine nous avons défini une ontologie de haut niveau contenant 12 concepts, 15 propriétés entre concepts et 9 entre concepts et types de données. Cette ontologie comporte des instances initiales qui correspondent à un environnement physique d'un utilisateur, qui n'est pas amené à être modifié par celui-ci. Cette ontologie est suffisante pour représenter notre domaine, dans la mesure où nous nous intéressons au fonctionnement d'un réseau de capteurs, et elle peut-être reprise pour différentes configurations, car seules les instances de départ changeront (voir (Sadoun *et al.*, 2012) pour une justification de cette conceptualisation). L'ensemble des individus sont identifiables à partir de leurs propriétés de localisation et de type.

7.2 Description des ressources

En l'absence de corpus suffisamment grand portant sur la spécification d'exigences dans le domaine des environnements intelligents, nous avons constitué un corpus d'apprentissage d'environ 5 millions de mots à partir de livres électroniques (e-books) de domaines et styles littéraires différents, issus de la *Bibliothèque numérique Anacleto*[7]. La diversité de ce corpus permet d'acquérir des règles pour des styles d'écriture différents. Par ailleurs, les propriétés recherchées sont suffisamment générales et transdomaines pour qu'on puisse partir d'un corpus non spécialisé pour acquérir les règles d'extraction. En contrepartie, sa généralité limite l'extraction de la terminologie pour les concepts spécifiques au domaine. Afin de disposer d'un corpus d'évaluation, nous avons développé une plate-forme[8] de collecte de spécifications. Les spécifications collectées représentent environ 80 phrases (1558 mots).

7.3 Résultats

7.3.1 Acquisition des règles d'extraction

Nous avons acquis en tout 126 règles d'extraction, dont 31 pour l'identification d'instances de propriétés complètes et 95 pour l'identification d'instances de propriétés partielles. Chaque règle a été générée à partir des n chemins syntaxiques les plus fréquents. Nous avons fixé ce paramètre de façon expérimentale à 4. Néanmoins, les mentions de certaines propriétés sont moins fréquentes dans les textes et ce nombre peut s'avérer trop bas. En l'augmentant, des chemins non pertinents peuvent engendrer des règles. Afin d'éviter ce bruit éventuel, nous fixons une seconde limite correspondant au nombre de dépendances syntaxiques composant le chemin. En effet, plus un chemin est long, et donc plus les termes sont distants dans la phrase, moins il y a de possibilité que ce chemin représente une propriété sémantique. Dans nos expériences, le nombre de dépendances maximal d'un chemin a été fixé à 4.

7.3.2 Évaluation de l'acquisition de termes

La terminologie amorce contient 109 termes, 18 termes préférés et 91 alternatifs. Ces termes sont associés aux instances de concepts et propriétés modélisés dans l'ontologie. Certains termes sont la dénomination d'individus préexistant dans l'ontologie, par exemple les localisations et les différents types reconnus par les capteurs.

Le tableau 1 illustre l'acquisition des termes sur trois catégories sémantiques par expansion de la terminologie. Ces catégories sémantiques représentent dans l'ordre les phénomènes à reconnaître, les processus physiques qui correspondent aux appareils connectés aux réseaux de capteurs et les actions possibles[9]. Les règles d'ex-

7. (http ://www.gutenberg.us/)

8. http ://perso.limsi.fr/sadoun/Application/fr/SmartHome.php

9. Actuate-on a comme sous-propriété : Turn-on, Turn-off, Increase et Decrease

traction ont été appliquées sur le corpus d'apprentissage comme cela est décrit en section 5. Le première colonne exprime les termes amorces issus de la terminologie de départ. La seconde colonne le nombre de termes différents extraits et la troisième colonne les termes pertinents.

	Amorce	Extrait	Pertinent
Phenomenon	18	965	49
Physical-process	33	625	23
Actuate-on	19	413	27

TABLE 1: Termes extraits pertinents

En raison de la généralité du corpus et la spécificité du domaine, la précision de l'extraction est assez basse. La sélection des termes pertinents est réalisée manuellement à partir de l'examen des n premiers termes retournés par la formule du calcul de pertinence cf. section 5, avec n fixé à 25%.

7.3.3 Évaluation de l'extraction des triplets candidats

L'extraction des triplets résulte de l'application sur le corpus de test des règles acquises. L'application des règles complètes est prioritaire par rapport aux règles partielles. Cette extraction est effectuée en ne considérant que les termes pertinents dans la terminologie.

Les résultats de l'extraction de triplets candidats sont décrits dans le tableau 2. La première colonne indique le nombre d'instances à reconnaître. Les colonnes suivantes indiquent le nombre de triplets correctement identifiés, et les triplets identifiés à tort. La première ligne représente les résultats pour les trois propriétés *Located-in*, *Fixed-in* et *Occured-in* qui associent une localisation à chacun des concepts *Localisation*, de *Physical-process* et *Phenomenon*. La propriété *Has-type* associe un *Type* aux phénomènes. Les deux dernières lignes représentent les instances de concepts *Phenomenon* et *Actuator* qui sont classées et identifiées lors du raisonnement sur leurs instances de propriétés.

Nous observons que la précision est très élevée (0.95). Cela montre la pertinence des règles acquises. De plus le rappel obtenu (0.63) est relativement élevé compte tenu du fait que l'acquisition des règles d'extraction et de la terminologie s'est faite à partir d'un corpus non spécifique au domaine. À partir des instances de propriété créées dans l'ontologie, 22 instances de *Phenomenon* et

	Pertinent	Correct	Incorrect	P	R	F-M
Loc	115	75	9	0.89	0.65	0.75
Has-type	62	35	0	1	0.56	0.72
Actuate-on	90	51	0	1	0.56	0.71
Total	267	164	9	0.95	0.61	0.7
Phenomenon	62	22	0	1	0.35	0.51
Actuator	42	17	0	1	0.40	0.57

TABLE 2: Extraction de triplets candidats

17 instances de *Actuator* ont été classées correctement, aucune n'a été incorrectement classée. Les instances de *Phenomenon* ont été identifiées comme appartenant à 10 individus différents. Les instances d'*Actuator* ont été identifiées comme appartenant à 7 individus différents.

7.4 Application

Les instances de propriété sont créées à partir des triplets extraits des textes. Lors de leur création, les termes représentant les domaine et image de la propriété sont nommés de la manière suivante : Si la formulation préférée d'un terme dénote un individu de l'ontologie alors il prend le nom de l'individu. sinon son nom est composé à partir du numéro de la phrase dans laquelle il apparaît et de son numéro de nœud dans l'arbre de dépendances syntaxiques. Par exemple, si le terme apparaît dans la phrase numéro 2 et son numéro de nœud dans l'arbre syntaxique est 3 alors il sera nommé 2-3. Cela permet de nommer les instances de façon unique et de mettre au même niveau les instances de concept explicites ou implicites dans les textes, issues des règles complètes ou partielles. Ainsi l'instance 2-3 qui apparaît dans les propriétés *Occured-in(2-3 , kitchen)* et *Has-type(2-3 , movement)* est déduite comme appartenant au concept *Phenomenon* de deux façons : à partir du domaine des propriétés *Occured-in* et *Has-type* définies sur le concept *Phenomenon* (cf. 6.1.1) et à partir de l'axiome d'équivalence : *Phenomenon* $\equiv$ *Has-type.Type* $\wedge$ *Occured-in.Location* qui définit une contrainte nécessaire et suffisante (cf. 6.1.2) pour reconnaître une instance du concept *Phenomenon*. L'instance est ensuite identifiée par rapport aux autres instances du concept *Phenomenon* à partir de la règle SWRL suivante :
Occured-in $(i_{P_1}\ ,\ i_L) \wedge$ *Occured-in*$(i_{P_2}\ ,\ i_L) \wedge$ *Has-type*$(i_{P_1}\ ,\ i_T) \wedge$ *Has-type*$(i_{P_2}\ ,\ i_T) \rightarrow$ *Sa-*

meAs(i_{P_1}, i_{P_2})

Cette règle exprime la contrainte d'unicité (cf. 6.2.2) inhérente aux individus du concept *Phenomenon*. Lors du raisonnement elle s'exprime concrètement de la manière suivante :
Occured-in(2-3 , kitchen) $\wedge$ *Occured-in* (i_2 , kitchen) $\wedge$ *Has-type*(2-3 , movement) $\wedge$ *Has-type*(i_2 , movement) $\rightarrow$ *SameAs*(2-3,i_2)

8 Conclusion

Nous avons présenté une approche de peuplement d'ontologie visant à représenter de manière formelle des connaissances issues de spécifications d'exigences. Cette approche est centrée sur l'identification de mentions d'instances de propriété dans les textes. Cette identification est faite à l'aide de règles d'extraction acquises à partir des chemins syntaxiques récurrents entre les termes dénotant les instances de concept et de propriété. Elles exploitent des connaissances lexicales, syntaxiques et sémantiques. Ces règles permettent d'identifier des instances de propriétés même lorsque l'une des instances du domaine ou de l'image est implicite. De plus, ces règles permettent de lever l'ambiguïté des termes extraits en capturant le contexte sémantique dans lequel ils apparaissent. Lors du raisonnement au sein de l'ontologie, les propriétés permettent de classer les instances de concepts et de les identifier de manière unique grâce aux contraintes d'unicité modélisées sous OWL. L'approche proposée a été conçue pour être indépendante du domaine et s'adapter facilement à d'autres langues. A partir d'une ontologie modélisée, elle ne nécessite qu'un corpus d'apprentissage ainsi qu'un ensemble de termes de départ. De plus seul le parseur utilisé et l'ensemble de termes de départ sont dépendants de la langue des textes à analyser.

References

BAJWA, I. S., LEE, M. et BORDBAR, B. (2012). Resolving syntactic ambiguities in natural language specification of constraints. *In Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*.

FOUGÈRES, A.-J. et TRIGANO, P. (1997). Rédaction de spécifications formelles : Èlaboration à partir des spécifications écrites en langage naturel. *In Cognito-Cahiers Romans de Sciences Cognitives*, 1(8):29–36.

GUISSÉ, A., LÉVY, F. et NAZARENKO, A. From regulatory texts to brms : how to guide the acquisition of business rules ? *In Proceedings of the 6th international conference on Rules on the Web : research and applications*.

HASEGAWA, T., SEKINE, S. et GRISHMAN, R. (2004). Discovering relations among named entities from large corpora. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

IJNTEMA, W., SANGERS, J., HOGENBOOM, F. et FRASINCAR, F. (2012). A lexico-semantic pattern language for learning ontology instances from text. *Web Semantics : Science, Services and Agents on the World Wide Web*, 15:37–50.

ILIC, D. (2007). Deriving formal specifications from informal requirements. *In Proceedings of the 31st Annual International Computer Software and Applications Conference - Volume 01*, COMPSAC '07, pages 145–152. IEEE Computer Society.

ILIEVA, M. et BOLEY, H. (2008). Representing textual requirements as graphical natural language for uml diagram generation. *In SEKE'08*, pages 478–483.

KOF, L. (2010). Requirements analysis : concept extraction and translation of textual specifications to executable models. *In Proceedings of the 14th international conference on Applications of Natural Language to Information Systems*, NLDB'09.

LIN, D. et PANTEL, P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering*.

MAKKI, J., ALQUIER, A.-M. et PRINCE, V. (2008). Ontology population via nlp techniques in risk management. *In ICSWE*.

MICH, L., FRANCH, M. et INVERARDI, P. (2004). Market research for requirements analysis using linguistic tools. *Requirement Engineering*, 9(1):40–56.

NAKAMURA-DELLOYE, Y. et STERN, R. (2011). Extraction de relations et de patrons de relations entre entités nommées en vue de l'enrichissement d'une ontologie. *In TOTh*.

PETASIS, G., KARKALETSIS, V., PALIOURAS, G., KRITHARA, A. et ZAVITSANOS, E. (2011). Ontology population and enrichment : State of the art. *In Knowledge-Driven Multimedia Information Extraction and Ontology Evolution'11*.

SADOUN, D., DUBOIS, C., GHAMRI-DOUDANE, Y. et GRAU, B. (2012). Formalisation en OWL pour vérifier les spécifications d'un environnement intelligent. *In Actes de la conférence RFIA*.

THONGKRAU, T. et LALITROJWONG, P. (2012). Ontopop : An ontology population system for the semantic web. *IEICE Transactions*, 95-D(4):921–931.

Session : Medical terminologies

Discovering Semantic Frames for a Contrastive Study of Verbs in Medical Corpora

Ornella Wandji
CNRS UMR 8163 STL
Université Lille 3
59653 Villeneuve d'Ascq, France
ornwandji@yahoo.fr

Marie-Claude L'Homme
OLST, Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal H3C 3J7
Québec, Canada
mc.lhomme@umontreal.ca

Natalia Grabar
CNRS UMR 8163 STL
Université Lille 3
59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

Abstract

The field of medicine gathers actors with different levels of expertise. These actors must interact, although their mutual understanding is not always completely successful. We propose to study corpora (with high and low levels of expertise) in order to observe their specificities. More specifically, we perform a contrastive analysis of verbs, and of the syntactic and semantic features of their participants, based on the Frame Semantics framework and the methodology implemented in FrameNet. In order to acheive this, we use an existing medical terminology to automatically annotate the semantics classes of participants of verbs, which we assume are indicative of semantics roles. Our results indicate that verbs show similar or very close semantics in some contexts, while in other contexts they behave differently.

1 Introduction

The field of medicine is heterogeneous because it gathers actors with various backgrounds, such as medical doctors, students, pharmacists, managers, biologists, nurses, imaging experts and of course patients. These actors have different levels of expertise ranging from low (typically, the patients) up to high (*e.g.*, medical doctors, pharmacists, medical students). Moreover, actors with different levels of expertise interact, but their mutual understanding might not always be completely successful. This specifically applies to patients and medical doctors (AMA, 1999; McCray, 2005; Zeng-Treiler et al., 2007), but we assume that similar situations apply to other actors.

In this study, we propose to perform a comparative analysis of written medical corpora, which are differenciated according to their levels of expertise. More specifically, we concentrate on the study of selected verbs used in these corpora and aim to characterize the syntactic and semantic features of their participants. Most of the participants are arguments (or, in terms of Frame Semantics, *core frame elements*). They often correspond to noun phrases. The description of verbs is based on the Frame Semantics framework (Fillmore, 1982). We assume that verbs are an excellent starting point for modeling the contents and semantics of sentences. The study is perfomed with French data. In the following, we briefly present previous work on verbs in specialized languages (section 2) and on Frame Semantics (section 3). We also describe the material that we use (section 4) and the method developed to process it (section 5). We then give an account of the results (section 6), and conclude with some directions for future work (section 7).

2 Verbs in specialized languages

Traditionally, the study of specialized languages focuses on nominal entities (typically, nouns and noun phrases), commonly used for the compilation of terminologies, ontologies, thesauri or vocabularies. This situation can be explained by the needs raised by specific applications (*i.e.*, indexing or information retrieval are typically based on nominal entities), but it can also be explained by theoretical and methodological approaches that were designed for processing nominal entities. Nevertheless, an increasing number of researchers now address the study of verbs and of their role

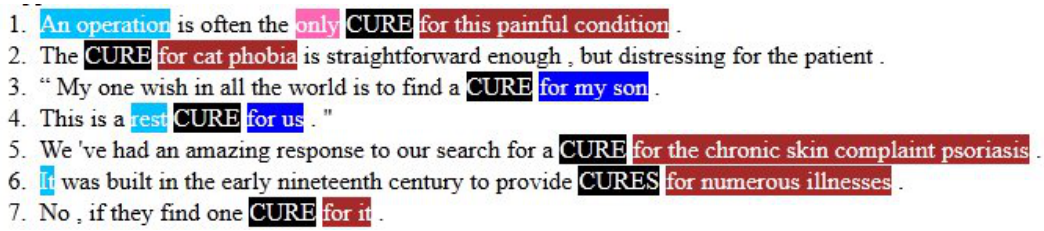


Figure 1: Example of the FrameNet annotations of the lexical unit CURE.

in specialized fields. Specific methods were developed in order to exploit verbs in terminological descriptions: in banking (Condamines, 1993), computer science (L'Homme, 1998), environment (L'Homme, 2012) and law (Lerat, 2002; Pimentel, 2011). The approaches taken by these authors differ, but they all agree on the importance of supplying a characterization of the arguments of specialized verbs. Notice also that TermoStat[1] (Drouin, 2003) can extract verbs from specialized corpora. Indeed, it has been demonstrated that verbs play an important role in Natural Language Processing (NLP) tasks, such as the detection of interactions between proteins or more generally in the extraction of semantic relations (Godbert et al., 2007; Rupp et al., 2010; Thompson et al., 2011; Miwa et al., 2012; Roberts et al., 2008).

3 Frame Semantics

The study of verbs we propose is based on Frame Semantics (FS) (Fillmore, 1982). This framework is increasingly used for the description of lexical units in different languages, mainly in English (Gildea and Jurafsky, 2002; Atkins et al., 2003; Basili et al., 2008), but it was soon extended to other languages (Padó and Pitel, 2007; Burchardt et al., 2009; Ohara, 2009; Borin et al., 2010; Koeva, 2010). Until recently, French has been neglected with regard to this framework. In addition to the description of general language, this framework can be adapted to take into account data from specialized languages (Dolbey et al., 2006; Schmidt, 2009; Pimentel, 2011). Other resources include a fine-grained characterization of the semantics and syntax of lexical units. For instance, while focussing on verbs (as opposed to FrameNet that takes into account all "frame-bearing units"), VerbNet (Palmer, 2009) implements a description of verbs and their argument structure within a similar framework.

FS puts forward the notion of "frames", which are defined as conceptual scenarios that underlie lexical realizations in language. For instance, in FrameNet (Ruppenhofer et al., 2006), the lexical database that implements the principles of FS, the frame CURE is described as a situation that comprises specific Frame Elements (FEs), (such as HEALER, AFFLICTION, PATIENT, TREATMENT, MEDICATION), and includes lexical units (LUs) such as *cure* (noun and verb), *alleviate, heal, healer, incurable, nurse, treat*.[2] In addition to the description of the frame, FrameNet provides annotations for LUs that evoke it (Figure 1).

According to our hypothesis, an FS-like modeling should allow us to describe the syntactic and semantic properties of specialized verbs and, by doing so, uncover linguistic differences observed in corpora of different levels of expertise.

4 Material

We use two kinds of material: corpora distinguished by their levels of expertise (section 4.1) and semantic resources (section 4.2), that are used for the semantic annotation of corpora.

4.1 Corpora building and processing

We study four medical corpora dealing with the specific field of cardiology. These corpora are distinguished according to their discoursive specificities and levels of expertise (Pearson, 1998). The first three corpora are collected through the CISMeF portal[3], which indexes French language medical documents and assigns them categories according to the topic they deal with (*e.g.*, cardiology, intensive care) and to their levels of expertise (*i.e.*, for medical experts, medical students or patients), the forth corpus is extracted from the

[1] *http://olst.ling.umontreal.ca/~drouinp/termostat_web/*

[2] *https://framenet.icsi.berkeley.edu/fndrupal*

[3] *http://www.cismef.org/*

Corpus	Size (occ of words)
C_1 / *expert*	1,285,665
C_2 / *student*	384,381
C_3 / *patient*	253,968
C_4 / *forum*	1,588,697

Table 1: Size of the corpora.

