**LREC 2018 Workshop**

# GLOBALEX 2018
# Lexicography & WordNets

# PROCEEDINGS

Edited by

Ilan Kernerman, Simon Krek

ISBN: 979-10-95546-28-3
EAN: 9 791095 546283

8 May 2018

Proceedings of the LREC 2018 Workshop
Globalex 2018: Lexicography & WordNets

8 May 2018 – Miyazaki, Japan

Edited by Ilan Kernerman, Simon Krek

https://globalex.link/globalex2018/

GLOBALEX 2018 is jointly organized by:

**GLOBALEX Preparatory Board**
http://globalex.link/

**Global WordNet Association (GWA)**
http://globalwordnet.org/

**European Lexicographic Infrastructure – ELEXIS**
http://www.elex.is/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

# Organising Committee

- Francis Bond, GWA; Nanyang Technological University
- Ilan Kernerman, GLOBALEX & ELEXIS; K Dictionaries (co-chair)
- Simon Krek, GLOBALEX & ELEXIS; "Jožef Stefan" Institute (co-chair)
- Luis Morgado da Costa, GWA; Nanyang Technological University

# Programme Committee

| | |
|---|---|
| • Michael Adams | Indiana University |
| • Michal Boleslav Méchura | Masaryk University |
| • Francis Bond | Nanyang Technological University |
| • Sonja Bosch | University of South Africa |
| • Julia Bosque-Gil | Madrid Polytechnic University |
| • Philipp Cimiano | University of Bielefeld |
| • Thierry Declerck | Austrian Academy of Sciences and DFKI |
| • Gerard de Melo | Rutgers University |
| • Anne Dykstra | Fryske Akademy |
| • Thierry Fontenelle | Translation Centre for the Bodies of the EU |
| • Alexander Geyken | Berlin-Brandenburg Academy of Sciences and Humanities |
| • Rufus Gouws | Stellenbosch University |
| • Jorge Gracia | University of Zaragosa |
| • Orin Hargraves | University of Colorado |
| • Kris Heylen | KU Leuven |
| • Aleš Horák | Masaryk University |
| • Hitoshi Isahara | Toyohashi University of Technology |
| • Miloš Jakubiček | Lexical Computing |
| • Jelena Kallas | Institute of the Estonian Language (EKI) |
| • Diptesh Kanojia | IIT Bombay |
| • Ilan Kernerman | K Dictionaries |

# Programme Committee

- Annette Klosa — German Language Institute (IDS)
- Iztok Kosem — University of Ljubljana
- Simon Krek — "Jožef Stefan" Institute
- Lothar Lemnitzer — Berlin-Brandenburg Academy of Sciences and Humanities
- Robert Lew — Adam Mickiewicz University
- Nikola Ljubešić — University of Zagreb
- Stella Markantonatou — Institute for Language and Speech Processing (ATHENA)
- John Philip McCrae — National University of Ireland, Galway
- Erin McKean — Wordnik
- Julia Miller — University of Adelaide
- Verginica Mititelu — Romanian Academy Research Institute for Artificial Intelligence
- Antoni Oliver — Open University of Catalunia
- Vincent Ooi — National University of Singapore
- Danie Prinsloo — University of Pretoria
- Ewa Rudnicka — Wroclaw University of Technology
- Michael Rundell — Lexicography Masterclass and Macmillan Dictionaries
- Klaas Ruppel — Institute for the Languages of Finland (KOTUS)
- Kevin Scannell — Saint Louis University
- Egon Stemle — European Academy of Bozen/Bolzano
- Carole Tiberius — Institute for Dutch Language (INT)
- Yukio Tono — Tokyo University of Foreign Studies
- Lars Trap Jensen — Society for Danish Language and Literature (DSL)
- Tamás Váradi — Hungarian Academy of Sciences
- Elena Volodina — Gothenburg University
- Shigeru Yamada — Waseda University

# Preface

GLOBALEX 2018 Workshop follows on the first GLOBALEX Workshop at LREC 2016. It is organized jointly by representatives of the GLOBALEX Preparatory Board, Global WordNet Association (GWA), and ELEXIS (H2020 project on European Lexicographic Infrastructure). The workshop begins with a short introduction and an overview of its main theme of Lexicography and WordNets, followed by 11 oral presentations and 6 posters, and concludes with an open discussion including presentations of GLOBALEX and ELEXIS.

The field of lexicography is continuously shifting to digital media – with effects on all stages of research, development, design, evaluation, publication, marketing and usage – while modern lexicographic content is becoming increasingly interoperable with numerous computational domains and solutions as part of large-scale knowledge systems and collaborative intelligence.

At the same time, new interlinked linguistic resources are being created to meet requirements for language technology (LT), leading to better federation, interoperability and flexible representation. In this context, lexicography constitutes a natural part of the Linguistic Linked (Open) Data (LLOD) scheme, currently represented by WordNets, FrameNets, and LT-oriented lexicons, ontologies and lexical databases. The various attempts that have been taken in the last decades to embed lexicography in a theoretical framework are leading the current search for a new research paradigm and common standards, including also the interoperability with LT systems and applications.

This second iteration of GLOBALEX Workshop continues to explore the development of global standards for the evaluation of lexicographic resources and their incorporation into new LT services and other devices. It seeks to promote cooperation with related fields of LT for all languages worldwide, and is intended to bridge existing gaps within and among such different research fields and interest groups.

The full papers and abstracts included in this volume cover nearly all the workshop presentations and discuss its main topic of Lexicography and WordNets as well as issues regarding the globalization and digitization of lexicography, etymology and historical lexicography, and the interoperability of lexicography with other disciplines and external resources and domains, mainly linguistic linked data and terminology.

*Ilan Kernerman and Simon Krek*
Co-chairs and editors



*Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets", Ilan Kernerman, Simon Krek (eds.)*

# Programme

09:00 – 09:30 Opening

Ilan Kernerman and Simon Krek, *Introduction*
Francis Bond, *WordNets and Lexicography*

09:30 – 10:30 Session 1

Thierry Declerck, John McCrae, Roberto Navigli, Ksenia Zaytseva and Tanja Wissik, *ELEXIS – European Lexicographic Infrastructure: Contributions to and from the Linguistic Linked Open Data*

Katrien Depuydt and Jesse de Does, *The Diachronic Semantic Lexicon of Dutch as Linked Open Data*

Sabine Tittel and Christian Chiarcos, *Linked Open Data for the Historical Lexicography of Old French*

10:30 – 11:00 Coffee break

11:00 – 12:40 Session 2

Aikaterini-Lida Kalouli, Livy Real and Valeria dePaiva, *WordNet for "Easy" Textual Inferences*

Armin Hoenen, *Attempts at Visualization of Etymological Information*

Sonja Bosch and Gertrud Faaβ, *Options for a Lexicographic Treatment of Negation in Zulu*

Elsabe Taljard and Daniel Prinsloo, *Developing a Multi-Functional e-Spelling Dictionary – a South African Perspective*

Daniel McDonald and Eveline Wandl-Vogt, *Blockchain Lexicography: Prototyping the Collaborative, Participatory Post-Dictionary*

12:40 – 14:20 Lunch break

# Programme

14:20 – 15:20 Session 3

Marie-Claude L'Homme, Nathalie Prévil and Benoît Robichaud, *A Methodology for Locating Translations of Specialized Collocations*

Sara Carvalho, Rute Costa and Christophe Roche, *Natural Language Definitions: An Example from the Biomedical Domain*

Monica Monachini and Anas Fahad Khan, *Towards the Construction of a Lexical Data and Technology Ecosystem: The Experience of ILC-CNR*

15:20 – 16:00 Posters

Raquel Amaro, *Integrating Prepositions in WordNets: Relations, Glosses and Visual Description*

Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo, Mohamed Khemakhem and Laurent Romary, *Presenting the Nenufar Project: A Diachronic Digital Edition of the Petit Larousse Illustré*

Ariel Gutman, Alexandros Andre Chaaraoui and Pascal Fleury, *Crafting a Lexicon of Referential Expressions for NLG Applications*

Mohamed Khemakhem, Axel Herold and Laurent Romary, *Enhancing Usability for Automatically Structuring Digitised Dictionaries*

Pilar León-Araúz and Antonio San Martín, *The EcoLexicon Semantic Sketch Grammar: From Knowledge Patterns to Word Sketches*

Bolette Pedersen, Sanni Nimb, Sussi Olsen and Nicolai Sørensen, *Combining Dictionaries, WordNets and other Lexical Resources – Advantages and Challenges*