Doctissimo forum *Hypertension Problemes Cardiaques*[4]. The size of corpora in terms of occurrences of words is indicated in Table 1.

- C_1 or *expert* corpus contains expert documents written by medical experts for medical experts. These documents usually correspond to scientific publications and reports. They show a high level of expertise;
- C_2 or *student* corpus contains expert documents written by medical experts for medical students. These documents usually correspond to didactic support created for medical students. This corpus shows a middle level of expertise: it contains technical terms that are usually introduced and defined;
- C_3 or *patient* corpus contains non-expert documents usually written by medical experts or medical associations for patients. These documents usually correspond to patient documentation and brochures. They show a lower level of expertise: technical terms may be replaced by their non-technical equivalents and be exemplified and defined;
- C_4 or *forum* corpus contains non-expert documents written by patients for patients. This corpus contains messages from the forum indicated above. We expect the corpus to show an even lower level of expertise, although technical terms may also be used.

These corpora are used for the observation and contrastive analysis of selected verbs. C_1/C_4 and C_2/C_3 have comparable sizes.

4.2 Semantic resources

The Snomed International terminology (Côté, 1996) is structured into eleven semantic axes, which we exploit to build the resource that contains the following semantic categories of terms:

- $\mathcal{T}$: 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 (donnor)*);
- $\mathcal{P}$: Procedures (*e.g.*, *césarienne (caesarean), transducteur à ultrasons (ultrasound transducer), télé-expertise (tele-expertise)*);
- $\mathcal{L}$: Living organisms, such as bacteries and viruses (*e.g.*, *Bacillus, Enterobacter, Klebsiella, Salmonella*), but also human subjects (*e.g.*, *patients (patients), traumatisés (wounded), tu (you)*);
- $\mathcal{J}$: Professional occupations (*e.g.*, *équipe de SAMU (ambulance team), anesthésiste (anesthesiologist), assureur (insurer), magasinier (storekeeper)*);
- $\mathcal{F}$: Functions 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 (*e.g.*, *prothèses (prosthesis), tube (tube), accident (accident), cathéter (catheter)*).

Terms from these categories are exploited to semantically annotate our corpora. The only semantic category of Snomed that we ignore in this analysis contains modifiers (*e.g.*, *aigu (acute), droit (right), antérieur (anterior)*), which are meaningful only in combination with other terms. In relation to FS, we expect these categories to be indicative of frame elements (FEs), while the individual terms should correspond to lexical units (LUs). For instance, the Snomed category *Disorders* should allow us to discover and group under a

[4] *http://forum.doctissimo.fr/sante/hypertension-problemes-cardiaques/liste_sujet-1.htm*

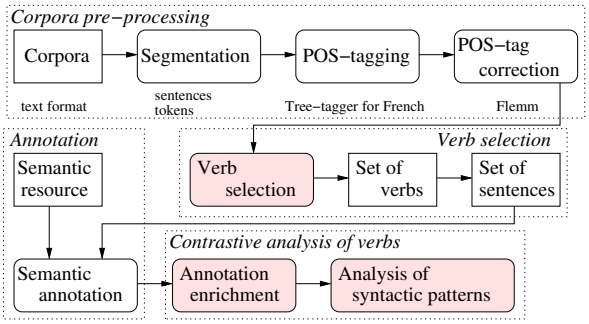


Figure 2: General schema of the method.

single label LUs (*e.g.*, *hypertension* (*hypertension*), *obésité* (*obesity*)) related to the FE DISORDER.

5 Method

The objective is first to discover the descriptions of verbs in a way compatible with FS and then to compare them. The description of verbs depends on the recognition and annotation of noun phrases, such as those provided by the Snomed terminology, which have syntactic dependencies with these verbs. The study is automated as we rely on NLP methods. The proposed method comprises four steps (Figure 2): corpora pre-processing (section 5.1), verb selection (section 5.2), semantic annotation (section 5.3), and contrastive analysis of verbs (section 5.4). On the schema, the three coloured boxes show steps that require human knowledge and that are performed manually; all the other steps are carried out automatically.

5.1 Corpora pre-processing

The corpora are all collected online and properly formatted. They are then tokenized into sentences and words: we expect this may improve POS-tagging. POS-tagging is performed with the French Tree-tagger (Schmid, 1994): its output contains words assigned to parts of speech (*e.g.*, verbs, nouns, adjectives) and lemmatized to their canonical forms (*e.g.*, singular and masculine adjectival forms, infinitive verbal forms). In order to improve the results, we check the output of the POS-tagging with the Flemm tool (Namer, 2000).

5.2 Verb selection

Sets of lemmatized verbs are extracted and their frequencies are computed in the four processed corpora. The verb selection process is carried out according to the following principles:

1. Removing forms that do not correspond to verbs:
   - POS-tagging and lemmatization errors: *e.g.*, *cardiologuer, dolipraner, rhumer*,
   - foreign words, usually also wrongly POS-tagged and lemmatized: *e.g.*, *casemixer, databaser, headacher*,
   - misspellings: *e.g.*, *souaiter, souhiter*.
2. Removing verbs which do not convey a medical meaning (*e.g.*, perception, movement, modal, state verbs);
3. Checking the meaning of the verbs in a medical dictionnary (Manuila et al., 2001): the verbs or their nominal forms have to appear in the dictionnary, as suggested in previous work (Tellier, 2008). For instance, the verb *consulter* is not recorded in the dictionnary but its nominal form *consultation* is: this verb can be then kept at this step;
4. Keeping those verbs with a frequency of 30 occurrences in the corpora. The main corpora considered are C_1 *expert* and C_4 *forum* corpora, while the other two corpora are expected to show at least 10 occurrences of the verbs. As a matter of fact, the frequency indicator is used mainly to guarantee that the verbs have a sufficient number of occurrences and appear in a high number of contexts, these showing a fair level of variability.

After the selection process, we obtain *causer* (*cause*), *traiter* (*treat*), *détecter* (*detect*), *développer* (*develop*), *doser* (*dose*) and *activer* (*activate*) among the remaining verbs. Sentences containing the selected verbs are extracted from each corpus.

5.3 Semantic annotation

The sets of sentences collected at the previous step are annotated using the Ogmios platform (Hamon and Nazarenko, 2008), which integrates and combines several NLP tools. In addition to the syntactic annotation, semantic annotation is obtained after the projection of the semantic resource described in section 4.2: the categories label the participants (that are likely to correspond to FEs), while the specific terms correspond to LUs. Thus, we assume that semantic categories provided by Snomed are useful for the description of semantic frames in medical corpora and that terms from

Step	Number
0. Raw list of verbs	6,218
1. Removing errors and foreign words	3,179
2. Removing non-medical verbs	556
3. Checking the verb meaning	47
4. Checking the frequencies	21

Table 2: Results of the verb selection at each step.

this terminology are useful for the automatic detection of relevant LUs. In a way, our approach is similar to previous work on automatic labeling of semantic roles (Gildea and Jurafsky, 2002; Padó and Pitel, 2007), although in our study we focus on specialized domain material, both corpora and resource, and we have no preconception about the semantic roles associated with medical verbs. Indeed, we exploit the entire Snomed International terminology (except the modifiers).

5.4 Contrastive analysis of verbs

The semantically annotated sentences are then analyzed manually in order to verify if the semantic roles and lexical units are correctly recognized. Wherever necessary, these annotations are enriched manually. This may apply to both missing or unrecognized LUs and FEs. Once the semantic annotation and labeling are completed, verbs from different corpora are analyzed in order to study the differences and similarities which may exist between their uses in these corpora.

6 Results and Discussion

The results are discussed along the following lines: verb selection (section 6.1), semantic annotation (section 6.2), and contrastive analysis of verbs (section 6.3).

6.1 Verb selection

Table 2 indicates the numbers of verbs selected at each step. We can see that an important number of verbs that were removed corresponds to errors, misspellings, and non-medical verbs. The subset of verbs which convey medical meanings corresponds to 0.76% (n=47) of the original set. The final subset contains 21 verbs. From this subset, we selected four verbs for a fine-grained analysis: *observer, détecter, développer*, and *activer*. These verbs were selected for two reasons: they

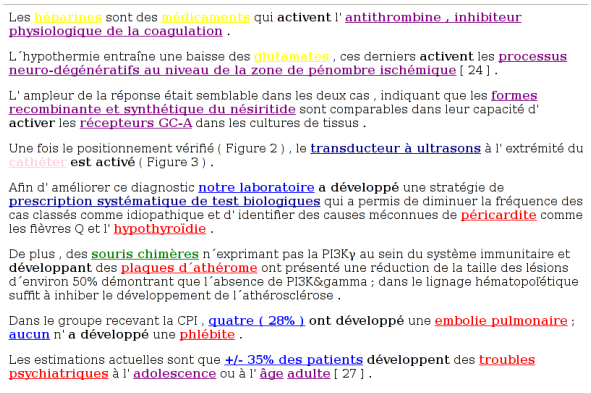


Figure 3: Examples of annotations in C_1. Verbs are in bold characters, semantic labels for arguments with different colours: DISORDERS in red, FUNCTIONS in purple, CHEMICALS in yellow, LIVING ORGANISMS in green, PHYSICAL AGENTS in pink.

were found a high number of contexts (respectively 270, 74, 193 and 85 contexts in C_1 and C_4 corpora) and these contexts seem to be diversified.

6.2 Semantic annotation

Sentences corresponding to the selected verbs have been automatically annotated with semantic classes that are indicative of FEs. The resulting annotation was checked and enriched manually: few errors are detected (*e.g.*, in English-language sentences, *or* (*où in French*) annotated as CHEMICALS (*gold*)). The main limitation is due to the incompleteness of annotations (*facteur* (*factor*) instead of *facteur V de Leiden* (*Factor V Leiden*)) and missing LUs (*e.g.*, *site d'insertion* (*insertion site*) as TOPOGRAPHY, *risque* (*risk*) as FUNCTION, *les traumatisés crâniens* (*people with brain injury*) as LIVING ORGANISMS), usually not recorded in the terminology. An example of the completed annotations is presented in Figure 3. We can observe that these annotations are evocative of those in Figure 1. In Figure 3, the verbs are in bold characters, while different FEs appear in different colours: DISORDERS in red, FUNCTIONS in purple, CHEMICALS in yellow, LIVING ORGANISMS in green, PHYSICAL AGENTS in pink. The syntactic information is also associated with the corresponding LUs but not presented in the figure. The LUs mainly correspond to nouns or noun phrases.

Another limitation discovered at this step is due to the erroneous POS-tagging. For instance, among the 32 contexts of the verb *activer* in C_4, 15 correspond to its adjectival forms (*e.g.*, *j etais une*

Verb	C_1	C_4
observer	$\mathcal{L}$, $\mathcal{J}$, $\mathcal{F}$, $\mathcal{S}$, $\mathcal{A}$, $\mathcal{D}$	$\mathcal{L}$, $\mathcal{J}$, $\mathcal{F}$, $\mathcal{A}$
détecter	$\mathcal{L}$, $\mathcal{A}$, $\mathcal{J}$, $\mathcal{P}$, $\mathcal{F}$, $\mathcal{D}$, $\mathcal{T}$	
activer	$\mathcal{C}$, $\mathcal{F}$, $\mathcal{P}$	$\mathcal{L}$, $\mathcal{P}$, $\mathcal{T}$
développer	$\mathcal{P}$, $\mathcal{D}$, $\mathcal{L}$, $\mathcal{F}$	$\mathcal{L}$, $\mathcal{D}$, $\mathcal{F}$, $\mathcal{T}$

Table 3: The most frequent arguments of verbs.

personne tres active (*I have been a very active person*), *marche active* (*active walking*)). These are not analyzed in the current study. Hence, the resulting number of contexts that were analyzed for this verbs is lower than that of the three other verbs.

6.3 Contrastive analysis of verbs

The contrastive analysis is performed manually. The most frequent labels for FEs of the four verbs analyzed appear in Table 3. We can observe for instance that LIVING ORGANISM $\mathcal{L}$ is usually the most frequent label and appears in both corpora. Typically, it corresponds to human subjects (people communicating in forum discussions in C_4, medical staff and patients observed by the medical staff in C_1). In C_1, PROCEDURES, DISORDERS and CHEMICALS also occupy an important place. Interestingly, with the verb *détecter*, the labels for FEs are identical in both corpora.

Table 4 shows the most frequent patterns of FEs with N_0 (subject) and N_1 (object) functions. We can see that some patterns are common to the two corpora studied (examples (1) to (4)). In the examples presented, the misspellings are genuine.

(1) $\mathcal{P}$ $\mathcal{D}$ with *détecter*: *j'ai acheter un tensiomètre*$_{\mathcal{P}}$ *qui détecte les anomalie cardiaque*$_{\mathcal{D}}$ (*I bought a blood pressure monitor*$_{\mathcal{P}}$ *that detects cardiac abnormality*$_{\mathcal{D}}$)

(2) $\mathcal{J}$ $\mathcal{D}$ with *détecter*: *suite a plusieurs analyses le Medecin*$_{\mathcal{J}}$ *a détecter une péricardite aigüe*$_{\mathcal{D}}$ (*after several tests the Doctor*$_{\mathcal{J}}$ *detected acute pericarditis*$_{\mathcal{D}}$)

(3) $\mathcal{D}$ as N_1 with *développer*: *Un syndrome de détresse respiratoire aiguë*$_{\mathcal{D}}$ *s'est développé* (*Acute respiratory distress syndrome*$_{\mathcal{D}}$ *appeared*)

(4) $\mathcal{D}$ $\mathcal{D}$ with *détecter*: *Une prééclampsie précoce ou sévère*$_{\mathcal{D}}$ *augmente le risque de développer une hypertension chronique*$_{\mathcal{D}}$ *et des maladies cardiovasculaires*$_{\mathcal{D}}$. (*Early or severe pre-eclampsia*$_{\mathcal{D}}$ *increases the risk to develop chronic hypertension*$_{\mathcal{D}}$ *and cardiovacular diseases*$_{\mathcal{D}}$.)

Verb	N_0	N_1	C_1	C_4
observer	$\mathcal{L}$	$\mathcal{D}$	20	3
		$\mathcal{D}$	38	1
	$\mathcal{J}$	$\mathcal{F}$	16	2
	$\mathcal{J}$	$\mathcal{D}$	4	2
détecter	$\mathcal{J}$	$\mathcal{D}$	6	39
	$\mathcal{P}$	$\mathcal{D}$	19	14
	$\mathcal{P}$	$\mathcal{F}$	2	–
	$\mathcal{J}$	$\mathcal{F}$	–	6
	$\mathcal{A}$	$\mathcal{D}$	–	2
activer	$\mathcal{L}$	$\mathcal{P}$	–	3
	$\mathcal{F}$	$\mathcal{T}$	–	2
	$\mathcal{T}$	$\mathcal{F}$	–	1
	$\mathcal{C}$	$\mathcal{F}$	3	–
	$\mathcal{F}$	$\mathcal{F}$	4	–
	$\mathcal{J}$	$\mathcal{J}$	1	–
développer	$\mathcal{L}$	$\mathcal{D}$	12	25
		$\mathcal{P}$	37	–
		$\mathcal{D}$	14	12
	$\mathcal{F}$	$\mathcal{D}$	3	4
	$\mathcal{D}$	$\mathcal{D}$	2	3
		$\mathcal{T}$	–	4

Table 4: The most frequent patterns of arguments of verbs within C_1 and C_4, with their frequencies.

On the other hand, other patterns are specific to a given corpus (examples (5) to (8)).

(5) $\mathcal{T}$ as N_1 with *développer* in C_4: *Certaines personnes réussissent à développer des branches de leurs coronaires*$_{\mathcal{T}}$ (*Some people can develop branches of their coronaries*$_{\mathcal{T}}$)

(6) $\mathcal{P}$ as N_1 with *développer*: in the *expert* corpus, a lot of PROCEDURES (*méthodes de surveillance du foetus (methods for foetus survey), stratégie diagnostique individualisée (strategies for personalized diagnosis), télémédecine (telemedicine)*) are developed with high priority within biomedical research, while this fact is missing in forum discussions

(7) $\mathcal{F}$ $\mathcal{F}$ with *activer* in C_1: *les formes recombinante et synthétique du nésiritide*$_{\mathcal{F}}$

sont comparables dans leur capacité d'activer les récepteurs GC-A$_{\mathcal{F}}$ (*recombinant and synthetic forms of nesiritide*$_{\mathcal{F}}$ *are comparable by their capacity to activate GC-A receptors*$_{\mathcal{F}}$)

(8) $\mathcal{C}$ $\mathcal{F}$ with *activer* in C_1: *Les héparines*$_{\mathcal{C}}$ *sont des médicaments*$_{\mathcal{C}}$ *qui activent l'antithrombine, inhibiteur physiologique de la coagulation*$_{\mathcal{F}}$ (*Heparine*$_{\mathcal{C}}$ *is a medication*$_{\mathcal{C}}$ *that activates antithrombin, physiological inhibitor of the coagulation*$_{\mathcal{F}}$)

Interestingly, the example (5) shows an occurrence of a different meaning of *développer* from that shown in the previous examples. Notice that we have also extracted non-medical meanings of the verbs (examples (9) and (10)), that cannot be labeled with the semantic resource we use.

(9) *Tazzy, tu peux développer ???* (*Tazzy, could you develop???*)

(10) *Santé Canada a développé une nouvelle brochure sur la déclaration des effets indésirables...* (*Health Canada designed a new brochure for the declaration of adverse reactions...*)

More generally, the verb *développer* is used in six patterns common to the two corpora, and eight and five patterns specific to C_1 and C_4 respectively, while the verb *détecter* appears in six common patterns and six specific to each of the corpora. No common pattern was identified for the verb *activer*: the syntactic and semantic properties of this verb are thus different in the two studied corpora, which may also be due to the small set of available contexts. Another difference between these two corpora is that in C_4, we can find some contexts in which verbs do not instantiate all the expected FEs: some syntactic positions remain empty.

On the whole, our observations indicate that the studied verbs present several common patterns within C_1 and C_4. This means that, in this situation, these verbs, although they have a medical meaning, can be correctly understood by patients. When the FEs are partially instantiated, differ from one corpus to the other, or when they show an important difference in terms of frequency, we assume that this may indicate situations in which the understanding may be partial or even unsuccessful. In this case, more thorough explanations are needed by patients to fully understand their health condition and required treatment.

7 Conclusion and Future work

We proposed an NLP approach to automatically discover the participants of verbs and label them using an existing medical terminology assuming that the semantic classes of the terminology are indicative of frame elements (FEs) within the framework of Frame Semantics. The study was performed with medical corpora differentiated according to their levels of expertise: high expertise in C_1 and low in C_4. The contrastive analysis of verbs was done on the basis of automatic annotations completed manually when necessary. The analysis indicates that some verbs share FEs in the studied corpora, while they usually select different FEs according to corpora.

For future work, we plan to add to this study the analysis of C_2 and C_3, which we expect may show intermediate patterns or provide a transition between C_1 and C_4. We also plan to extend this study to other verbs. Up to now, we studied verbal arguments in two syntactic positions (N_0 and N_1), which seems to suffice for the four verbs presented in this paper, but more complex patterns are likely to appear with other verbs. Moreover, automatic distinction between core FEs and non-core FEs (Hadouche et al., 2011), and between the syntactic positions of the labeled entities are other directions for future work.