16:00 – 16:30 Coffee break

16:30 – 18:00 Discussion

Simon Krek, Ilan Kernerman and Francis Bond, *GLOBALEX – ELEXIS – Conclusions*

# Table of Contents





*Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets", Ilan Kernerman, Simon Krek (eds.)*

# Table of Contents

# Options for a Lexicographic Treatment of Negation in Zulu

**Sonja Bosch**[1], **Gertrud Faaβ**[1,2]

[1]University of South Africa /[2]University of Hildesheim
PO Box 392, Unisa 0003, Pretoria, South Africa/Universitätsplatz 1, 31141 Hildesheim, Germany
boschse@unisa.ac.za, gertrud.faass@uni-hildesheim.de

**Abstract**

Dictionaries of today should offer much more than just knowledge about single words, they should rather be regarded as language information tools. However, in most electronic dictionaries of today, complex morphological constructions are not considered, thus users of dictionaries are usually expected to analyse such complex words themselves and to query base forms. With such a task, language learners, especially beginners, are often out of their depth. The question arising now, in particular with regard to learners' dictionaries, is whether and how can we enable an electronic dictionary to analyse complex constructions providing information on their structure and on their meaning? Taking Zulu negation as an example of a complex morphological construction, we first examined the frequency of this phenomenon in the corpora available and found an impressive number of them. So in the latter part of this paper, we try to find options for a practical implementation in electronic dictionaries.

**Keywords:** negation, Zulu, corpus-based queries, lexicographic treatment, learners' dictionaries

## 1. Introduction

Negation is described by Crystal (1994:231) as "A process or construction in GRAMMATICAL and SEMANTIC analysis which typically expresses the contradiction of some or all of a sentence's meaning." As an important instrument of language use, one would therefore expect aspects of negation to be dealt with in dictionaries. However, as Dahl (1979) states, negation phenomena appear to be at the border of lexicon and grammar, thus, one could argue that grammatical issues are not a matter for lexicography. Electronic dictionaries, on the other hand, are nowadays seen rather as language information tools, that is, they are to contain and to present extra-lexicographic data about a language as well (cf. Prinsloo et al. 2012), so negation again comes into play. Kovarikova et al. (2012:827) point out that "The main advantages of such a dictionary - almost unlimited size, interconnectivity of entries, easy referencing both within the dictionary and to a corpus - can also be used to describe negation … with all its aspects." Although only very few works are available on the lexicographic treatment of negation, novel (paper) dictionary conventions for the handling of negative verbal morphemes in Northern Sotho are proposed by Prinsloo and Gouws (1996), while van Son et al. (2016) address the need for building a dictionary of affixal negations and regular antonyms.

In a language such as English, word formation rules are relatively straightforward in the sense that inflections and derivations are usually constructed by adding suffixes to a root. Therefore, written words are usually roots, or commence with stems, and these roots can be looked up with ease in an English dictionary. An agglutinating language such as Zulu, however, has a much richer morphological structure, comprising an extensive and productive system of affixation that "pushes" roots into the middle of a word. Just looking up a Zulu word therefore requires the ability of a language learner to take the word apart by stripping the prefixes and suffixes and identifying any morphophonological changes that took place, in order to extract a stem that can be looked up in the dictionary.

The aim of this paper is therefore to investigate methods of enabling an electronic Zulu (learners') dictionary (with inflected forms) to analyse complex constructions, in this case negation phenomena, providing information on their structure and on their meaning (thereby supporting both perception and production). We firstly provide some background on negation in Zulu, followed by an investigation into the frequency of this phenomenon of negation as reflected in Zulu corpora. This is followed in section 4 by a description of negation as treated in existing Zulu dictionaries. In section 5 we suggest requirements for improved (electronic) Zulu bilingual (learners') dictionaries, and present options for a practical implementation in electronic dictionaries with detailed exemplification. These options are based on our findings in the foregoing two sections, and also include existing data and software. Finally, a conclusion and notes on future work are presented.

## 2. Background on Negation in Zulu

### 2.1 Orthography

Zulu follows a conjunctive orthography, which means that bound morphemes are attached to the words (unlike other South African languages e.g. the Sotho languages), and thus cannot occur independently as separate words.

Furthermore, the order of occurrence of morphemes is fixed, as in other agglutinating languages such as Turkish.

Orthographic words are of a polymorphemic nature of affixes attached to the root or core of the word, while monomorphemic words are limited to the following parts of speech: ideophone, conjunction and interjection. It is also noteworthy that morphophonological changes may occur between lexical and surface levels.

Kosch (2006:42) emphasizes that mother-tongue speakers of a language are familiar with the structure and sound patterns of their language, and therefore intuitively select allomorphs that are conditioned by the relevant phonological rules. For language learners however, this

selection may seem unnatural and they need to learn consciously when certain sound changes need to be made. An example of a morphophonological change such as vowel elision in the formation of negatives is demonstrated in (1) where the vowel of the SC01_neg is deleted before the vowel initial VRoot *-akh-* in order to present as *akakhi* on the surface:

(1)

| *a* | *ka* | *akh* | *i* |
|---|---|---|---|
| neg | SC01_neg | VRoot | VEnd (neg) |
| not | 3rd person sg. | build | |

'he/she does not build / is not building'

## 2.2 Morphological Negation

Zulu is characterised, among others, by a rich morphological structure including a noun class system which classifies nouns into a number of noun classes, as indicated by noun prefix morphemes. Noun prefixes play a significant role in the morphological structure of the language in that they connect the noun to other parts of speech (e.g. verbs, adjectives, possessives and pronouns) in the sentence. This linking takes place by means of a system of so-called concordial agreement morphemes which are derived from the noun prefixes and usually bear a close resemblance to them.

The two main forms that negation takes in the Bantu languages of Guthrie's so-called zone S (i.e. those of Southern Africa) including Zulu, are described by Gowlett (2003:636) as (i) the use of a preconcordial negative marker, with or without a concurrent suffixal marker; (ii) the use of a post-concordial negative marker, with or without a concurrent suffixal marker. This applies not only to verb constructions, but also to so-called copulative constructions that include adjectives and relatives[1].

According to Kosch (2006:106) the positive form of the verb is not clearly identifiably marked by affixes, while overt marking does occur in the negative form. Negativising strategies may vary in different moods such as the participial (sometimes referred to as the situative in grammatical descriptions) and the subjunctive mood; the imperative form of the verb, and tenses such as the past, perfect and future tenses, as illustrated in the following examples:

(2a) Participial -
*(uma) ehamba* > *e**nga**hambi* 'if he goes/does not go'

(2b) Subjunctive mood -
*(ukuze) ahambe* > *a**nga**hambi* 'so that he goes/does not go'

(2c) Imperative form -
*vala* > ***unga**vali* 'close/do not close) (singular)
*valani* > ***ninga**vali* 'close/do not close) (plural)

(2d) Past tense -
*bahamba* > ***a**bahamb**anga*** 'they went/did no go'

(2e) Perfect tense -
*bahambe/bahambile* > ***a**bahamb**anga*** 'they went/did no go'

(2f) Future tense -
*bazohamba* > ***a**ba**zu**hamba/aba**zuku**hamba* 'they will go/ will not go'
*bayohamba* > ***a**ba**yu**hamba/aba**yuku**hamba* 'they will go/ will not go'

(2g) Stative -
*silele* >***a**silele* 'he/she/it is asleep/ is not asleep'

As evident in the examples above, with the exception of the stative form in (2g), one or more prefixes take a negative form in conjunction with a change in final suffix of verb, therefore Zulu verbal negation strategies can be summarised as showing either dyadic negation (3) or polyadic negation (4):

(3a)

| | *ngi-* | *-hamb-* | *-a* |
|---|---|---|---|
| | SC1p | VRoot | VEnd |
| | person Sg. | go | |

'I go'

(3b)

| *a-* | *ngi-* | *-hamb-* | *-i* |
|---|---|---|---|
| neg | SC1p | VRoot | VEnd (neg) |
| not | 1st person Sg. | go | |

'I do not go'

(4a)

| *u-* | *-ya-* | *-hamb-* | *-a* |
|---|---|---|---|
| SC01 | long pres tense | VRoot | VEnd |
| 3rd person sg | | go | |
| he/she goes | | | |

'he/she goes/is going'

(4b)

| *a-* | *ka-* | *-hamb-* | *-i* |
|---|---|---|---|
| neg | SC01_neg | VRoot | VEnd (neg) |
| not | 3rd person sg. | go | |

'he/she does not go / is not going'