Our findings may be helpful in several contexts: improving mutual understanding between medical staff and patients, creating two-fold dictionaries with expert and patient expressions, adapting the content of scientific literature for patients. This last context may also provide an interesting application and the possibility for the evaluation of the proposed analysis of verbs.

References

AMA. 1999. Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, 281(6):552–7.

S Atkins, M Rundell, and H Sato. 2003. The contribution of framenet to practical lexicography. *International Journal of Lexicography*, 16(3):333–357.

R Basili, C Giannone, and D De Cao. 2008. Learning domain-specific framenets from texts. In *ECAI Workshop on Ontology Learning and Population*.

L Borin, D Dannélls, M Forsberg, M Toporowska Gronostaj, and D Kokkinakis. 2010. The past meets the present in the swedish framenet++. In *14th EURALEX International Congress*, pages 269–281.

A Burchardt, K Erk, A Frank, A Kowalski, S Padó, and M Pinkal, 2009. *Using FrameNet for the semantic analysis of German: Annotation, representation, and automation*, pages 209–244.

A Condamines. 1993. Un exemple d'utilisation de connaissances de sémantique lexicale: acquisition semi-automatique d'un vocabulaire de spécialité. *Cahiers de lexicologie*, 62:25–65.

RA Côté, 1996. *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

AM Dolbey, M Ellsworth, and J Scheffczyk. 2006. BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. In *KR-MED*. 87-94.

P Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

C Fillmore, 1982. *Frame Semantics*, pages 111–137.

D Gildea and D Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288.

E Godbert, M Malik, and J Royauté. 2007. Analyse des formes prédicatives dans des textes biomédicaux, pour l'identification d'interactions géniques. In *JOBIM*, pages 81–86.

F Hadouche, S Desgroseilliers, J Pimentel, M.-C. L'Homme, and G Lapalme. 2011. Identification des participants de lexies prédicatives : évaluation en performance et en temps d'un système automatique. In *TIA 2011*.

T Hamon and A Nazarenko. 2008. Le développement d'une plate-forme pour l'annotation spécialisée de documents web: retour d'expérience. *TAL*, 49(2):127–154.

S Koeva. 2010. Lexicon and grammar in bulgarian framenet. In *LREC'10*.

P Lerat. 2002. Qu'est-ce que le verbe spécialisé? le cas du droit. *Cahiers de Lexicologie*, 80:201–211.

MC L'Homme. 1998. Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie*, 73(2):61–84.

MC L'Homme. 2012. Adding syntactico-semantic information to specialized dictionaries: an application of the FrameNet methodology. *Lexicographica*, 28:233–252.

L. Manuila, A. Manuila, P. Lewalle, and M. Nicoulin. 2001. *Dictionnaire médical*. Masson, Paris. 9^e édition.

A McCray. 2005. Promoting health literacy. *Journal of American Medical Informatics Association*, 12:152–163.

M Miwa, P Thompson, and S Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–65.

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

KH Ohara, 2009. *Frame-based contrastive lexical semantics in Japanese FrameNet: The case of risk and kakeru*, pages 163–182.

S Padó and G Pitel. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. In *TALN 2007*.

M Palmer. 2009. Semlink: Linking propbank, verbnet and framenet. In *GenLex-09*.

J Pearson. 1998. *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. John Benjamins, Amsterdam/Philadelphia.

J Pimentel. 2011. Description de verbes juridiques au moyen de la sémantique des cadres. In *TOTH*.

A Roberts, R Gaizauskas, M Hepple, and Y Guo. 2008. Mining clinical relationships from patient narratives. *BMC Bioinformatics*, 9(11):3–.

CJ Rupp, P Thompson, WJ Black, J McNaught, and S Ananiadou. 2010. A specialised verb lexicon as the basis of fact extraction in the biomedical domain. In *Interdisciplinary Workshop on Verbs: The Identification and Representation of Verb Features*.

J Ruppenhofer, M Ellsworth, MRL Petruck, C R. Johnson, and J Scheffczyk. 2006. Framenet ii: Extended theory and practice. Technical report, FrameNet. Available online http://framenet.icsi.berkeley.edu.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, pages 44–49, Manchester, UK.

T Schmidt, 2009. *The Kicktionary – A Multilingual Lexical Resource of Football Language*, pages 101–134.

C Tellier. 2008. Verbes spécialisés en corpus médical: une méthode de description pour la rédaction d'articles terminologiques. Technical report, Université de Montréal.

P Thompson, J McNaught, S Montemagni, N Calzolari, R del Gratta, V Lee, S Marchi, M Monachini, P Pezik, V Quochi, CJ 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.

Q Zeng-Treiler, H Kim, S Goryachev, A Keselman, L Slaugther, and CA Smith. 2007. Text characteristics of clinical reports and their implications for the readability of personal health records. In *MEDINFO*, pages 1117–1121, Brisbane, Australia.

Quand le patient devient expert : usages des termes dans les forums médicaux

Valérie Delavigne
Université Paris 3-Sorbonne nouvelle
Laboratoire EA 1483 - Recherche sur le Français contemporain (CLESTHIA)
59, rue du Forgeron
F-76190 Ecretteville les Baons
valerie.delavigne@univ-paris3.fr

Résumé

Dans le cadre d'une étude sur la circulation et la validation sociale des termes, nous souhaiterions revisiter la question de l'expertise. Dès lors que la circulation des termes s'élargit, leur signification est sujette à des négociations nouvelles. Une analyse contrastive sur un corpus de discours de vulgarisation et de forums médicaux montre une économie spécifique de ces discours, repérable par des choix lexicaux, énonciatifs et argumentatifs divers. Les forums se constituent ainsi en un espace de diffusion de terminologie médicale, mais aussi en un espace de validation, avec des logiques de légitimation spécifiques qui font évoluer le contrôle social du sens.

1 Introduction

Dans le cadre d'une étude exploratoire sur la circulation et la validation sociale des termes, nous souhaiterions revisiter la question de l'expertise. Une analyse contrastive sur un corpus de discours de vulgarisation et de forums médicaux montre une économie spécifique de chacun de ces discours, repérable par des choix lexicaux, énonciatifs et argumentatifs divers. Les forums se constituent ainsi en un espace de diffusion de terminologie médicale et en un espace de validation.

Nous nous intéressons en tant que socioterminologue (Gaudin, 2013, Delavigne 2001a) aux questions liées à la circulation des unités terminologiques dans divers discours. Des enquêtes menées sur des terrains variés : génie génétique, chimie (Gaudin, 1996), énergie nucléaire (Delavigne 2001a), astronomie (Nicolae, 2013), cancérologie (Delavigne, 2013) permettent de remettre en question la réputation de précision des vocabulaires scientifiques et techniques, opinion coutumière des sphères professionnelles et techniques où ils sont en usage ; ce préjugé est fragile et tombe vite dès lors qu'on les soumet à un examen attentif. De surcroît, dès lors que des documents techniques, scientifiques ou médicaux franchisent les portes d'un service, d'un site, d'une entreprise, d'une organisation, que deviennent les termes ?

Notre définition du terme est socioterminologique : c'est une unité lexicale dont la spécificité est à relier à son statut dans une communauté discursive donnée. Ce statut se manifeste dans le discours par des marques repérables (énoncés définitionnels, reformulations, connotations autonymiques, thématisations, etc.). *Passage* qui renvoie à une intertextualité et à des cultures, le terme ne devient tel que par décision du locuteur ou de l'analyste, qui le juge pertinent pour un savoir, un système de connaissances ou une pratique. Cette position épistémologique d'un terme comme signe caractérisé par une signification socialement normée est aisément admise aujourd'hui.

Définir le terme par son statut sociolinguistique nécessite d'envisager autrement les questions d'expertise et de validation terminologique. Ces questions ont déjà été posées, mais l'exploration de nouveaux corpus en renouvelle la problématique.

2 Statut et type d'expertise

Dans la masse des formes diversifiées produites vers l'extérieur des communautés discursives, les termes sont amenés à circuler. Dès lors, comment se joue leur négociation hors de leur terreau d'origine ? Quelles variations rencontre-t-on en fonction des genres discursifs ? des différents supports ?

Dans une communauté donnée, le consensus terminologique fonctionne. Cependant, la fréquentation de divers terrains montre une complexité qui ne se laisse pas saisir au premier abord : l'observation révèle bien des dissensus. C'est ainsi que la découverte des exoplanètes a obligé à redéfinir la notion de planète sans qu'un accord ne se fasse jour pour borner la signification du terme (Nicolae, 2013). Néanmoins, tant que les termes restent cantonnés à leur sphère organisationnelle, les terminologies jouent leur fonction. C'est ce qui permet aux communautés de se reconnaitre, d'être entre-soi : les identités culturelles des acteurs s'impriment dans la matérialité discursive (Delavigne, 2013).
De surcroit, et c'est ce qui nous intéresse ici, les choses se troublent singulièrement lorsque les terminologies doivent sortir des sphères organisationnelles et circuler hors des circuits pour lesquelles elles sont initialement prévues. C'est hors des communautés d'usagers que les problèmes se posent de façon cruciale, lorsque les terminologies sortent de leur enclave. Les choses se passent au jointif, à la frontière, dans l'entre-deux : sans voisin, il n'y a pas grand-chose à négocier. Il est avantageux d'examiner là où les terminologies se dissolvent, se désagrègent : à la surface de séparation. Dès lors, le consensus n'est plus à l'ordre du jour, des concurrences dénominatives peuvent apparaitre, des négociations discursives se font jour.

2.1 Légitimité et illégitimité de l'expertise

Quand un doute sur un mot, une notion apparait, on se tourne vers l'expert. L'expert appartient à une catégorie particulière de locuteurs auxquels on pose un problème technique dont on pense que les réponses résident dans sa discipline. C'est un locuteur *garant*.

> "La Garantie est l'instance de validation qui fonde l'évaluation des « données » : cette instance est une norme sociale qui peut être juridique, scientifique, religieuse ou simplement endoxale." (Rastier, 2011 : 54)[1]

La problématique de la garantie mène à des questions de valeur et de légitimité. D'où l'expert tire-t-il sa légitimité ? Un détour par les dictionnaires ne laisse pas d'être intéressant. Il s'avère que l'expert n'y est pas décrit simplement par celui qui sait, mais aussi par une expérience reconnue. Retenons des éléments définitoires repérés le mot *expérience*. Autrement dit, au-delà de la connaissance, du savoir, l'expert serait aussi celui qui a éprouvé une pratique ; l'expert est garant de la qualité d'un contenu cognitif, mais aussi expérientiel. C'est cette expérience qui lui confère sa légitimité.

2.2 La figure de l'expert

La figure de l'expert a été analysée par Gérard Petit sur un corpus de discours médiatiques portant sur la maladie de la vache folle. Il montre la dualité du terme qui active une définition objective en référant à une profession et souvent mêlée à des traits axiologiques. Il décrit comment le « vocable *expert* intègre un paradigme de termes parfois reformulants et coréférentiels : *chercheur, scientifique, spécialiste, chimiste, botaniste, biologiste, sociologue, épidémiologiste, médecin, microbiologiste, vétérinaire, virologiste, zoologiste, anatomopathologi(st)e*. En particulier il entre fréquemment en co-occurrence avec *scientifique*, autre figure du spécialiste. » (Petit, 2000).
Cette convocation de l'expert médiateur, socialement cautionné, en fait un « gestionnaire discursif » entre sa communauté d'origine et celui du public présumé. Cette gestion discursive se repère dans les discours par diverses marques.

2.3 La validation sociale des termes

Les rôles se voient ainsi répartis entre experts et non-experts par un pacte social. Putnam parle à ce propos de division linguistique du travail. C'est vers l'expert que se tourne le non-expert pour savoir si un terme, une définition sont justes. Des espaces de validation semblent bien circonscrits : dictionnaires, ouvrages de référence, appel à l'expertise, etc. font « loi ». C'est en outre ce clivage de la communauté linguistique qui légitime l'existence même de la vulgarisation telle qu'on l'entend d'ordinaire.
Cependant, en modifiant le corpus, les modalités discursives restent-elles les même ? Si l'on en retrouve des traces au travers des domaines différents que nous avons fréquentés : énergie nucléaire, cancérologie, corpus de vulgarisation, le support en modifie-t-il la nature ? Comment s'y joue le recours à l'expertise ? Pour donner quelques éléments de réponse à ces question, nous examinerons deux corpus ayant trait à l'information médicale.

[1] Nous renvoyons au modèle de la donnée décrit dans Rastier (2011).

3 Les corpus

Cette contribution se centrera en effet sur un terrain que nous fréquentons depuis plusieurs années, la cancérologie. Depuis les premiers états généraux sur les cancers, en 1998, les rapports à l'information médicale ont considérablement évolué et les thématiques de santé ont envahi l'espace public, dans un contexte législatif autour des droits des patients. De véritables « industries du contenu » (Romeyer, 2008) se sont spécialisées dans le domaine de la santé, proposant une offre d'information surabondante, émanant tout aussi bien d'acteurs institutionnels, d'associations de malades, d'établissements de santé, des médias, de laboratoires pharmaceutiques, de mutuelles, d'assurances, etc. La nature des informations proposées présente une forte hétérogénéité, allant de la mise à disposition d'informations médiales aux informations pratiques, en passant par des conseils divers et variés.

3.1 La plateforme Cancer Info

Cette offre s'est avérée incomplète et souvent peu fiable (Ménoret, 2007 ; Carretier *et al*., 2010). Ce constat a motivé l'émergence d'un programme visant à mettre à la disposition des personnes atteintes de cancer une information médicale validée, compréhensible et régulièrement actualisée, fondée sur des « recommandations » destinées aux professionnels de santé[2]. Ce programme, développé par la Fédération nationale des Centres de Lutte contre le Cancer, puis l'Institut national du cancer, agence sanitaire et scientifique, a pour objectif de produire des outils textuels : guides, fiches d'information, dictionnaire…, destinés à compléter et à renforcer l'information prodiguée par les équipes médicales. Ces outils sont diffusés dans les établissements de santé concernés (centres de lutte contre le cancer, hôpitaux publics et privés) et auprès d'associations de patients, et disponibles sur la plateforme Cancer Info (www.ecancer.fr).

Le contenu de ce site, en tant que site institutionnel, offre un exemple de choix pour examiner le fonctionnement d'une vulgarisation médicale spécifique : une vulgarisation institutionnelle[3]. L'ensemble des textes y est en effet « validé » par un groupe de travail composé de professionnels de santé, de patients, d'anciens patients et de proches de personnes malades, ce qui peut constituer un gage de « bonne qualité » de la vulgarisation. Première part de notre corpus, nous l'avons constitué en « corpus de référence » sur lequel peuvent s'adosser de façon contrastive les analyses de notre corpus de forums médicaux.

3.2 Les forums médicaux

Dans le panel des outils d'information sollicités par les patients, les forums médicaux prennent une place grandissante. En contournant le modèle traditionnel d'information ascendant, du médecin vers le patient, ils mettent en scène un patient qui ne recherche pas une information scientifique ou une explication, mais s'interroge sur des aspects pratiques de la maladie (Battaïa, 2012, Delavigne, 2013)

Nous reprendrons la définition que propose Marcoccia du forum de discussion, défini comme « document numérique dynamique, produit collectivement de manière interactive » (2004). Cet objet discursif interroge un certain nombre de concepts linguistiques, comme les notions de genre par exemple, ou de vulgarisation, ce que nous allons plus particulièrement explorer ici.

C'est en outre un type de discours particulier qui soulève de redoutables problèmes dès lors qu'on souhaite le soumettre à une analyse automatique. Un certain nombre de ses caractéristiques ont été mises en évidence (cf. Henri et Charlier, 2005). Notons-en seulement ici certaines : les formes discursives et sémiotiques spécifiques qu'il convoque et qui peuvent évoquer un « sous-genre », son hétérogénéité sémiotique avec ses variables d'expression de « bas niveau » : ponctèmes, emphase par la focalisation des capitales ou la graisse typographique, smileys et autres marqueurs d'interaction, son hybridation entre écrit et oral, ses particularité morphologiques et syntaxiques : néographies, accidents dactylographiques, ellipses, l'importance de certains déictiques… (Anis, 1999). Autant d'éléments qui viennent singulièrement compliquer une analyse automatique.

Notre corpus rassemble un certain nombre de discussions extraites de forums médicaux autour de la thématique du cancer : Doctissimo, Ligue contre le cancer, Jeunes Solidarité Cancer,

[2] Cet objectif est devenu une priorité de santé publique, concrétisée par les différents Plans cancer : « Donner l'accès à l'information pour que les patients qui le souhaitent puissent être acteurs dans leur combat contre la maladie » (Plan cancer 1) ; « Rendre accessible aux patients une information de référence sur les cancers afin d'en faire des acteurs du système de soins » (Plan cancer 2).

[3] Le recueil de données a pour terme juin 2012.

Atoute.org, Les Impatientes, Anamacap, France Lymphome espoir, Cancerdusein.org, sante-medecine.net, SantéAZ, aufeminin.com, e-santé.fr, Psychologies.com.

4 Aspects méthodologiques

Nous nous fondons sur l'hypothèse que les structures sociales (organisationnelles, institutionnelles ou autres) et les conditions matérielles de la communication contraignent les formes énonciatives. C'est, en suivant Rastier (2011), poser l'hypothèse de l'incidence du global sur le local.
Notre approche sociolinguistique des faits terminologiques s'inscrit dans le cadre des analyses de discours et de la sémantique de corpus. Les termes sont bien sûr à considérer non comme unités atomisées, mais au cœur de pratiques discursives variées et situées. Notre approche va vers un au-delà du terme - celui-ci étant un point d'entrée dans les discours - avec une visée discursivo-centrée.
Les analyses linguistiques des discours de vulgarisation ont bien décrit la façon dont ces textes mobilisent une intense activité de reformulation autour de certaines unités terminologiques. Par hypothèse, ces reformulations sont destinées à aider le destinataire à construire du sens. Elles peuvent être de plusieurs types : définitionnelles, désignationnelles, métaphoriques… Cette propriété est un moyen dont il est coutumier de se saisir comme moyen de dépistage des termes, ces traces formelles de cette activité reformulatrice constituant autant d'indices pour repérer des unités terminologiques (Delavigne, 2001b). Des « patrons » (Aussenac-Gilles et Condamines, 2009) permettent de localiser des « structures doubles » (Fuchs, 1982), des « paradigmes définitionnels » ou « désignationnels » (Mortureux et Petit, 1989). A partir d'un terme cristallisant autour de lui un certain nombre d'« événements discursifs », qui exhibent une énonciation en acte, il s'agit de repérer les traces de levée du « jargon » en nous attachant à examiner les lieux discursifs où il se dénoue.
Le corpus est exploité à l'aide du logiciel de traitement de données textuelles Nooj, développé par Max Silberztein[4]. Cet outil nous permet d'explorer les différentes relations autour des termes. Sans revenir sur les différentes approches outillées pour analyser le matériau langagier, disons seulement que de notre point de vue, Nooj présente l'avantage de pouvoir créer ses propres patrons de recherche sans l'effet de « boîte noire » de certains logiciels. C'est ainsi que nous avons créé des grammaires (transducteurs) pour exploiter notre corpus, ainsi qu'un dictionnaire des néographies, nécessaire pour analyser automatiquement le corpus[5].