Over and above the regular negated constructions as shown in (3) to (4), we also find a number of additional rules for specific verbs or verb forms. Whereas passive verbs in the perfect and past tense suffix the negative suffix *-anga* (as in 5a), passive verbs in the present tense, for example, may not use the negative verbal ending *-i* when being negated, but retain the positive *-a*, as in (5b):

(5a)

| *a-* | *yi-* | *-shay-* | *w* | *-anga* |
|---|---|---|---|---|
| neg | SC09 | VRoot | Pass | VEnd (neg) |
| not | 3rd person sg. | beat | | |

'It was not beaten'

(5b)

| *a-* | *ba-* | *-thand-* | *w* | *-a* |
|---|---|---|---|---|
| neg | SC02 | VRoot | Pass | VEnd |
| not | 3rd person pl. | like | Pass | |

'They are not liked'

In Figure 1 we summarise verbal negativising strategies used in Zulu.

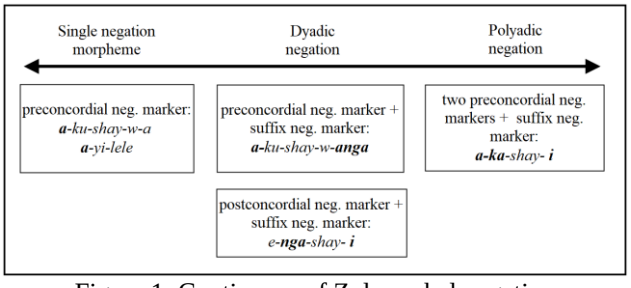



Figure 1: Continuum of Zulu verbal negation

[1] We will only be dealing with verbal negation in this paper.

Moreover, there are also so-called defective verb forms such as *-sho* 'say' which take irregular negative suffixes, for example *-ongo* instead of the regular *-anga* in the past tense. Further defective verb forms are *-thi* 'say; think' and *-azi 'know'*. They have an irregular verb ending *-i* which does not change when the verb is negated in the present or future tense, but is replaced by the negative suffix *-anga* in the past tense. See the following examples:

(6a) *ba- -sh- -o*
SC02 VRoot VEnd
3rd person pl. go
'They said'

(6b) *a- ba- -sh- -ongo*
neg SC02 VRoot VEnd (neg)
not 3rd person pl. go
'They did not say'

(6c) *a- ba- -az- -i*
neg SC02 VRoot VEnd
not 3rd person pl. know
'They do not know'

(6b) *a- ba- -az- -anga*
neg SC02 VRoot VEnd (neg)
not 3rd person pl. know
'They did not know'

### 2.3 Syntactic Negation

In English, verbs take different auxiliaries when forming the negative, there are so-called 'is' and 'have' forms. Thus, to correctly translate an identified negated verb form, we need to store the respective category of the (English) translation of each verb stem in the dictionary.
In Zulu, such categories do not exist, we however find several lexicalized negation word forms used in the imperative as shown in (7), similar to negating strategies of e.g. English. We categorize these as "syntactic negation".

(7a) *mus- a- uku- -hamb- -a*
VRoot VEnd SC15 VRoot VEnd
do not (imp) cl15(inf) go
'Do not go!' (semantically stronger than simple negation)

(7b) *yek- -a uku- -hamb- -a*
VRoot VEnd SC15 VRoot VEnd
stop (imp) cl15(inf) go
'Do not go!!' (semantically stronger than (7a))

## 3. Negation as Reflected in Zulu Corpora

It is well-known that in comparison to a language such as English for which corpora with billions of tokens are available, Zulu can be regarded as an under-resourced language (cf. Prinsloo, 2012:121, Quasthoff et al., 2016:89). To the best of our knowledge, there are only four Zulu corpora that are freely available:

(a) The raw University of KwaZulu-Natal (*UKZN)* isiZulu National Corpus[2], containing about 19.5 million tokens (no publication found). Of this corpus, no sentences but a word frequency list is downloadable;
(b) the raw Wortschatz Universität Leipzig Internet Corpus of Zulu (*LC*, Quasthoff et al. 2014) contains about 3.2 million tokens (2.77 million words);
(c) the *NCHLT* isiZulu Annotated Text Corpus (2014), which is based on government web pages and contains about 46,000 tokens (39,869 words). This corpus is available in different formats, we chose the version annotated with parts of speech;
(d) the UKWABELANA corpus (*UK*, Spiegler et al. 2010) containing about 21,400 words (no punctuation) which is very small by world standards, but is nevertheless also available in different formats. Again, we chose the version annotated with parts of speech.