5 Le devenir des termes

Dans les limites de cette présentation, nous nous focaliserons autour de quelques problèmes terminologiques : que deviennent les termes dans ces deux types de corpus ?
Le site Cancer Info déploie un contenu centré sur les cancers, leurs symptômes, les traitements proposés et leurs effets sur l'organisme[6]. Autour des termes se repèrent des marques de reformulation, bien décrites par ailleurs (Mortureux, 1982, 1993 ; Jacobi, 1999). Nous ne sous y attardons pas ici dans la mesure où nous nous servons de ce corpus comme instance de référence.
Les forums de santé présentent des objets de discours d'une autre nature. On y décèle des énoncés de soutien, des encouragements, des échanges de recettes (sur les moyens d'éviter les nausées provoquées par certaines chimiothérapies par exemple), des conseils pratiques, des renseignements sur les traitements ou encore de simples contacts avec d'autres malades ; il s'agit certes de comprendre sa maladie ou ses traitements, mais aussi de se rassurer, de partager… (Romeyer, 2008). On peut noter une forte densité de modalisateurs affectifs dont la fonction semble être de restaurer un lien phatique et conatif absent de ces énonciations asynchrones.

5.1 Les termes

Que deviennent alors les termes dans ces forums médicaux ? On pourrait penser dans une première approche que le discours médical s'y dissout. Or il s'avère que le forum est un lieu émaillé de terminologie médicale, reprise, commentée, évaluée, recatégorisée. Livrons-en un exemple :

[4] Le logiciel NooJ est téléchargeable gratuitement sur le site : http://www.nooj4nlp.net.

[5] Le format de cet article ne nous permet pas de décrire ces grammaires NooJ. Le dictionnaire s'appuie sur les unités que NooJ ne reconnait pas (« unknowns » dans la terminologie NooJ).

[6] Cf. *Cancer info : méthodologie d'élaboration des contenus* (www.e-cancer.fr).

> Aujourd'hui je suis allé passer le scanner de contrôle; PAS DE MÉTASTASE EN VUE ! Il y a un petit point : l'aérobilie qui causait par une communication anormale entre le tube digestif et le tractus biliaire, spontanée en cas de fistule, ou encore provoquée par une intervention chirurgicale.

Nous conservons la mise en forme et la graphie d'origine dont il faut rendre compte, ainsi que des usages spécifiques de la typographie et des ponctuèmes, extrêmement nombreux par rapport au corpus Cancer Info. On décèle une forte densité terminologique, accompagnée de traces d'intertextualité, qui marque la culture « périmédicale » des patients à l'œuvre dans ces discours (Delavigne, 2009).
Il n'y a cependant pas juxtaposition des termes dans les deux corpus. C'est ainsi qu'un terme comme le verbe *brûler*, relatif à un effet secondaire possible de la radiothérapie, est significativement présent du corpus forum (comme il l'est par ailleurs dans d'autres corpus oraux), mais absent du corpus Cancer Info.

5.2 Les reformulants

Autour de ces unités terminologiques se focalisent également des indices d'explication, de définition, de reformulation, en bref, des marqueurs de nature diverse. Examinons un extrait à propos du terme *thrombose lymphatique superficielle* :

> Alors les deux petites cordes qui me sont apparues sous le bras sont des TLS comme dans le titre du sujet. (…) ce sont les déchets non évacués par les ganglions qui ne sont plus la qui se sont thromboses au niveau des vaisseaux, ce qui empêche les mouvements et qui tire dans le bras. C'est une complication assez fréquente d'un curage ganglionnaire.

On aperçoit un certain nombre de structures définitoires typiques : « est un », « c'est un », « ce sont ». L'énoncé mobilise tout un paradigme désignationnel avec divers types de reformulants : « corde », « déchets non évacués », « complication assez fréquents ». Le terme se voit même siglé : « TLS ». Ce sont autant de traces d'une vulgarisation d'un autre type. Nous ne sommes en effet plus là dans une communication descendante du médecin vers le patient, mais dans une autre forme de vulgarisation, de pair à pair.

5.3 Les marqueurs de reformulation

L'examen d'autres extraits du corpus révèle des modalités différentes de certains marqueurs de reformulation, d'où l'exhibition vulgarisatrice est absente. Un repérage de ces marqueurs permet de mettre au jour des variantes d'usage d'un corpus à l'autre. On note ainsi moins de guillemets, moins de parenthèses. *Est un*, marqueur de reformulation présent dans les deux corpus - sans surprise -, présente des usages qui divergent nettement.
Quoique moins exhibée, la visée vulgarisatrice reste une caractéristique tout à fait prégnante du corpus forum :

> L'homéopathie est dite médecine douce mais ça ne veut pas dire sans effet ça reste un médicament !!!! Malgré ses, médicaments contre les nausées de l'homéopathie, de l'aloévera, son état se s'améliore pas

> Il arrive que l'homéopathie soulage les nausées due aux chiimios, en particulier Cocculine en doses.

Ce sont d'autres systèmes qui sont activés de façon plus présente : diaphores et autres séquences explicatives, réseau prédicatif comme la fonction.

5.4 L'émergence d'une expertise spécifique

Cette vulgarisation montre une expertise spécifique en train de se mettre en place. Cette expertise se décline avec des moyens variés. Elle laisse place notamment à une intertextualité marquée, explicite, comme dans l'extrait suivant dans lequel le locuteur fait appel à des documents de référence :

> Bonjour à tous. Alimentation et cancer Comment s'alimenter pendant les traitements ?Nous avons la réponse par la ligue voir le PDF.Affectueusement.

Ou implicite, comme celui-ci autour du terme *protocole* :

> J'ai été opéré à TOURS par le Docteur Gilgert avec un protocole ; La prise en charge thérapeutique est multidisciplinaire, elle concerne notamment : médecin généraliste, hépato-gastro-entérologue, chirurgien digestif, oncologue médical, radiologue, oncologue radiothérapeute, pathologiste, médecin nutritionniste, diététicien, gériatre, biologiste, anesthésiste,

> psychologue, personnels paramédicaux, infirmier, assistant socio-éducatif. Ce qui nous s'installe dans confiance.

On a là la convocation inexprimée d'un ailleurs textuel, en l'occurrence un extrait d'un guide pour les patients. C'est ainsi qu'on voit se mettre en place au fil des discussions une validation terminologique particulière. Ce peut être par un appel discursif à l'expert :

> Effets secondaires de la chimiothérapie. Je reprends la chimiothérapie jeudi, dès demain comme apéritif je vais boire un jus de citron pendant quatre jours pour éviter des nausées trop violentes, ce qui facilite la digestion, mon docteur en ontologie me dit que le citron est très bon, pressé plus eau et le boire chaud[7]

Ou par ce qu'on pourrait désigner comme une validation « interne » :

> J'ai un cancer du sein triple négatif... C,est un cancer qui ne répond ni à l,hormonothérapie, ni à l'herceptine que certaines ont après la chimio... Sinon, notre traitement est le même au niveau chimio...

Vulgarisation par des pairs pour des pairs, les formulations sont spontanément adaptées aux besoins des énonciateurs, ce dont témoignent certains commentaires épilinguistiques :

> Le chirurgien m'a bien expliqué tout ça et l'anesthésiste aussi mais ton témoignage est nettement plus clair

On a là l'émergence d'un patient expert qui s'adonne à une vulgarisation « de partage », facilitante pour l'appropriation terminologique. C'est ainsi qu'on voit se dessiner une évolution du modèle de la vulgarisation dans lequel se réaménagent les figures de l'interlocution. Les significations se négocient au plus près des pratiques des locuteurs. On passe en somme de la garantie issue de l'expertise à la garantie provenant de l'expérience.

6 La vulgarisation revisitée

Dès que la circulation des termes s'élargit, leur signification est sujette à des négociations nouvelles. Le rôle de la frontière est de réguler et de filtrer ; lieux de passage, les frontières sont aussi ceux où se négocie le sens. La méthode déployée sur deux corpus autour de l'information médicale permet ainsi de mettre en évidence un certain nombre de caractéristiques énonciatives autour des termes, variables selon les différents corpus. Elle montre la façon dont un lexique issu de la cancérologie se diffuse les discours des patients. Les modalités de la cooccurrence de termes et de leurs paraphrases sont certes variées ; néanmoins, l'analyse des questions discursives de vulgarisation révèlent comment les forums se constituent en un espace de diffusion de terminologie médicale, mais aussi en un espace de validation, avec des logiques de légitimation spécifiques qui font évoluer le contrôle social du sens.

Références

Jacques Anis. (Ed.). 1999. Internet, communication et langue française. Hermès. Paris.

Aussenac-Gilles, Anne Condamines. 2009. Variation syntaxique et contextuelle dans la mise au point de patrons de relations sémantiques, dans J.-L. Minel (ed.) : Filtrage sémantique. Hermes/Lavoisier, Paris :115-149.

Jean-Claude Beacco. 1993. L'explication d'orientation encyclopédique. Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires, (1):33-54. http://cediscor.revues.org/602

Jean-Claude Beacco. 2000. Écritures de la science dans les médias. Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires, (6):15-24.

Céline Battaïa. 2012. L'analyse de l'émotion dans les forums de santé. Actes de la conférence conjointe JEP-TALN-RECITAL, RECITAL :267-280.

Julien Carretier, Valérie Delavigne, Béatrice Fervers. 2010. Du langage expert au langage patient : vers une prise en compte des préférences des patients dans la démarche informationnelle entre les professionnels de santé et les patients, Sciences-Croisées :6. http://pagesperso-orange.fr/sciences.croisees

David Crystal. 2001. Language and the Internet. Cambridge University Press.

[7] Le lecteur appréciera la saveur du « docteur en ontologie » pour « oncologie »…

Fabienne Cusin-Berche & Florence Mourlhon-Dallies. 2000. Le débat autour des OGM sur Internet. Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires, 6:113–130.

Valérie Delavigne. 2013. Du vagabondage du jargon. Identités, langages et cultures d'entreprise. La cohésion dans la diversité ? 7e colloque international du GEM&L, Marseille, 21-22 mars 2013.

Valérie Delavigne. 2012. Peut-on "traduire" les mots des experts ? Un dictionnaire pour les patients atteints de cancer, Dictionnaires et traduction, Frank & Timme, Berlin :233-266.

Valérie Delavigne. 2001a. Les mots du nucléaire. Contribution socioterminologique à une analyse des discours de vulgarisation. Thèse de doctorat, Université de Rouen, 1186 p., 3 vol.

Valérie Delavigne. 2001b. « Repérage de termes dans un corpus de vulgarisation : aspects méthodologiques », Actes des quatrièmes rencontres Terminologie et Intelligence artificielle, 33-43.

François Gaudin. 1993. Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles, Presses universitaires de Rouen.

François Gaudin. 1996. Une approche sociolinguistique de la terminologie. Mémoire pour l'habilitation à diriger les recherches, URA CNRS 1164, Université de Rouen.

France Henri, Bernadette Charlier. 2005. L'analyse des forums de discussion : pour sortir de l'impasse. Symposium Symfonic, Amiens 20-22 janvier 2005. http://www.dep.u-picardie.fr/sidir/articles/index.php

Daniel Jacobi. 1999. La communication scientifique ; discours, figures, modèles. Presses Universitaires de Grenoble, coll. Médias & sociétés.

Michel Marcoccia. 2004. L'analyse conversationnelle des forums de discussion : questionnements méthodologiques, Les Carnets du Cediscor: 8:23-37.

Marie Ménoret. 2007. Les temps du cancer, Editions Le Bord de l'eau, Latresne.

Sophie Moirand. 2000. Variations discursives dans deux situations contrastées de la presse ordinaire. Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires, (6):45–62.

Marie-Françoise Mortureux. 1993. Paradigmes désignationnels. Semen, 8: 123-141.

Marie-Françoise Mortureux. 1982. Paraphrase et métalangage dans le dialogue de vulgarisation. Langue française, 53(1):48–61.

Florence Mourlhon-Dallies, Florimond Rakotonoelina, & Sandrine Reboul-Touré.2004. Les discours de l'internet : quels enjeux pour la recherche ? Les Carnets du Cediscor, 8:9-19.

Cristina Nicolae. 2013. Qu'est-ce qu'une planète? Sens et référence dans les discours scientifiques et de vulgarisation scientifique. Thèse de doctorat, Université de Rouen.

]

Building a Medical Ontology to support Information Retrieval: Terminological and metamodelization issues

Jean Charlet
AP-HP, Paris, France;
INSERM U872, Paris, France;
Jean.Charlet@upmc.fr

Gunnar Declerck
INSERM U872, Paris, France;

Ferdinand Dhombres
INSERM U872, Paris, France;
Université Pierre et Marie Curie, Paris, France;
Hôpital Armand Trousseau, AP-HP, Paris, France;

Patrick Miroux
Dpt Urgences CHU
Angers, France;

Pierre Gayet
Centre hospitalier de
Compiègne, France;

Pierre-Yves Vandenbussche
INSERM U872, Paris, France;
MONDECA, Paris, France.

Abstract

ONTOLURGENCES is a termino-ontological resource (TOR) developed for retrieving information in electronic Emergency Medical Record. This ressource describes the meaning of the relevant concepts of the field and the different labels of those concepts when they occur in documents. In this paper we show that: (i) the sustainability of such a resource requires a precise articulation between terms and concepts, and (ii) such a requirement can be met via the implementation of standardized procedures based on a meta-model architecture allowing the modeling of all necessary KOS and other knowledge structures.

1 Introduction

The use of terminological systems for the creation of ontologies raises several major issues (García-Silva et al., 2008). Obviously, ontologies and terminologies play a similar normative role. They aim at establishing a common vocabulary and make use of shared representations and concepts to allow the documents interoperability and facilitate knowledge building. However, ontologies and terminologies have clearly a different formal approach on Semantics. Ontologies are *concepts* architectures and are not organized lists of *terms*. Unlike terms, the concepts are characterized by *formal definitions*. The formal aspect enables the computerized treatment of the information. To use an ontology to normalize a document is, in that sense, to encode it by bringing a characteristic allowing the automated treatment of the information.

However, the creation of ontologies involves sometimes the use of terminologies, or even more radically, the use of corpus of text. If the ontology is to be integrated within an automated information treatment system, as for example the information retrieval (IR), the concepts should match with the terms appearing on the documents to enable the information treatment. The ontology should ensure the *coverage of the terminological domain*. The conceptual representation would otherwise be unusable.

The Lerudi (emergency services) Project intends to develop an Information System (IS) offering an overview of the Electronic Health Record (EHR) to the health professionals. Additionally, it aims at facilitating the quick reading of the EHR to allow quick medical decisions under tight time constraints. The field experimentation of that project is the reading of hospital files by an emergency regulating physician. Practically, Lerudi is IR system based on a Termino-Ontological Resource (TOR)[1] named ONTOLURGENCES. The TOR (a) plays the field model role by listing the relevant concepts; and (b) ensures the link between the concepts and their name in the EHR documents. This double function should not only enable an easy annotation and indexation of the patient files, but also facilitate the retrieval of information from the indexed records.

The ONTOLURGENCES development included 6 phases: (i) the building of the TOR ontological skeleton based on a corpus analysis method; (ii) the use of existing terminological and ontological

[1] A TOR is an ontology in which terms are linked to concepts in a systematic and exhaustive way. Several methods exist to link terms and concepts depending on the representation target (Reymonet et al., 2007).

resources to manually complete the TOR concepts system; (iii) the automatic enhancement and (iv) semi-manual TOR enhancement at the terms level; (v) the TOR enhancement of concepts in relation to the medicines; and finally, (vi) the implementation of validation and quality control procedures.

The first 2 phases of the TOR correspond to the usual ontology construction method, widely tried in our team and did not raise major issues. However, the 3 following phases that were specific to the TOR development were much more problematic. Specifically, the TOR terminological enhancement required external resources: *Knowledge Organization System* (KOS). These external resources are only useable in an architecture supporting a complex modeling of the target TOR. In particular, the architecture should accommodate the terms, the concepts and their interrelation, and simultaneously, the KOS used for the enhancement. The last stage corresponding to the quality control is also specific to this project and was necessary considering the various participants involved in the TOR construction.

Trough the detailed description of the process guiding this TOR construction and validation within a large team, we aim at showing that: (i) the sustainability of such a resource requires a concise articulation between terms and concepts; and (ii) such a requirement can be met via the implementation of standardized procedures based on a metamodel architecture allowing the modeling of all necessary KOS and other knowledge structures.

The rest of the paper is organized as follows: Section 2 briefly presents the advantages of using ontologies for retrieving information. The first two steps of the TOR construction and its specificities are presented in Section 3. Section 4 provides an overview of the UniMoKR model that enables the implementation of the TOR terminological and conceptual enhancement procedures; and its uses. Section 5 describes the TOR different enhancement phases[2]. The validation and quality control are detailed in the section 6. Finally, the paper concludes with a summary and a discussion in Section 7.

[2]Due to the lack of space, the step of conceptual enhancement of drugs branch of the ontology is not described in this paper.

2 Why using an ontology for information retrieval?

To begin with, we should ask ourselves what the point in using an ontology for an IR is. Specifically, the main advantage of an ontology is to allow an automated reasoning based on the conceptual structure and semantic relations between notions. Consequently, in addition to subsumption relations (*is-a formal relation*), we modeled the semantic relations between signs, diseases and medical specialties. These relationships enable an interface (*i.e.* a cloud of words) to display the medical specialties that characterize a given EHR.

An ontology for IR has also *de facto*, as any ontology, a structure that depends on the task (Charlet et al., 1996; van Heijst et al., 1997). This structure is not a quality in itself for the IR, but it nevertheless has two advantages: (i) a well-structured ontology is easier to maintain than a poorly structured ontology, (ii) a well-structured ontology enables valid reasonings. This second point is obviously expected from any ontology, but it is clear that it is not always satisfied. Another important property that has to possess an ontology for IR is the coverage of the terms relevant to express the notions of the target-domain. The following two examples will illustrate these points:

Example of the importance of the formal structure of the TOR. Considering the important following question that asks an emergency physician about a patient: "Has my patient already been infested by an enterobacteria in the past?". Consider that the patient's record contains a document annotated with the concept of "Salmonella". For the system to conclude that Salmonella is an enterobacteria, it is necessary that the TOR specifies that the concept "Salmonella" has a transitive relation of specialization with the concept of "enterobacteria". In this way, the answer to the question of the emergency physician will be positive, *even if the patient record is not directly annotated with the more general concept of enterobacteria*.

Example of the importance of the terminological coverage of the TOR. The annotation of the noun phrases "paracetamol", "Dafalgan" and "paraml." requires that the TOR has a unique concept representing these three syntagms

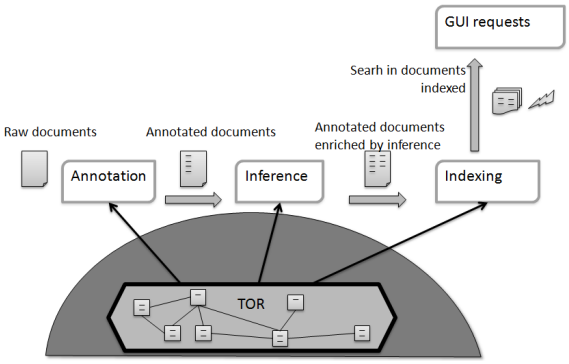


Figure 1: Uses of the TOR in the Lerudi project. The TOR supports the processes of annotation, indexing and inference.

and the availability of the terms related to the chemical molecule (paracetamol) and the proprietary drug or its brand name (Dafalgan).

It appears clearly that, the quality of the information displayed to the final user of the IR system crucially depends on the quality and richness of the TOR. The processes of annotation, indexing and inference rely on the formal structure and the terminological completeness (*i.e.* its capacity to cover the terms of the domain). The figure 1 below illustrates the different uses of the TOR in the Lerudi project.

3 Terminological and ontological resources used for designing ONTOLURGENCES

3.1 Domain of ONTOLURGENCES

ONTOLURGENCES has been built in several steps and by using different resources, and its target knowledge field has been clarified gradually. From the very beginning of the project, we realized that the knowledge field that had been originally set for the TOR (that is: the repertoire of concepts that had to be present in the TOR) had to evolve. We had left with the idea of building an ontology representing only the specific concepts used by the emergency physicians. But it turned out that, from the perspective of information retrieval in EHRs, such restriction a priori of the target knowledge field was a mistake. Indeed, the information system aims to allow the emergency physician to quickly find medically relevant concepts in EHRs. But these concepts can not be reduced to concepts specific to the medical emergency field, they can instead meet *any medical specialty*.

In the paragraphs below, we present the main phases of the development of ONTOLURGENCES and the terminological and ontological resources we have used. We do not discuss the problem of the organization of these stages and cycles of development. For this question, we may refer to (Dhombres et al., 2010).Suffice it to say that during the development process of ONTOLURGENCES, we followed the ARCHONTE method developed by B. Bachimont Bachimont et al. (2002).

3.2 The processing of textual data

In the ARCHONTE method, the domain ontology is built on the analysis of documents generated during the activity to be modeled. In our case, we have encountered great difficulties in accessing a corpus that could perform this function. The emergency services being not computerized, and the paper documents shorter and less numerous than in other services, it was difficult to find documents in sufficient numbers to make up the corpus in question.

Consequently, we used two other kinds of documents: the acts of the *Urgences* conference of the discipline and the *Guides to Good Practice*. Besides the difficulty we had to preprocess the corpus, the main problem was the coverage capacity of the corpus compared to the target. Indeed, the corpus of the conference proceedings, that was fully processed, has shown its limits in terms of scope. Conference papers are in many cases concerned with the ”rare bird”, that is with questions that are not representative of the problems that emergency physicians are confronted daily. A specific work has shown this clearly by comparing the terms most frequently detected in the corpus with the actual incidence of the emergency diseases (Gayet et al., 2010).

This issue of availability of the corpus should not be underestimated: in the areas where we can base the construction of the ontology on a corpus analyzed by tools of natural language processing (NLP), resorting to existing terminologies operates in the validation process of the work having been done. In the case of interest here, they occur much earlier in the development process.

3.3 Reusing the specialty thesaurus

For the PMSI[3] coding, the emergency physicians make use of an CIM-10 extract which contains about 1,000 terms. These terms covering an important part of the terminological repertoire used by emergency physicians for coding, it appeared necessary to incorporate them in the ontology. Consequently, a concept was created and defined for each of them.

One of the major limitation of the project is the fact that the CIM-10 terms are suitable for coding, but some of them are difficult to manage in an ontology because they encompass several heterogeneous concepts. For example, one can find terms such as "subject waiting to be admitted elsewhere, in a suitable establishment" or "Symptoms and signs involving cognitive functions and consciousness, other and unspecified". The concepts associated with such terms, because they articulate in a complex way a multitude of heterogeneous concepts, are difficult to model.

3.4 Reusing the CCAM

The french CCAM classification (commune classification of medical acts) has the benefit of having been designed by teams familiar with ontologies. Which means a priori that each concept of this classification has been validated by a formal representation (Rodrigues et al., 1999). The reuse of the CCAM thus enabled us to incorporate a classification made up in accordance with consistent principles to our TOR.

The problems rather came from the way the CCAM is organized and designations used for the acts, which are built for specified accounting policies and not at all suitable for their expression in medical documents - our target. Much work has thus consisted in renaming the terms associated with concepts (*cf.* § 3.6).

3.5 Reusing the SNOMED V3.5

The creation of the branch of diseases concepts is always a major part in the constitution of medical ontologies. As the necessary corpus for the design of such a branch were not available or did not cover the whole area, we decided to complete the work by integrating in ONTOLURGENCES the diagnoses branch of the SNOMED V3.5[4]. This procedure was mainly carried out by physicians and required more than 100 hours of work: The SNOMED V3.5 was notoriously too specific - what could be expected - but appeared also very badly organized - which was quite surprising. From the 25,000 diseases present in the The SNOMED V3.5, 6 500 have been preserved.

3.6 Additional methodological comments

To complete the description of the construction of ONTOLURGENCES, a few points of clarification are further needed:

1. ONTOLURGENCES was developed with the OWL2 description logic (DL) language and with the Protégé ontology editor;
2. The SKOS[5] language was used for the formalization of the terms. The SKOS language is representation language for knowledge organization systems such as thesauri, taxonomies, or any other type of controlled or structured vocabularies. This standard provides some primitives dedicated to the terminology with for each language, a preferred term `skos:prefLabel`, synonyms `skos:altLabel` and a definition `skos:definition`. Those primitives belonging to a standard commonly used are suitable for the representation of names and synonyms of the concepts of the ontology and can be perfectly mobilized within an ontology described in OWL.
3. The resources used in the construction of ontology are diverse. As far as possible, we memorize the origin of the concepts with an annotation that specifies the identifier of the concept in the original resource, *SnomedId*

[3] The french Information System Medicalization Program (PMSI) is intended to introduce concepts of analytic accounting in the administrative management of hospitals: diagnosis and procedures performed in a health facility are coded and recorded, reported to a patient and to the various costs in the structure.

[4] The SNOMED V3.5 is a multiaxial classification whose development has been initialized by Canadian anatomopathologists. Its aim is to represent the whole domain of medicine and related notions of society. It contains 105,000 concepts. SNOMED V3.5 exists in French and was chosen as the *reference terminology* by the French government (Rosenbloom et al., 2006). An ontology, the SNOMED-CT, has been derived from this classification by successive reorganizations and integrations of other terminologies. SNOMED-CT is not entirely available in French.

[5] The *Simple Knowledge Organization System* (SKOS) is developed within the W3C since 2003.

for SNOMED V3.5 or *FmaID* pour the FMA (*Foundational Model of Anatomy*).

4. The concepts of the ontology can be distinguished among those used for IR and the others. The latter are either high-level structuring concepts – *e.g.* IntentionalObject – or medical concepts too general to be discriminating – *e.g.* PhysicalExamination. This feature is described via a boolean annotation – terminologicalConcept – which specifies if the concept has a "terminological" character (it is potentially useful for IR) or not.

4 Meta-modelize to support enhancements

The ONTOLURGENCES ontology provides a conceptualization of the emergency field with terms to designate its concepts. This conceptualization can benefit from (i) the terms present in the KOS of Health to increase the detection of concepts in documents processed and from (ii) specific concepts about drug molecules in the ATC classification. To develop this new resource, you must be able to represent the KOS and ontology at the same level of description. Indeed, these resources are available in different formats and languages.

4.1 The UniMoKR metamodel

The diversity that exists in the nature, representation, and organization of the knowledge can be explained by different pasts, objectives, and uses. However, these KOS always intend to grasp information, to share it, and to support the human and computised processing. Thus, it is possible to extract a common model core from this obvious heterogeneity (i.e. a model common to all knowledge structuring). In the field of knowledge organization system representation, some norms and standards are in place and facilitate the interoperability (Miles, 2006; Clarke, 2008). Although SKOS and BS 8723 allow terminologies representation, none of them adress the issue of concepts group in a satisfactory manner[6]. We reuse in this project, the UniMoKR model designed in our previous work (Vandenbussche and Charlet, 2009)[7]. This model uses and extends modeling elements from SKOS, BS 8723 and is already used by research and commercial projects (Joubert et al., 2011; Vandenbussche et al., 2013).

The Termino-Conceptual part of UniMoKR model describes the relation between a *Concept* and its related *Preferred Term* and *Simple non preferred Terms* (aka synonyms) in each language. The Group Part enables not only the representation of a whole terminology, but also the representation of a terminology subset. It allows two different ways to characterize membership: by intension (concepts have to meet the restriction requirement to be part of a group; all concepts answering this request are implicitly members of the group) and by extension (concepts have to explicitly refer to this group via the relationship *inGroup*) Our modeling reified the SKOS original alignement relations and allows alignments representation generated by various sources as well as the representation of the associated metadata information. Finally, meta-classes intend to guaranty the UniMoKR model extensibility and to facilitate its re-use and adaptation: some artifacts particular to some terminologies are not taken into account in UniMoKR; however, they need to be represented to avoid the loss of information.

5 Linguistic enhancement of the ontology

As mentioned above, for the Lerudi information system to be operational in situation, it is necessary that the TOR ONTOLURGENCES covers almost all linguistic forms under which medical concepts relevant for emergency decisions appear in the EHR the system will have to deal with. Ultimately, the system must also be able to accommodate the "shortcuts" and "imperfections" of the language in which patient records are written, which for instance make use of abbreviations or may simply contain spelling errors.

The overall Lerudi system works as follows: the text of the various documents comprised in the EHR is processed by an algorithm that seeks to establish a correspondence (if necessary, by integrating NLP methods) between the phrases (treated as mere strings) and the system of concepts of the TOR. If a string has been matched with a concept, the concept will be used to index the document.

Now, medical records are usually written in natural language (or at least in this semi-standardized language suitable for concrete medical activities),

[6]For instance SKOS and BS 8723 models can not cope with SNOMED CT value sets or any concept groups defined in intension (`http://schemas.bs8723.org/`).

[7]The model is accessible at `http://bit.ly/15azC7k`

for the semantic interpretation process to reach a satisfactory level (or an optimal one: the optimum being set by the performance attained by an emergency physicist), it is often necessary to have available all lexical variations (synonyms, short forms, etc.) that may present the textual form of the concept. If a form encountered in the EHR has not been specified in the ontology, the record will not be indexed with the corresponding concept. The medical term will not be displayed by the interface. The emergency physicist will then have to put up with an incomplete or incorrect information.

To overcome this problem, two terminological enhancement processes of the TOR have been performed: (i) an automatic enhancement of the TOR by the adding of terms extracted from various KOS; (ii) a semi-automatic enhancement of the TOR by the adding of noun phrases extracted from the EHRs.

5.1 Enhancement of the TOR through the alignment with KOS

A first version of the enhanced TOR is realized through the aligment of the emergency domain ontology with few KOS relevant for the field, including CIM-10, SNOMED 3.5, MedDRA, ATC. By providing a controlled vocabulary, the KOS support the functions of analysis (annotation) of the EHRs. But, du to the difficulty to validate alignments, we decide to keep just the alignment to SNOMED V3.5. The alignment was performed with the alignment software ONAGUI (Mazuel and Charlet, 2010)[8] and by manually validating all the automatic alignments made.

Finally, during an export phase, the TOR, now optimized for annotation and indexation, is made available in the SKOS format. Once the concepts of the ontology enhanced with lexical forms from the KOS, the representation model of the TOR is converted to SKOS. This operation of conversion is performed with the model transformation method described at the section 4.1.

5.2 Enhancement of the TOR through the analysis of noun phrases

To improve the terminological completeness of ONTOLURGENCES TOR, a complementary semi-automatic enhancement procedure was introduced. This procedure incorporates the principles of the *bottom-up* methodology used by the designers of domain ontologies. It includes the following steps: (i) we first analyze with NLP tools the content of the documents produced by the operating health professionals (*i.e.* the EHRs), in order to extract (this time by mobilizing statistical methods) the noun phrases likely to be among the most structuring of the considered field of knowledge, that is the terms that are specific and essential to the field; (ii) Once these terms are identified, health professionals (emergency physicians): A) perform a filtering operation to retain only the terms actually belonging to the medical field and likely to be clinically relevant during the process of IR in EHRs and B) validate the relevance of the identified synonymous terms; (iii) these terms are then: A) added as synonyms (`skos:altLabel` tag) when they meet medical concepts already present in ONTOLURGENCES TOR, or B) converted into new concepts, when they refer to notions that do not yet have a conceptual representation in the TOR (in that specific case these terms correspond to the so-called *candidate-terms* of the *bottom-up* methodology. This conceptual conversion step requires to produce a formal definition of the concept being considered, which means firstly positioning the concept in the existing ontological hierarchy.

6 validation Processes

6.1 Why using validation procedures?

After one year of work, it appeared that the implementation of control procedures was necessary to maintain the quality of ONTOLURGENCES TOR, and that these procedures had to be replayed regularly. Indeed, (i) many stakeholders, physicians as well as modelers, are working together on the ontology, and despite all our efforts, we have not always been able to correctly apply the guidelines for the maintenance of quality and the homogeneity of the TOR. In addition, (ii) many instructions are binding and a person may apply them one day and forget them another.

In a first step, these procedures do not address the structure of the ontology. The main reason is that at this level of development of the TOR and given the skills of the team, the problems we encountered were first terminological problems. But it is clear that problems of structuration, also

[8] `http://sourceforge.net/projects/onagui/`

present, call for future treatments (*cf.* 7). Our procedures are based on patterns, or anti-patterns when managing mistakes to be avoided. This work falls under the current research area concerned with the control of the quality of ontologies, as can be read on more structural points in (Roussey et al., 2010) or (Rector et al., 2004).

6.2 Which meta-model?

The quality control procedures were designed to ensure that the TOR meets the criteria of a specific meta-model. As far as this part is concerned, the meta-model can be expressed by the list of following rules:

- Each concept must carry an annotation ¡*terminologicalConcept*¿ in Boolean format;
- Each terminological concept (*cf.* previous rule) must have one, and only one, `skos:prefLabel` in French.
- Each terminological concept must have zero or one Skos:prefLabel in another language. The other relevant languages are: English, for the communication, and Latin, widely represented in the etymology of medical concepts;
- Due to the IR algorithms functioning, two different concepts must not have the same `skos:prefLabel` or the same `skos:altLabel` (string of identical characters);

6.3 The procedures

The procedures are implemented by uploading the ontology to a SESAME store, and via SPARQL requests. Consequently, the quality criteria are verified in the *triplestore*:

- *Each terminological class must have a prefLabel*
- *Each class must have one, and only one, prefLabel in the same language.*
- *Each prefLabel must be associated with a language.*
- *Each altLabel must be associated with a language.*
- *Each class must carry a hiddenLabel.*
- *Each class must carry only one hiddenLabel in the same language.*
- *Two classes must have the same prefLabel for the same language.*
- *Two classes must not have the same altLabel for the same language.*
- *Two classes must not have the same hiddenLabel in the same language.*
- *Two classes must not have one identical prefLabel and altLabel in the same language.*
- *Tracking of multiple parent classes.* The fact that one concept has two parent concepts is not a problem in OWL. However, this pluriparentality can be symptomatic of flawed modeling. In our methodology, the ontology is first designed based on a differential approach. The double heritage appears with the implementation of the defined concepts. Two parents can be allocated to one concept as an intermediary solution before the enhancement of the modeling. However, this double heritage, either intended or not, must be tracked and listed.
- *Additional requests.* Most of the additional requests have two primary optimization purposes – (i) to standardize the frag-URI in relation to lowercase or uppercase letters, and (ii) to standardize the labels and remove, to the possible extend, the characters (such as brackets or parentheses) that could hinder proper matching.

7 Conclusion and perspectives

Lerudi is a project that applies a specific methodology to the requirements of the medical emergency environment. The goal of this project is to build a TOR capable of retrieving information efficiently. Trough the complete description of the building process and the TOR validation in a large team, we have shown that: (i) concepts and terms must be precisely articulated within such a resource; (ii) the developed meta-modeling architecture must allow the modeling of all necessary KOS and other knowledge structures; (iii) standardized procedures based on this architecture may be implemented to enable the modeling.

Finally, the integration of the KOS in the same format and the RDF transformation service (capable of operating pre-treatments) allow to generate a termino-ontological resource with a lexicalization able to carry out the annotation, inference

and indexation actions of the patients files. This project demonstrates the possibility to accommodate multiple KOS and to provide an efficient resource based on different request and transformation treatments.

References

Bruno Bachimont, Antoine Isaac, and Raphaël Troncy. 2002. Semantic Commitment for Designing Ontologies: A Proposal. In A. Gomez-Pérez and V.R. Benjamins, editors, *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, volume 2473 of *Lecture Notes in Artificial Intelligence*, pages 114–121, Sigüenza, Espagne, 1-4 Octobre. Springer Verlag.

Jean Charlet, Bruno Bachimont, Jacques Bouaud, and Pierre Zweigenbaum. 1996. Ontologie et réutilisabilité : expérience et discussion. In Nathalie Aussenac-Gilles, Philippe Laublet, and Chantal Reynaud, editors, *Acquisition et ingénierie des connaissances : tendances actuelles*, chapter 4, pages 69–87. Cepaduès-éditions.

S.G.D. Clarke. 2008. Iso 2788+ iso 5964+ much energy= iso 25964. *Bulletin of the American Society for Information Science and Technology*, 35(1):31–33.

Sylvie Després and Michel Crampe, editors. 2010. *Actes des 21es Journées Ingénierie des Connaissances*, Nîmes, France, June, 08-11,. Presse des Mines.

Ferdinand Dhombres, Jean-Marie Jouannic, Marie-Christine Jaulent, and Jean Charlet. 2010. Choix méthodologiques pour la construction d'une ontologie de domaine en médecine prénatale. In Després and Crampe (Després and Crampe, 2010).

Andrés García-Silva, Asunción Gómez-Pérez, Mari-Carmen Suárez-Figueroa, and Boris Villazón-Terrazas. 2008. A pattern based approach for re-engineering non-ontological resources into ontologies. In John Domingue and Chutiporn Anutariya, editors, *The Semantic Web*, volume 5367 of *Lecture Notes in Computer Science*, pages 167–181. Springer Berlin Heidelberg.

Pierre Gayet, Jean Charlet, L Josseran, Laurent Mazuel, and Patrick Miroux. 2010. Représentation de la médecine d'urgence dans le corpus des abstracts du congrès urgence. In *Actes du congrès URGENCES 2010*. Poster.

Michel Joubert, Tayeb Merabti, Pierre-Yves Vandenbussche, Hocine Abdoune, Badisse Dahamna, Marius Fieschi, and Stefan Darmoni. 2011. Modeling and integrating terminologies into a french multi-terminology server. In *Poster presented at MedInfo*.

Laurent Mazuel and Jean Charlet. 2010. Alignment between domain ontologies and snomed: three case studies. In Charles Safran, Heimer F. Marin, and Shane R. Reti, editors, *MEDINFO 2010 - Proceedings of the 13 th World Congress on Medical and Health Informatics - Partnerships for effective e-Health solutions*, volume 160, Cape Town, South Africa. IOS Press. Poster.

A. Miles. 2006. Skos: requirements for standardization. In *DC-2006: Proceedings of the International Conference on Dublin Core and Metadata Aplications*, pages 55–64.