For a better comparability and to simplify searches, we downloaded corpora (b), (c) and (d) and encoded them with the Corpus WorkBench (Evert and Hardie, 2011). In the case of the UKZN corpus, we made use of the word frequency list.
Table 1 shows the number of occurrences of the syntactic verb negation described above (*musa/musani/yeka/yekani* followed by a verb in the imperative). As in most corpora, we do not find many texts of the type "conversation" in which imperatives occur, thus these phenomena are not very frequent.

| ***Type of negation*** | ***UKZN*** | ***LC*** | ***NCHLT*** | ***UK*** |
|---|---|---|---|---|
| *musa uk…a* | n.a.[3] | 69 | 3 | 0 |
| *musani uk…a* | n.a. | 7 | 0 | 0 |
| *yeka uk…a* | n.a. | 17 | 0 | 0 |
| *yekani uk…a* | n.a. | 2 | 0 | 0 |

Table 1: Frequency of occurrence of syntactic negation

A more frequent way of negating the imperative is the (semantically) weaker morphological negation form *unga…i* (singular) or *ninga…i* (plural) as described in example (2c). The frequency of occurrence of this strategy in the Wortschatz Universität Leipzig Internet Corpus of Zulu (~ 3,2 mio tokens) is 1,264 which is fairly high in comparison to Table 1.

| ***Type of negation*** | ***UKZN*** | ***LC*** | ***NCHLT*** | ***UK*** |
|---|---|---|---|---|
| *unga…i* | 13,824 | 1,140 | 11 | 28 |
| *ninga…i* | 1,955 | 124 | 0 | 1 |
| Total | 15,779 | 1,264 | 11 | 29 |

Table 2: Frequency of occurrence of *unga…i* and *ninga…i*

[2] https://iznc.ukzn.ac.za/ [2017-12-25]
[3] Not applicable because the available data of UKZN consists of wordlists and not sentences.

To find the cases of morphological negation, a number of scripts were developed which make use of regular expressions. These describe the different verb forms in their full paradigm of inflection. Taking the verb forms of *-thanda* ([to] like) as an example, we find the list of present tense indicative conjugation forms shown in (8a). The appropriate regular expression in (8b) encodes these forms, but does not include the root *-thand-*.

(8a) *angithandi, awuthandi, asithandi, anithandi, akathandi, abathandi, awuthandi, ayithandi, alithandi, wathandi, asithandi, azithandi, ayithandi, azithandi, aluthandi, abuthandi, akuthandi*

(8b) (a[bkw]a|angi|a[lnsyz]i|a[blkw]u).+i

Thus we plan to find and count all negated verb forms following the regular conjugation pattern in the corpora. We do not differentiate between upper and lower case letters, but we exclude forms matched by the regular expression of which we know that they are not negated verbs. There are, for example, also deverbative nouns beginning with *aba-* and ending in *-i* (e.g. *abafazi*), of which we generated a stop list. We also exclude relative and adjective constructions like *ababanzi* or *abaningi*. However, ambiguous forms like *ababhali* (verb as well as deverbative noun) remain in the query as noise (see Annexure A for a list). Lastly, we use pos="v" as a selection condition (only for the corpora where parts of speech are annotated). The results are found in Table 2 and they show that negated verb forms are a frequent matter (at least in written text) worth describing in more detail in dictionaries.
Table 3 also shows that there would be sufficient data in the corpora for finding examples to be linked to the entries of dictionaries.

| ***Type of negation*** | ***UKZN*** | ***ZULU*** | ***NCHLT*** | ***UK*** |
|---|---|---|---|---|
| imperative | 9,534 | 1,545 | 11 | 28 |
| present tense | 121,554 | 20,967 | 105 | 123 |
| participial/subjunctive | 86,471 | 11,504 | 83 | 129 |
| recent past | 15,483 | 1,869 | 10 | 20 |
| recent past continuous | 3,939 | 554 | 1 | 7 |
| remote past | 13,758 | 1,711 | 10 | 15 |
| recent past remote cont. | 7,832 | 486 | 0 | 26 |
| recent past perfect | 163 | 25 | 0 | 0 |
| remote past perfect | 506 | 62 | 0 | 2 |
| future tense | 13 | 0 | 0 | 0 |
| future tense continuous | (no.of aux verbs) 5,022[4] | 24 | 0 | 0 |
| future tense perfect | | 0 | 0 | 0 |