Alan Rector, Nick Drummond, Matthew Horridge, Jeremy Rogers, Holger Knublauch, Robert Stevens, Hai Wang, and Chris Wroe. 2004. Owl pizzas: Practical experience of teaching owl-dl: Common errors & common patterns. In *In Proc. of EKAW 2004*, pages 63–81. Springer.

Axel Reymonet, Jérôme Thomas, and Nathalie Aussenac-Gilles. 2007. Modélisation de ressources termino-ontologiques en owl. In Francky Trichet, editor, *Journées Francophones d'Ingénierie des Connaissances (IC)*, pages 169–180, http://www.cepadues.com/, juillet. Cépaduès Editions.

Jean-Marie Rodrigues, B Trombert-Paviot, A Rector, R Baud, L Clavel, V Abrial, H Idir, and J.-M. Very. 1999. GALEN, il existe quelque chose après les mots : leur signification et au delà le savoir médical. *Innovation Stratégique en Information de Santé*, (2–3):48–62.

S Trent Rosenbloom, R A Miller, and K B Johnson. 2006. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *J Am Med Inform Assoc*, 13(3):277–88.

Catherine Roussey, François Scharffe, Oscar Corcho, and Ondrej Zamazal. 2010. Une méthode de débogage d'ontologies OWL basées sur la détection d'anti-patrons. In Després and Crampe (Després and Crampe, 2010), pages 43–54.

Gertjan van Heijst, A. Th. Schreiber, and Bob J. Wielinga. 1997. Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies*, 45(2/3):183–292.

Pierre-Yves Vandenbussche and Jean Charlet. 2009. Méta-modèle général de description de ressources terminologiques et ontologiques. In *Ingénierie de la Connaissance (IC)*.

Pierre-Yves Vandenbussche, Sylvie Cormont, Antoine Buemi, Jean Delahousse, Jean Charlet, and Eric Lepage. 2013. Implementation and management of a biomedical observation dictionary in a large healthcare information system. *J Am Med Inform Assoc*.

Session : Terminologies and ontologies

Experiments in synonymy: term extraction and mapping to concepts

Michel Généreux and **Amália Mendes**
Centro de Linguística
Universidade de Lisboa
Lisbon, Portugal
genereux@clul.ul.pt
amalia.mendes@clul.ul.pt

Thierry Hamon
LIM&BIO (EA3969)
Université Paris 13, Sorbonne Paris Cité
Paris, France
thierry.hamon@univ-paris13.fr

Abstract

We describe experiments using distributional semantics to extract and map simple terms from a corpus of genomics in Portuguese. A list of salient terms is first extracted from the corpus and manually verified by a geneticist. We then suggest a list of possible matches for the salient terms among the Unified Medical Language System (UMLS) semantic types and concepts using a Vector Space Model (VSM). Experiments show that salient terms can be extracted efficiently, while mapping them to specific concepts from a large set semantic types proved much more difficult.

1 Introduction

The meta-thesaurus UMLS is a multi-lingual collection of biomedical vocabularies which include concepts associated to Semantic Types (provided by the Semantic Network) and defined with one or more terms synonyms or term variations[1]. The thesaurus is by and large populated with terms from the English language, while other major languages such as Portuguese are lagging behind in terms of volume and quality. This work is a first step to overcome this state of affairs, by investigating how far a small to medium size corpus, combined with an automated approach based on distributional semantics and limited human intervention, can provide new terms to be included in the Portuguese section of the thesaurus (UMLS-POR). A good review of pattern-based approaches to semantic relation extraction can be found in (Auger and Barrière, 2008), while (Hamon and Grabar, 2008) offers a good example of an approach tapping on existing terminology to bring out the synonymy relations between words. Here we focus on simple terms.

[1]There is an abundant literature on term variations (Weller et al., 2011), but in this work we will treat them as synonyms.

2 Corpus

The Genomics corpus (GENOMICA) was compiled at CLUL (Centro de Linguística da Universidade de Lisboa) during 2003 and 2004. The GENOMICA corpus comprises 611 texts with 1,086,772 tokens and 66,817 words for a word/token ratio of 6.15%. The corpus has been tokenized, tagged for part-of-speeches and lemmatized using a tool we developed at CLUL. The identification of relevant texts for inclusion in the corpus was a challenge since few texts in this area of knowledge are actually produced in Portuguese. While geneticists mostly write their publications in English, in international conferences and journals, genomic knowledge has to be accessible to Portuguese speakers, laymen or students. We relied on the help of a Portuguese geneticist to locate materials in Portuguese and also to validate the texts as belonging to the area of genetics and genomics. The corpus includes scientific abstracts, papers and books, PhD dissertations, documents from the Portuguese Society for Human Genetics, excerpts from courses in genetics, support materials for students taking genetics courses, legal text and Supreme Court rulings on genomics, articles and book for scientific dissemination as well as two newspapers interviews with a Portuguese geneticist. A small sample of the corpus consists of transcriptions of spoken data: one faculty class, one conference and one interview (around 11,000

tokens). Four texts are translations to Portuguese, a total of 39,058 tokens. Given the rather modest size of our corpus, and to alleviate the problem of sparse data, we have decided to normalize terms from UMLS and GENOMICA to lower-case.

3 Experiments

These experiments are about term extraction and concept mapping. Our objective is to extract from the GENOMICA corpus salient terms pertaining to the domain of genomics and map them to concepts from the UMLS-POR. Terms extracted and mapped are validated by humans. The experimental flow-chart, including extraction and mapping, is shown in figure 1. The approaches adopted are not novel, but their combination may be.

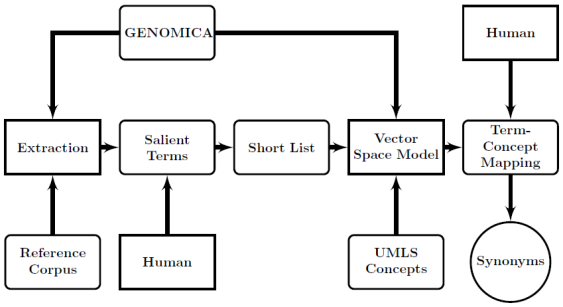


Figure 1: Experiment flow-chart

3.1 Extracting salient terms from the corpus

To extract relevant terms from GENOMICA, that is terms belonging to the field of genomics, we use an approach that computes the salience of terms based on log-odd ratio. The log-odd ratio compares the frequency of terms in a specialized corpus (GENOMICA) with that in a reference (general) corpus. High positive values indicates strong salience. We used a random subset of 10,416 files from the CRPC (Reference Corpus of Contemporary Portuguese) as reference corpus[2]. Our subset comprises 8,916,910 tokens for 175,220 words, a word/token ratio of 2.0%.

We asked our human expert, a geneticist, to review the top 2,000 terms we extracted from GENOMICA and labelled them as either "yes" (for a term definitely belonging to genomics), "maybe" or "no". The labelling task outcome was 1,403 "yes", 115 "maybe" and 482 "no", so a precision of 70% (1,403/2,000). Somewhat surprisingly, given that the UMLS-POR comprises a total of 23,350 terms distributed in 16,491 concepts, we find that only a small fraction (21%) of the terms extracted from GENOMICA and labelled "yes" are already present in UMLS-POR, which indicates that the thesaurus is in need of an update. We tackle the task of assigning some of the remaining extracted terms labelled "yes" not already in UMLS-POR to an appropriate UMLS concept in the next section.

3.2 Mapping: approach and evaluation

We undertake the task of assigning automatically new terms to UMLS-POR using a VSM approach comparing the contexts of occurrences of terms. VSMs are built around word-context matrices where rows are words (terms) to compare and columns are words used as context (context-words) for the comparison. Context-words are any words (with the exception of stopwords) neighbouring a term. Our approach is typical of VSM and can be summarized as follows: we count occurrences of context-words around each term and then compute the cosine distance between any pair of terms, assuming that words with the same contexts are closer in meaning.

Contexts are usually defined as words surrounding each term. In our case, the context can be more specifically parametrized by size and form. By *size* we mean the window size, how many words on each side of a term we are considering. By *form* we refer to either word-form or lemma. Another parameter to consider is the scoring method, how you count the number of occurrences of context-words around a term. We look at parameter tuning in the next section.

3.2.1 Parameter tuning

The first step in building matrices for VSM involves tuning parameters. We experimented with different values for three parameters: the window size, the scoring system and the form for context-words. We tested these parameters for twelve terms from five different concepts as found in the ULMS-POR (see table 1).

After building a matrix for all twelve terms summarizing their "behaviour" (scores) in GENOMICA, we computed the distances between

[2]The full corpus is available for search online at `http://alfclul.clul.ul.pt/CQPweb/`.

Terms	Concepts	English
procriação cruzamento geração	C0006159	breeds cross generation
dna adn	C0012854	deoxyribonucleotide deoxyribonucleotide
nascituro feto fetos	C0015965	unborn child fetus fetuses
gravidez gestação	C0032961	pregnant gestation
arn rna	C0035668	ribonucleic acid ribonucleic acid

Table 1: Twelve terms from five different concepts

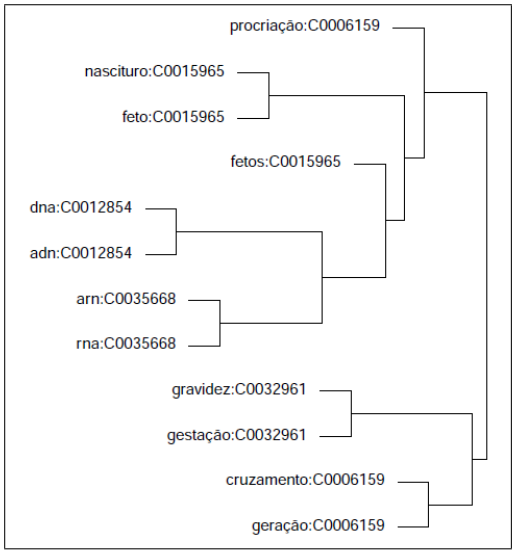


Figure 2: Clustering twelve terms (semantic distance)

each pair of terms and clustered the terms[3]. After numerous trials, we found that a window size of two (on each side of the term), counting each occurrence as 1 (without consideration for location within the window) and the use of word-form give the best result. The clustering obtained with these values for the three parameters set as above is shown in figure 2. The clusters mirror very closely the memberships between terms and concepts. The same settings are therefore used for the rest of the experiments.

It is worth mentioning that other works have also found that the immediate context of a word is more important than the distant context for determining the meaning of a word. In (Rapp, 2003), a window size of two words was used to achieve 92.5% correct on the 80 TOEFL questions.

3.2.2 Preliminary evaluation and search space reduction

Before we assign concepts to terms extracted from GENOMICA, we thought useful to get an evaluation on how effective VSM can be in assigning concepts to terms which are extracted directly from UMLS-POR, and for which we know to which concept they belong. For example, as we have seen in table 1, the concept C0006159 (*the production of animals or plants by selective pairing*) lists three synonyms in the thesaurus: *cruzamento*, *procriação* and *geração*. Each synonym may appear a certain number of times in GENOMICA. The idea is to evaluate the distance computed using VSM for synonyms but also unrelated word-pairs in relation to their frequency in GENOMICA. Table 2 presents this evaluation.

Term frequency	Synonyms	Unrelated pairs	Sample size
[2, 5]	86.5	92.8	30
[6, 40]	73.8	84.5	26
[41, ∞]	55.1	71.8	36

Table 2: Average distance between pairs of words

For example, when we compute the average cosine distance between two synonyms each with frequency (in GENOMICA) between 2 and 5, we find 86.5, while two unrelated words have a distance of 92.8. The results are consistent with the fact that semantic distance is smaller for synonyms appearing often in GENOMICA, for which a more reliable statistical model can be built.

Our objective is to compute distance between terms and concepts, so we can assign, as reliably as possible, each term a concept. Therefore, we computed the distance between 26 terms already in UMLS-POR and 329 concepts[4]. After ranking distances from lowest to highest, we found that in 38% of cases, the correct concept can be found among the top 3%. This finding will help our human expert reduce the search space while reviewing new extracted terms.

[3]Using R hierarchical clustering function hclust().

[4]The selection was made so that terms and concepts (i.e. their synonyms) appear at least 41 times in GENOMICA.

3.3 Assigning UMLS concepts to extracted terms from GENOMICA

This task is complicated by the fact that there is no guarantee that there exists a relevant concept in the thesaurus. In such cases, we should not find any strong association between the term and concepts, which in practice translates to short cosine distances as computed by the VSM. Let us also recognize that not all semantic types defined in UMLS are relevant for genomics. According to (Cohen et al., 2007), the relevant semantic types are *Gene or Genome*, *Nucleic Acid*, *Nucleoside or Nucleotide*, *Biologically Active Substance*, *Idea or Concept*, *Cell Component*, *Amino Acid*, *Peptide or Nucleotide*, *Genomic Function* and their subhierarchies[5]. If we name the semantic types mentioned previously UMLS-GENOMICS-POR, then we are left with a set 3,854 terms distributed in 2,907 concepts, compared with the full UMLS-POR (23,350 terms distributed in 16,491 concepts).

Finally, we can set out to assign concepts from UMLS-POR (UMLS-GENOMICS-POR) to extracted terms from GENOMICA which do not currently exist in the thesaurus. In the previous section we saw that a frequency of at least 41 was a good threshold for building a fairly reliable statistical model. To satisfy this constraint, we must reduce our UMLS-POR (UMLS-GENOMICS-POR) search space from 16,491 (2,907) concepts to 330 (96). Our final task is to assign the top (i.e. with the highest log-odd ratio) 141 (135) terms labelled "yes" (and with frequency > 40) by our human expert to one of the 330 (96) concepts. Our human assessor reviewed the ranking provided by our VSM and found that the correct match was often found substantially below the top 3% ranked terms for UMLS-POR (UMLS-GENOMICS-POR), limiting the usefulness of the approach and requiring a prohibitively high human expertise and revision. However, when we looked solely at the average distance calculated by VSM, we observed that genomic terms are on average 67% closer to concepts from UMLS-POR (UMLS-GENOMICS-POR) than terms outside the field of genomics. This shows that the approach is good, but yet not good enough for practical purposes.

[5](Yu et al., 1999) takes a similar restrictive views about concepts relevant to genomics in the UMLS.

4 Discussion and Conclusion

Given a reference corpus with a decent size and coverage, the extraction of salient terms from a specialized genomic corpus, albeit small, can be achieved with good precision (70%). Mapping those relevant terms to a set of predefined concepts from the UMLS thesaurus turned out to be much more difficult, as human validation required for the proposed ranking would be excessive. Given that the approach gives fairly good results on easier tasks, we believe that a larger corpus and the inclusion of methods developed elsewhere for the treatment of term variations should alleviate human intervention and render the approach more efficient. It should also be interesting to investigate in which respect heterogeneity plus scarcity of texts, but also how human intervention have impacted the results. A Portuguese version of Wordnet is also available[6], if needed. We are currently working on extracting and mapping complex terms.

References

Auger A. and C. Barrière. 2008. *Pattern-based approaches to semantic relation extraction: A state-of-the-art*. Special Issue on Pattern-based Approaches to Semantic Relation Extraction, Terminology. vol. 14, number 1, pp. 1–19.

Cohen B., Y. Chen and Y. Perl. 2007. *Updating the Genomic Component of the UMLS Semamtic Network*. AMIA Symp. Proc. 2007, pp. 150–154.

Hamon T. and N. Grabar. 2008. *Acquisition of elementary synonym relations from biological structured terminology*. Proc. of Computational Linguistics and Intelligent Text Processing - 9th Int. Conference, CICLing. Haifa, Israel, pp. 40–51.

Rapp, R. 2003. *Word sense discovery based on sense descriptor dissimilarity*. Proc. of the 9th Machine Translation Summit, pp. 315–322.

Weller, M., A. Gojun, U. Heid, B. Daille, and R. Harastani. 2011. *Simple methods for dealing with term variation and term alignment*. Proc. of the 9th International Conference on Terminology and Artificial Intelligence. Paris, France. pp. 87–93.

Yu H., C. Friedman, A. Rhzetsky and P. Kra. 1999. *Representing genomic knowledge in the UMLS semantic network*. Proc. of AMIA, pp. 181–185.

[6]`http://www.clul.ul.pt/clg/wordnetpt/index.html`

User experimentation with terminological ontologies

Louise Pram Nielsen
Department of International Business Communication
Copenhagen Business School
Denmark
lpn.ibc@cbs.dk

Abstract

This paper outlines work-in-progress research suggesting that domain-specific knowledge in terminological resources can be transferred efficiently to end-users across different levels of expertise and by means of different information modes including articles (written mode) and concept diagrams (graph mode). An experimental approach is applied in an eye-tracking laboratory, where a natural user situation is replicated for Danish professional potential end-users of a terminology and knowledge bank in a chosen pilot domain (taxation).

1 Introduction

Modern lexicography and terminology are converging (Cabré, 1999). Traditionally, terminology and lexicography have been separate research fields with different approaches to compilation and presentation of data. However, modern technology offers unlimited opportunities to meet the needs for several target groups in one database by offering the possibility of choosing between different presentations, in theory, providing means for knowledge transfer across different information modes.

Madsen and Thomsen (2008) argue that systematic terminology work ensures consistency across the entries of a given database (ISO 704:2009). This improves the quality of the information tool considerably compared to other types of specialized reference works. Thus, the end-user is presented with consistent information in a *written mode* representing the terminological information such as definitions, synonyms, equivalents and sources. In practice, however, concept clarification usually takes a more graphical starting point, in particular, when terminology and knowledge banks follow the principles of terminological ontologies (concept systems) previously discussed by Madsen and Thomsen (2008). Here terminological ontologies are defined as domain-specific ontologies where certain aspects of terminology theory have been formalized: Characteristics are modeled by formal feature specifications (attribute-value pairs) and subdivision criteria that correspond to the attributes of the feature specifications. In other words, terminologists may structure their knowledge by means of concept systems, then develop consistent definitions, and subsequently process them into article-like entries using a knowledge management tool (Madsen, 1999). This implies that the graphical structuring is central to terminology method and theory, and end-users should have access into the underlying ontology or concept system (*graph mode*) as a complementary source of knowledge.

This short paper presents experiments with the primary purpose of exploring Danish end-users' understanding of concept systems. The preliminary results imply that domain-specific knowledge in terminological resources can be transferred efficiently to users across different levels of expertise and by means of different information modes. The paper outlines work-in-progress research and will focus on the information mode variable. The paper is organized as follows: in section 2 the method is outlined; in section 3 the eye-tracking experiments are described, and in section 4 the preliminary results

suggested by the experiments are presented followed by a conclusion in section 5.

2 Method

An experimental approach is applied with Danish professional potential end-users participating in an eye-tracking experiment.

2.1 Information modes

Terminology work is concept-oriented (ISO 704:2009; Madsen, 1999), which means that synonyms are registered in one entry in the database, while lexicography is word-oriented, i.e. one dictionary entry comprises all meanings of an entry-word. With the use of databases, however, the possibilities for presentation do no longer depend on the structure of the data collection, and thus it is possible to present data from a terminological ontology with a concept-oriented structure in a word-oriented user interface. Terminology resources contain lexicographic (written) information such as definitions, sources, synonyms and equivalents, but the terminology work offers a complementary graphical information mode displaying the concept position and relations to other concepts (ISO 704:2009). Therefore, both information modes carry knowledge that can be transferred to end-users.

2.2 Experiment stimuli

Concept diagrams carry the same amount of information across eight blocks. Each block represents a taxation term: direct tax; land tax; middle-bracket tax; personal income tax; energy tax; excise duty; green tax; motor vehicles tax. The stimulus template is shown in Figure 1.

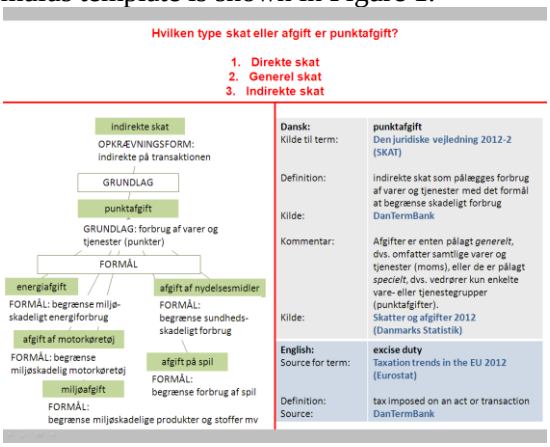


Figure 1. Stimulus template

Concept diagrams include 5-7 concepts structured in three levels filling out half of the screen. Articles are equally sized presenting more detailed information (term, definition, equivalent, and comment including sources of both the Danish source language and the English target language). The eye-tracking stimuli are thus constructed to provide participants with a double-mode design and each stimulus comprises three primary areas of interest (AOIs): the AOI-question (at the upper part of the screen); the AOI-diagram (placed below); the AOI-article (placed on the opposite side of the diagram). Figure 1 shows an example of excise duty and the question translated into English is "What type of tax or duty is excise duty? 1. Direct tax; 2. General tax; 3. Indirect tax."

In the experiment, the article entry and concept diagram of the stimuli are static images. In the real system, users should be allowed functionalities to unfold concept diagrams and articles further. The stimuli of the eye-tracking experiment can be seen as design artefacts closely resembling articles and concept diagrams of an existing knowledge management tool *i-Term* (DANTERMcentre). This approach constitutes a usual starting point of human work interaction design (Clemmensen, 2011). In addition, it is being assumed that users have entered the terminology and knowledge bank correctly and found the concept represented in the relevant diagram (graph mode) or article (written mode) necessary for concept clarification.

2.3 Question types

The experiment begins with a reading task using a specialized text in participants' first language (Danish). Then 48 multiple-choice concept-clarifying questions (trials) resulting from six types of questions about concept clarification pertaining to each of the eight chosen domain-specific terms (blocks) are posed. Questions can be answered by consulting information in either one of the information modes, or the answer lies in both: The six question types include: sub-ordinates (First diagram-based question); sub-division criteria (Second diagram-based question); equivalents (First article-based question); comments (Second article-based question); super-ordinate (First diagram- and article-based question); characteristics (Second diagram- and article-based question). The six question types are randomly distributed across the eight blocks. The eight blocks are also randomized, and so is the display-side of the information

modes (diagram to the right and article to the left side of the screen or vice versa).

3 Eye-tracking experimentation

The sample comprises 40 Danish professional potential end-users of the terminology and knowledge bank in the taxation domain.

3.1 Experimental design

An experimental approach is applied in an eye-tracking laboratory, where a natural user situation is replicated. The approach is guided by the triangulation principle resulting in both quantitative and qualitative data (Holmquist et al., 2011) that will contribute to the understanding of professional end-users' performance and perception which contribute to the subsequent interface design process, in particular, for the development of personas and scenarios (Nielsen, 2002).

Prior to the experiment, participants' domain-specific expertise is measured in a combined assessment comprising self-assessment and a test revealing their declarative knowledge in the taxation domain. In particular, participants are asked to fill out a background questionnaire comprising a declaration of consent, background information (age, gender, education, industry, typical tasks during their professional working day), introduction to concept clarification and a terminology warm-up exercise.

During the experiment, participants are asked multiple-choice questions pertaining to concept clarification in the taxation domain, while they are presented with the double-mode stimuli and their eye-movements are being recorded. A remote SensoMotoric Instrument (*SMI*) eye-tracker, which supports gaze sampling rates of 50 Hz, is used for the recordings of participants' on-screen eye-movements. The experiment is built in the psychology software *E-prime*, which facilitates randomization, records user responses, and informs participants whether they answered correctly or not.

After the experiment, a retrospective interview is conducted with the participants. Here they evaluate their performance, preference and needs pertaining to concept clarification, including their use of taxation texts, in their work. In total, the experiment lasts about one hour.

3.2 Sampling across expertise

When compiling specialized dictionaries, it is necessary to distinguish between different types of users, i.e. experts, semi-experts and laymen (Gouws, 2009). Therefore, it has been crucial in the sampling of participants for the eye-tracking experiment that they represent different levels of expertise ranging from high expert to low non-expert level. In the sample, half of participants are staff members from *The Central Customs and Tax Administration (SKAT)* working in the taxation domain as e.g. legal advisers, economists, software developers, business analysts, communicators, translators or generalists. The remaining participants are professional staff members from e.g. private companies, universities and other government organizations. All participants have Danish as their first language, and all questions and concept diagrams are in Danish.

The background questionnaire primarily assesses declarative knowledge skills in the taxation domain, whereas the eye-tracking experiments also require procedural knowledge or logical reasoning skills. Expertise variables should reflect the expertise needed in the experiments. In order to overcome any discrepancy between the declarative nature of the assessment and the procedural nature of the expertise needed, the expertise assessment also comprise participants' information seeking skills and their weekly number of electronic searches in search engines, terminological resources such as encyclopedia, dictionary or term banks.

3.3 Eye-tracking measures

In eye-tracking research, the recorded eye movements are analyzed by means of detecting events, i.e. measures accounting for scan paths (where do participants look and do they revisit AOIs) and fixation duration (what do participants look at and do they fixate on AOIs) (Holmquist et al., 2011). In particular, the eye-to-mind-hypothesis (Just and Carpenter, 1980) uses eye-movements (fixations) to indicate the cognitive effort needed to process and understand stimuli.

4 Preliminary results

Preliminary results reveal a "learning effect" which reduces the response times of participants across the 48 trials without reducing the relative number of correct answers. Moreover, high relative average fixation duration per trial in the AOI-diagram and the AOI-article on diagram and article questions respectively, suggests that

users “know” where to look for answers and can access information in the graph mode.

Potential interactions were observed during the experiments, but need further testing as part of the inferential statistical analyses:

Participants assessing their level of expertise to be high on the expertise measure (experts) are quite critical towards the stimuli. Experts have a high success rate, but they are sometimes confused by the simplified concept diagrams and article entries of this experiment and express verbally their disagreement. In addition, experts express a high preference for the detailed and precise articles compared to diagrams, which they might be confused by. Once they have learned to navigate the experiment, they start appreciating the advantages of diagrams, especially if they were new to a field, including the double-mode interface design.

Participants assessing their level of expertise to be low on the expertise measure (non-experts) have hardly any opinion on the taxation domain. Non-experts tend to be overwhelmed by the complex taxation domain and spend quite a long time understanding the questions and information modes. However, if the long response time is disregarded, non-experts perform quite well. The learning effect also applies to non-experts who learn to navigate the information modes across the experiment.

An inherent impatience is revealed during the experiments. It seems to lead participants to fuzzy scan paths and random guesses if they do not locate an answer inside the stimuli space. This impatience is due to the time pressure that participants feel they are performing under and the fact that the answer “do not know” is not available to them. Opposed to the inherent impatience, which tends to shorten the response time, an inherent insecurity tends to prolong the response time. The inherent insecurity makes participants search for answers they already know or have already found.

5 Conclusion

Despite the inherent drawbacks due to the experimental design, it can be concluded that domain-specific knowledge is transferred across written and graph modes to both experts and non-experts. It should be noted that complete descriptive and inferential statistical analyses are in progress. In addition, the eye-tracking experiments constitute a first step, which needs to be followed by future research on a dynamic system version offering participants the possibility of interacting with the article entries and concept diagrams of the terminology and knowledge bank.

Acknowledgments

The research is funded by the VELUX FOUNDATION and constitutes sub-project three of the DanTermBank project which aims at developing the foundations for the establishment of a terminology and knowledge base in Denmark.

References

Bodil Nistrup Madsen. 1999. Terminologi 1. Principper og metoder. Copenhagen, Gads Forlag.

Bodil Nistrup Madsen and Hanne Erdman Thomsen. 2008. Terminological Principles Used for Ontologies. In: Proceedings from the International Conference on Terminology and Knowledge Engineering. Managing Ontologies and Lexical Resources, pp. 107-122.

DANTERMcentre. DOI= http://www.iterm.dk/.

ISO 704 :2009. Terminology work. Principles and methods. International Organization for Standardization.

Kenneth Holmquist; Marcus Nyström; Richard Andersson; Richard Dewhurst; Halszka Jarodzka; Joost Van De Weijer. 2011. Eye tracking. A comprehensive guide to methods and measures. Oxford University Press.

Lene Nielsen. 2002. From user to character: an investigation into user-descriptions in scenarios. In Proc. DIS2002, ACM, 99-104.

M. A. Just and P. A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension, Psychological Review 87(4), 329-354.

Maria Teresa Cabré. 1999. Terminology. John Benjamins Publishing Company.

Rufus H. Gouws. 2009. Integrated dictionary use of specialised dictionaries for learners. Lexikos. 19:72-93.

Serge Verlinde, Patrick Leroyer, and Jean Binon. 2010. Search and You Will Find. From Stand-Alone Lexicographic Tools to User Driven Task and Problem-oriented Multifunctional Leximats. International Journal of Lexicography vol. 23 (1): pp. 1-17.

Thorkil Clemmensen. 2011. Designing a simple folder structure for a complex domain. Human Technology, 7(3), 216-249.

An Ontology-Driven Methodology to Reuse, Link and Merge Terminological and Language Resources

Antonio Pareja-Lora
DSIC, ILSA (Universidad Complutense de Madrid) / ATLAS (UNED)
Madrid, Spain
aplora@ucm.es

Abstract

In the last years, ontologies have proven to be a very useful way to model concepts in terminological works, even surpassing some other well-known concept modeling frameworks, such as the ones based on UML (cf. ISO/DIS 24156-1). Indeed, ontologies (Gruber, 1993; Borst, 1997) can be easily applied to link and merge different terminological and/or language resources. Towards this end, an ontological concept-model is built and placed on top of these resources, serving as a common umbrella that allows linking their terms appropriately. This paper presents some problems found when linking together a number of terminological resources (e.g. about Pragmatics) by means of ontologies. We also present the methodology followed to build these linking ontologies, and a number of recommendations that derived from their development. They helped us evaluate, reuse, link and merge these terminological resources and make them interoperate. Moreover, they have helped us identify their terminological gaps and fill them in conveniently.

1 Introduction

Pragmatics is a relatively young area of Linguistics. In effect, according to Crystal (1992), at the beginning of the nineties, 'no coherent pragmatic theory' had 'been achieved, mainly because of the variety of topics it has to account for', such as (i) speech acts (Searle, 1975); (ii) deixis, presuppositions and implicatures (Levinson, 1983; Grice 1975; 1989); or (iii) pragmatic coherence relations (Hovy and Maier, 1995; Romera, 2004; Asher and Lascarides, 2003; Prévot, 2004).

One of the first global and comprehensive views of pragmatics can be found in Yule (1996). Yet, the different theories and approaches to pragmatics are even now quite fragmentary and disconnected. Thus, the way to link the pragmatic categories derived from these theories is not obvious. Besides, most of these theories are under development and lack a proper sub-classification of the pragmatic objects and phenomena that they study (Pareja-Lora, 2012b; 2013a). They are usually based on some examples that support their assumptions, but which are insufficient to apply these theories to unrestricted texts and dialogues. Furthermore, the sub-classifications and the particularizations of these theories must be frequently re-defined *ad hoc* for each project (Pareja-Lora et al., 2013b) and are often incomplete and/or biased towards each project assumptions.

In other words, most terminological resources that account for pragmatic categories are partial, disconnected and/or poorly detailed. Thus, they require being fully subcategorized and developed, as well as being conveniently linked, in order to complement each other and overcome their fragmentation. This was the main aim of the research presented in Pareja-Lora (2012b; 2013a): the creation of a comprehensive conceptualization that linked these pragmatic categories, suitable for the pragmatic annotation of texts and dialogues. This was achieved by means of the development of some ontology modules, since they have already been successfully applied for similar purposes (Chiarcos, 2008; Buitelaar et al., 2009).

In this paper we present the methodology followed to build and link these ontology modules and other similar ones (Pareja-Lora, 2012a; 2012c; Pareja-Lora and Aguado de Cea, 2010), together with a number of recommendations and lessons learned from their development and our previous experiences in this area. They helped us evaluate, reuse and merge these terminological resources as well as make them interoperate.

Moreover, they also helped us identify their terminological gaps and decide how to fill them in systematically.

This paper is organized as follows. Section 2 details how we built the ontology module of pragmatic categories, suitable for pragmatic annotation, by reusing, merging and linking some other terminological and language resources. Section 3 presents the ontology-driven recommendations for terminological works that we identified as we built this ontology, as well as other related ones. Finally, section 4 unfolds the conclusions of this research.

2 Reusing Terminological and Language Resources for Pragmatic Annotation

As explained in the Introduction, this research originated from the need to link and to supplement several terminological and language resources dealing with pragmatics categories. Towards this end, we built some interrelated ontological modules that could give a coherent and interoperable view over them and help in their annotation. For their development, we used both the NeOn Toolkit[1] and Protégé[2], and we followed the NeOn Methodology[3] (Suárez-Figueroa et al., 2012).

In this section we detail, by means of examples, the main problems that we had to face when developing one of the ontological modules aforementioned, namely the one including pragmatic relations and units. First, we present the problems associated to linking and merging the different terminological resources that we reused. Second, we discuss how we identified and filled the resulting terminological gaps.

2.1 Linking and Merging Terminological and Language Resources

As mentioned above, this section presents how we overcame the problems that we had to face when developing our ontological module of pragmatic relations and units. In particular, the part dealing with pragmatic relations had to provide an ontological hierarchy of object properties as a result, despite object property hierarchies are not very common as yet (this is a fairly new feature of W3C/OWL 2 (2012)[4]). This section introduces the methodology we followed to build it (Subsection 2.1.1); how the top-level hierarchy of pragmatic units was generated (Subsection 2.1.2); and how we linked both hierarchies together (Subsection 2.1.3).

2.1.1 Linking and Merging the Resources with Pragmatic Relations

The top-level classes of the ontological modules and the relations that link them were fairly easy to identify and model. Firstly, we retrieved a taxonomy of discourse and pragmatic coherence relations, elaborated by Hovy and Maier (1995). This taxonomy surveyed the discourse and pragmatic coherence relations identified in the literature so far. All we had to do was to formalize the pragmatic counterpart of this taxonomy in our ontology modules. This process was quite straightforward:

i) We included two new object properties in the ontology module, namely `PragmaticRelation` and `PragmaticCoherenceRelation`.
ii) We linked them both, stating that the latter is an `rdfs:subPropertyOf` of the former.
iii) Then, we followed a similar process to formalize the rest of the taxonomy: (a) we created a new object property for each pragmatic coherence relation; and (b) we linked it to its parent object property/ies in the ontology, by means of the corresponding `rdfs:subPropertyOf` axiom(s).

Secondly, some authors refer to pragmatic coherence relations differently. For instance, the term *interpersonal coherence relation* is sometimes used instead (Hovy and Maier, 1995). So, to account for these other equivalent terms, we added a new `rdfs:subPropertyOf PragmaticRelation` to the ontology, i.e. `InterpersonalCoherenceRelation`, and added also an `owl:equivalentTo` statement, that linked the latter to `PragmaticCoherenceRelation`.

We could have simply annotated `PragmaticCoherenceRelation` with two labels, namely `"pragmatic coherence relation"@en` and `"interpersonal coherence relation"@en` instead, by means of the

[1] http://neon-toolkit.org.
[2] http://protege.stanford.edu/.
[3] We provide some motivation for choosing a methodology for ontology development in Section 3.1. The reasons why we chose this particular methodology and these tools are also mentioned there.
[4] See also http://www.w3.org/TR/owl2-primer/.

rdfs:label annotation property. However, we found it more adequate to model this relationship at the conceptual level, since (a) it allows to develop a sub-ontology with the terminology of each author and/or theory for the corresponding ontological items (objects or object properties), and link all equivalent terms and/or ontological items also at the conceptual level; (b) any change in the definition of one of these terms that turns it into a new ontological item requires simply to remove the corresponding owl:equivalentTo or owl:equivalentClass statement; or else, in the latter case, substitute it with the definition of a new suitable relation (i.e. object property); and (c) the use of the rdfs:label is fairly language-dependent. Thus, our approach is more abstract and modular, as well as theory- and language-independent. The same reasoning was applied in what follows.

Yet, the InterpersonalCoherenceRelation object property and its related term were theory-dependent. Including only this equivalent would have entailed biasing the ontology to some extent. In effect, some potential users of the ontology would not find their own way to refer to this concept in the ontology, and discard (re-)using it. So, just in case, we decided to (a) try and identify in the literature any other equivalent terms referring to this concept; and (b) model and link them analogously. Eventually, we found the term participation framework relation (Schiffrin 1987; Redeker, 1990) to be equivalent. So we added the concept Participation-RelationFramework to the ontology and linked it to the other two equivalent concepts by means of two owl:equivalentTo statements as well.

Thirdly, we had to find out other types of pragmatic relations. We used two different terminological sources towards this end: a terminological resource in the domain of linguistics, i.e. SIL's *Glossary of linguistic terms*[5] (SIL/GLT) and a Pragmatics book, i.e. Yule (1996). From SIL/GLT we extracted the terms Exophora, Homophora, and Deixis (and its subtypes); and from Yule (1996), Hedging (and its subclasses), Mitigation, and AdjacencyPairRelation (and its subtypes). Finally, we modeled them in this ontological module as well (as object properties, like pragmatic coherence relations).

[5] By *SIL International* (http://www-01.sil.org/linguistics/GlossaryOflinguisticTerms/Index.htm).

2.1.2 Linking and Merging the Resources with Pragmatic Units

Then, we had to detail the pragmatic units that are inter-related by the pragmatic relations discussed above. So we added a class hierarchy to this ontological module to formalize this type of pragmatic elements. Building this hierarchy overall was not easy, since the literature does not identify clearly what pragmatic units are or how they can be sub-classified (cf. Yule, 1996). However, we found out that

1. The different pragmatic coherence relations discussed in Hovy and Maier (1995) are signaled in text and/or dialogues by means of so-called functional units (Romera, 2004). These pragmatic functional units (PFUs henceforth) can be sub-classified according to their function, that is, the pragmatic relation that they signal.
2. Pragmatic coherence relations link together the main units of discourse, that is, macropropositions (Tannen, 1982).
3. Some authors refer to pragmatic units as pragmatemes (Mel′čuk, 2001).
4. There is yet another and most prominent type of pragmatic units, namely speech acts, that had to be included in our ontology as well.

So we included in the ontological module the PragmaticUnit class and the concepts PragmaticFunctionalUnit [6], Macroproposition, Pragmateme and SpeechAct as subclasses of the former concept. We also (a) created a sub-ontology of PFUs, quite similar to the sub-ontology of pragmatic relations identified in Hovy and Maier (1995); and (b) included the subclasses of SpeechAct traditionally identified in the literature, namely Representative, Expressive, Commissive, Directive and Declaration (Searle, 1975; Yule 1996).

2.1.3 Linking Pragmatic Units by Means of Pragmatic Relations

Finally, we had to link these pragmatic units by means of the pragmatic relations that hold among them. This was achieved by including in the ontology the corresponding rdfs:domain and rdfs:range statements. For example, we constrained (a) the domain and the range of a PragmaticCoherenceRelation to Macroproposition; and (b) the domain and the

[6] With "PFU@en" as its skos:altLabel.

range of AdjacencyPairRelation to SpeechAct.

2.2 Identifying and Filling Terminological Gaps

Up to this point, we had already developed a couple of ontological taxonomies, whose concepts were not uniformly sub-classified. Thus, we had to discover if some other concepts and/or sub-ontologies were missing and, if so, model them as well. We found out that, for example, the hierarchy of speech acts had not been sufficiently detailed for their application to pragmatic annotation yet. So it was built practically from scratch as explained below. The remaining sub-ontologies and gaps identified in our ontology were filled following the same approach.

2.2.1 Developing the Speech Act Sub-Ontology

In order to build the sub-ontology of speech acts, we followed a top-down approach, since we had at least (1) the top-level sub-classification of speech acts into representatives, expressives, commissives, directives and declarations (Searle, 1975; Yule, 1996); (2) some examples of these sub-classes of speech acts included in Yule (1996) and in SIL/GLT; and (3) a contrastive (English-Spanish) taxonomy of expressive and speech verbs[7] (henceforth, the TESV).

However, the TESV could not be reused before being previously evaluated for its suitability as for a speech act taxonomy. Some of its verbs (e.g. pronounce) did not refer to real speech acts and, besides, we had to identify other possible speech acts not included in this classification (its potential gaps). Thus, some general-purpose language resources were (re-)used towards this end (see below).

Nevertheless, evaluating our hierarchy of speech acts from the entries against the definitions of a general-purpose language resource (e.g. OALD[8]) was not easy. To start with, we found out that some of the related definitions were circular, tautological and/or inconsistent.

For instance, the sources commented above state that Command and Advice are subclasses of Directive. So, to build this part of the hierarchy, we looked up their corresponding verbs in OALD[9] and found the following relevant definitions:

- **advise**:
  1. to *tell* somebody what you think they should do in a particular situation.
- **command**:
  1. to *tell* somebody to do something.

Clearly, the *genus* of these two definitions (Faber and Mairal, 1999) is 'to tell'. This, recursively, led us to look for a relevant definition of 'tell', and we found out that it entailed a tautological definition:

- **tell**:
  1. to *order* or *advise* somebody to do something.

Indeed, 'advise' refers to 'tell', and 'tell' refers to 'advise'. Besides, when we looked up 'order' in this dictionary, we found out that it contained another circular definition involving 'tell':

- **order**:
  1. to use your position of authority to *tell* somebody to do something or say that something must happen.

Besides, in this case, the definition is also inconsistent, since 'order' might entail 'a position of authority' of the speaker that is not a real characteristic of 'tell'.

Fortunately, we could solve these problems by evaluating these terms with MWLD[10], which does not include circular and/or tautological definitions for them. Thus, we built the sub-classification of speech acts of Searle (1975) and Yule (1996) as follows:

1. We searched for the top-level performative verbs in the TESV (such as order, command, promise, thank);
2. We compiled a list of their related terms in WordNet Online[11] (Fellbaum, 1998) and in the Merriam-Webster Dictionary[12].
3. We looked up their definition in the MWLD, and then applied the algorithm presented in the next sub-section to each definition, in or-

[7] Included in the deliverables of the project "Desarrollo de una lógica léxica para la traducción asistida por ordenador a partir de una base de datos léxica inglés, alemán, francés, español, multifuncional y reutilizable" DGICYT Research Grant PB 94/0437. Spanish Ministry of Science and Education.

[8] http://oald8.oxfordlearnersdictionaries.com/ (Oxford Advanced Learner's Dictionary).

[9] Their noun entries were less informative and, besides, very often referred to the corresponding verb entry.

[10] http://www.learnersdictionary.com/ (Merriam-Webster Learner's Dictionary).

[11] http://wordnetweb.princeton.edu/perl/webwn.

[12] http://www.m-w.com.

der to ensure the completeness of the hierarchy.

2.2.2 An Algorithm to Build Terminological Gap-Free Ontologies

To develop our hierarchy, we used an algorithm that we designed following the approach in Dik (1989) to construct stepwise lexical definitions (also followed in Mairal-Usón and Periñán-Pascual (2010) to define stepwise conceptual decompositions).

Thus, we built the hierarchy of concepts for the terms we had already selected as follows. First, we took the term definitions and split them into their *genus* and *differentiae*. Second, we further split their *differentiae* into their basic components (that is, into their semantic features and/or characteristic type values (ISO//DIS 24156-1)). Third, we grouped together the terms that shared both their *genus* (*G*) and one *differentia* basic component (*DBC*). Fourth, we included a concept *C* in the ontology for *G* (if not yet present), another concept, *C'*, for the compositional semantics of *G* + *DBC*, and an `owl:subClassOf` statement that linked *C'* to *C*. Fifth, we removed the terms for *C* and *C'* from the list (if present). And sixth, we iterated this process until we had incorporated a concept for each term of the list into the ontology[13]. In addition, we attached its associated differentia basic components to each concept in the ontology, as data properties and property values – see General Recommendation 6.

3 Ontology-Driven Best Practices for Terminological Works

The hybrid terminological and ontological research works described above (as well as some other related works, not described here for space reasons[14]) allowed us to identify some ontology-driven best practices that should be applied on ontology-based terminological works. They can be classified, according to their scope, as (1) general recommendations; and (2) recommendations for terminological resource merge and link. They are discussed in a dedicated subsection below.

3.1 General Recommendations

The main general recommendations for developing ontology-based terminological works are the following:

1. **Choose a convenient ontology development methodology**, such as METHONTOLOGY (Gómez-Pérez et al., 2004) or the NeOn Methodology (Suárez-Figueroa et al., 2012). Using a methodology will help make the terminological work more systematic. Most terminologists might find the NeOn Methodology most convenient, since it identifies a particular ontology development scenario dealing with the reuse and re-engineering of non-ontological resources. However, METHONTOLOGY has also been instantiated and applied to the transformation of thesauri into ontologies (García-Torres et al., 2008).
2. **Select an ontology development tool (ODT)**, such as Protégé, WebProtégé[15] or the NeOn Toolkit, **that suits your particular needs and capabilities**. Not all ODTs are suitable for everybody and/or every ontology development. Some of them (e.g. the NeOn Toolkit) are more user-friendly than others; some others (e.g. Protégé) might be more informative and helpful for advanced users; and some others (e.g. WebProtégé) are more suitable for collaborative ontology development. In any case, ODTs that do not include sufficient and/or robust import/export functionalities should be discarded. Otherwise, the availability and maintainability of the resulting ontology will be jeopardized.
3. Concept identifiers should be as explicit as possible. Therefore, **whenever possible, use widespread terms as concept identifiers, that is, to name concepts**. This will help the future users of the ontological resource evaluate and reuse it. If no suitable term is available to name a concept, create a new term to name it. For this, use the name of its parent in the ontology (e.g. the name of the concept for its *genus*), preceded by its corresponding characteristic type value, as detailed in the *differentia* of its associated definition.
4. **Use pervasively the `rdfs:label` property to annotate the concepts of the ontology**

[13] In this case, synonyms were discarded. However, in other cases, a new step must be added after the fifth step, in order to add the necessary `owl:equivalentClass` or `owl:equivalentTo` statements.

[14] E.g. Pareja-Lora (2012a; 2012c) and Pareja-Lora and Aguado de Cea (2010).

[15] http://webprotege.stanford.edu/.

with the term(s) that designate it (Declerck, and Gromann, 2012)[16].

5. **Whenever possible, use the `rdfs:isDefinedBy` annotation property to provide a suitable definition for your concepts and/or of the terms that label them**. *When available, refer to a standard definition of the concept and/or the labelling term*, such as the ones provided in ISO standards, and often included in the ISO Concept Database[17]. Other additional but interesting details (e.g. examples of use of the term in context) can be supplied by means of the other RDFS predefined annotation properties, namely `rdfs:comment` and `rdfs:seeAlso`.
6. **In order to ensure the completeness of your conceptual model, identify the *genus* and the *differentia* in your `rdfs:isDefinedBy` annotations for concepts** (Faber and Mairal, 1999) and, then,
   a. check that the concept formalizing the *genus* is the parent in the ontology of the concept being defined;
   b. find other similar *differentia* values that might sub-classify the *genus* concept into more specific concepts[18], and check that all the possible children of the *genus* concept have already been formalized in the ontology (formalizing them if not yet present);
   c. identify an XSD datatype that represents the *differentia* values found (`xsd:boolean`, for example), or else create a new one that includes them all;
   d. define a suitable object property that has the *genus* concept as domain and this datatype as range[19];
   e. For each children of the *genus* concept: assign to the object property the corresponding value of the *differentia* (one of the values included in the datatype);

   This will also help you build your ontology with stepwise concept definitions and/or decompositions (*cf.* Dik, 1989; Mairal-Usón and Periñán-Pascual, 2010) as a basis. It will make the meaning of your concepts more explicit and help inference engines when reasoning on your ontology or querying it.
7. **Taxonomise as much as possible**. This criterion, when associated to the previous one, helps attach to each concept only its particular properties and inherit the properties and property values of its ancestors in the ontology taxonomy. Hence, it avoids redundancy and allows for property share and reuse.

3.2 Recommendations for Terminological Resource Merge

When merging different terminological resources by means of an ontology,

1. **Follow a conceptual, comprehensive, general and/or eclectic approach**. Even though the terms are different, the concept that they designate might be the same or very similar. This is most frequent when the terminological resources capture the terms of different theoretical frameworks and/or approaches in a given domain. In this case, you can
   a. **Select a preferred term among the alternative equivalent terms, as done in the development of thesauri** (Gil-Urdiciain, 2004), that is,
      i. decide which is the most well-know or the most widespread term;
      ii. use it to name the concept in the ontology.
   b. **Or else, create a new theory-neutral identifier to refer to this concept in the ontology.** This second approach might be more obscure, but it might help when having to choose is complex and/or not politically correct. In this case, follow the indications in General Recommendation 3.
2. **Link within the ontology all the terms that designate the same concepts in different terminological resources**, be they for the same language or not, by means of `owl:equivalentClass` and `owl:equivalentTo` statements, respectively, for classes and for object properties. This will establish a relationship between your resources at the conceptual level, not at the lexical level. Accordingly, your resources will be more abstract, modular and, in general, more general and robust.

[16] For a most interesting discussion on the difference between (i) concepts, (ii) their names and/or identifiers in an ontology and (iii) the lexical items or the terms commonly used to refer to these concepts, see L'Homme and Bernier-Colborne (2012).

[17] https://www.iso.org/obp/ui/.

[18] If they can be found, then the *differentia* is, in fact, a value of a terminological characteristic or criterion of subdivision.

[19] By means of an `rdfs:domain` statement or an `rdfs:range` statement (as applicable).

3. **Fill in the conceptual and/or terminological gaps you identify, including as many concepts in your ontology as required** to fill in these gaps, preferably following General Recommendation 6.
4. The `rdfs:label` property allows specifying the language to which a term belongs. Therefore, **when merging several multilingual terminological resources, (i) use as many `rdfs:label` statements as required to annotate the concepts in the ontology with all the languages being covered by your resources; and (ii) indicate the language for these terms in the `rdfs:label` statements correspondingly** (Declerck, and Gromann, 2012)[20].

4 Conclusions

In this paper, we have presented how we built a set of ontology modules that include pragmatic categories. These ontology modules have been built on several other terminological and language resources, in order to overcome the biases, disconnections and lack of details and subclassification that this area shows altogether. In particular, our ontological modules provide a comprehensive abstract model over the terminology of pragmatics and pragmatic categories, which has helped link, merge, complement and extend these different resources, and make them interoperate.

We have also shown the methodology followed to build these ontology modules. In particular, we have shown the usefulness of object property hierarchies in this type of terminological works and how they can be easily created. Our methodology includes also a set of recommendations and lessons learned so far when modeling this and other ontology-based terminological resources in the linguistic domain.

These recommendations detail how to build hierarchies of concepts, which constitute the backbone of conceptual models in terminology works and/or in ontologies. As ontological terminology works become more frequent, these recommendations will become more necessary and useful to (a) reuse, link and merge other language or terminological resources; and (b) build them from scratch. They describe also how these hierarchies should be systematically evaluated in order to identify terminological gaps and fill them. Besides, they provide some useful hints to name and annotate ontology concepts with their designating terms and definitions, both in monolingual and multilingual environments. Overall, they provide a sound basis for ontologies to be used as terminology managers, which is one of the possible ways in which ontological and terminological research works might converge in the future.

[20] Alternatively, if more information about the term (e.g. its (morpho-)syntactic, lexical and/or combinatorial properties) should be provided, you can associate a **lemon**-based lexicon (http://lemon-model.net/) to your ontology(ies).

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press. Cambridge. UK.

Willem N. Borst. 1997. *Construction of Engineering Ontologies*. PhD thesis. University of Twente. Enschede. Netherlands.

Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. 2009. Towards Linguistically Grounded Ontologies. In *The Semantic Web: Research and Applications* (*Lecture Notes in Computer Science, Vol. 5554/2009*). Berlin / Heidelberg: Springer.

Chiarcos, Christian. 2008. An Ontology of Linguistic Annotations. In *LDV Forum (GLDV-Journal for Computational Linguistics and Language Technology)*, 23(1):1-16.

David Crystal. 1992. *A Dictionary of Linguistics and Phonetics* (3rd edition). Oxford: Blackwell.

Thierry Declerck and Dagmar Gromann. 2012. Towards the Generation of Semantically Enriched Multilingual Components of Ontology Labels. In *Proceedings of the 3rd Workshop on the Multilingual Semantic Web (MSW3)*. Boston, USA, November 2012.

Simon C. Dik. 1989. *The Theory of Functional Grammar. Part I: The Structure of the Clause*. Dordrecht: Foris Publications.

Pamela Faber and Ricardo Mairal. 1999. *Constructing a lexicon of English verbs*. Berlin: Mouton de Gruyter.

Christiane Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Alberto García-Torres, Antonio Pareja-Lora and Daniel Pradana-López. 2008. Reutilización de tesauros: el documentalista frente al reto de la web semántica. In *El Profesional de la Información*, 2008, January-February, vol. 17 (1), pp. 8-21.

Blanca Gil-Urdiciain. 2004. *Manual de lenguajes documentales*. Gijón: TREA.

Asunción Gómez-Pérez, Mariano Fernández-López and Ócar Corcho. 2004. *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web*. London: Springer-Verlag.

Herbert P. Grice. 1975 (1989). Logic and conversation. *Ibid.* Reprinted in *Studies in the Way of Words* (H. P. Grice, ed.), 22–40. Cambridge, MA: Harvard University Press.

Thomas R. Gruber. 1993. A Translation Approach to Portable Ontologies. In *Journal on Knowledge Acquisition*, Vol. 5(2): 199-220.

Eduard Hovy and Elisabeth Maier. 1995. Parsimonious or Profligate: How Many and Which Discourse Structure Relations? [http://www.isi.edu/natural-language/people/hovy/papers/93discproc.pdf].

International Organization for Standardization. 2013. Graphic notations for concept modeling in terminology work and its relationship with UML – Part 1: Guidelines for using UML and mind-mapping notation in terminology work (ISO/DIS 24156-1).

Marie-Claude L'Homme and Gabriel Bernier-Colborne. 2012. Terms as labels for concepts, terms as lexical units: A comparative analysis in ontologies and specialized dictionaries. In *Applied Ontology*, 2012, January, vol. 7 (4), pp. 387-400.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Ricardo Mairal-Usón and Carlos Periñán-Pascual. 2010. Role and Reference Grammar and Ontological Engineering. In *Los caminos de la lengua: Estudios en homenaje a Enrique Alcaraz Varó*. Alicante: Universidad de Alicante, Servicio de Publicaciones, pp. 649-666.

Igor Aleksandrovič Mel′čuk. 2001. *Communicative Organization in Natural Language: the Semantic-Communicative Structure of Sentences*. Amsterdam, Philadelphia: John Benjamins Publishing.

Antonio Pareja-Lora. 2012a. *Providing Linked Linguistic and Semantic Web Annotations – The OntoTag Hybrid Annotation Model*. Saarbrücken: LAP – LAMBERT Academic Publishing.

Antonio Pareja-Lora. 2012b. OntoLingAnnot's Ontologies: Facilitating Interoperable Linguistic Annotations (Up to the Pragmatic Level). In *Linked Data in Linguistics. Representing Language Data and Metadata*. Heidelberg: Springer, pp. 117-127.

Antonio Pareja-Lora. 2012c. OntoLingAnnot's LRO: An Ontology of Linguistic Relations. In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)*. Madrid, June 2012, pp. 49-64. [http://www.oeg-upm.net/tke2012/proceedings, paper 04]

Antonio Pareja-Lora. 2013a. The pragmatic level of OntoLingAnnot's ontologies and their use in pragmatic annotation for language teaching. In J. Arús, M.E., Bárcena, and T. Read (eds.) *Languages for Special Purposes in the Digital Era*. Springer [IN PRESS].

Antonio Pareja-Lora, María Blume and Barbara Lust. 2013b. Transforming the Data Transcription and Analysis Tool Metadata and Labels into a Linguistic Linked Open Data Cloud Resource. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013)*. Pisa, September 2013 [IN PRESS].

Antonio Pareja-Lora and Guadalupe Aguado de Cea. 2010. Modelling Discourse-related terminology in OntoLingAnnot's ontologies. In *Proceedings of the TKE 2010 workshop "Establishing and using ontologies as a basis for terminological and knowledge engineering resources"*. Dublin, August 2010.

Laurent Prévot. 2004. *Structures sémantiques et pragmatiques pour la modélisation de la cohérence dans des dialogues finalisés*. Thèse de doctorat de l'université Paul Sabatier. Toulouse. France.

Gisela Redeker. 1990. Ideational and pragmatic markers of discourse structure. In *Journal of Pragmatics*, 14: 367-381.

Magdalena Romera. 2004. *Discourse Functional Units: the Expression of Coherence Relations in Spoken Spanish*. Munich: LINCOM.

Deborah Schiffrin. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

John Searle. 1975. Indirect speech acts. In *Syntax and Semantics*, 3: Speech Acts (P. Cole and J. L. Morgan, eds.): 59–82. Academic Press. New York. USA.

Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez and Mariano Fernández-López. 2012. The NeOn Methodology for Ontology Engineering. In M.C. Suárez-Figueroa et al. (eds.) *Ontology Engineering in a Networked World*. Berlin Heidelberg, Springer-Verlag.

Deborah Tannen. 1982. *Analyzing Discourse: Text and Talk*. Georgetown: Georgetown University Press.

W3C. 2012. OWL 2 Web Ontology Language – Document Overview (Second Edition) [http://www.w3.org/TR/owl2-overview/].

George Yule. 1996. *Pragmatics*. Oxford University Press. Oxford. UK.

Author table