[4] These forms are generated with a preceding auxiliary word (*asobe, basobe* etc.). As there's only a word list in UKZN available, we search for those.

| | | | | |
|---|---|---|---|---|
| Total | 264,275 | 38,747 | 220 | 350 |
| no. of words | 19,553,511 | 2,771,207 | 39,867 | 21,416 |
| % verb negations | 1.35 | 1.40 | 0,55 | 1.63 |

Table 3: Frequencies of occurrences of morphological verb negation

## 4. Negation as Reflected in Zulu Dictionaries

In this section we address the treatment of negation in a variety of Zulu dictionaries ranging from paper to online dictionaries and compare it to the findings of negation as reflected in the available corpora discussed in the foregoing section. Dictionaries are fundamental resources for language learning, however, lexical resources for Zulu are still very limited, and machine-readable lexicons are not freely available.
In Table 4, we show how some well-known Zulu paper dictionaries, namely the bilingual general dictionary of Doke et al. (2005), the bilingual learners' dictionaries of Dent and Nyembezi (1969) and of De Schryver (2010), and the monolingual general dictionaries of Nyembezi (1992) and Mbatha (2006) deal with the negation phenomena of Zulu verbs. It is conspicuous that negation is treated inconsistently in the various dictionaries.

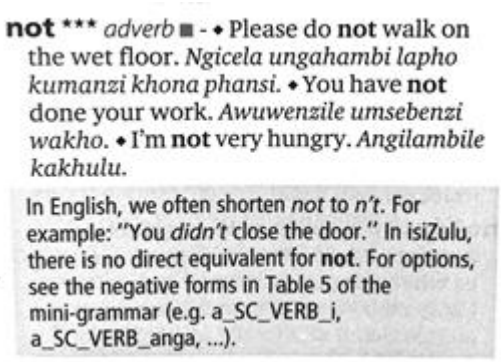
| ***Dictionary*** | ***morph. negation*** | ***syntactic negation*** | ***outer matter*** |
|---|---|---|---|
| Doke et al. | Yes | Yes | Notes / Tables |
| Dent & Nyembezi | Yes, two examples provided with 'not' | Yes | No info on negation |
| De Schryver (ed) | Occasional examples, Textboxes | Yes, Textbox | Mini-Grammar |
| Nyembezi | No | Yes | No info on negation |
| Mbatha | No | only *phinde* | No info on negation |

Table 4: Negation in printed dictionaries

Doke et al. (2005) list syntactic negation by means of the two (auxiliary) verb stems *musa* 'don't' and *yeka* 'leave off; stop; let go'. In the case of *musa*, the plural *musani* is also listed, as well as the information for the user that this verb is used to form negative imperatives 'don't; you mustn't'. The outer matter also contains notes and tables dealing with negation.
In a scholar's dictionary such as that of Dent and Nyembezi (1969) one would expect some outer matter information on negative constructions to guide scholars. The following is the only information available: the two (auxiliary) verb stems *musa* 'don't' and *yeka* 'leave off;

stop; let go' as well as the conjunctive *phinde* 'never' are included in the Zulu-English side of the dictionary, while a lookup under 'not' on the English-Zulu side, actually provides two negated verb constructions *angiboni* 'I do not see' and *asibonanga* 'we did not see'.

In De Schryver's (2010) bilingual school dictionary, morphological as well as syntactic negation are included in the dictionary with occasional examples and textboxes referring the user to the mini-grammar in the outer matter that contains tables of negative forms. This is illustrated in Figure 2 below.


**not** *** *adverb* ■ - ♦ Please do **not** walk on the wet floor. *Ngicela ungahambi lapho kumanzi khona phansi.* ♦ You have **not** done your work. *Awuwenzile umsebenzi wakho.* ♦ I'm **not** very hungry. *Angilambile kakhulu.*

In English, we often shorten *not* to *n't*. For example: "You *didn't* close the door." In isiZulu, there is no direct equivalent for **not**. For options, see the negative forms in Table 5 of the mini-grammar (e.g. a_SC_VERB_i, a_SC_VERB_anga, ...).

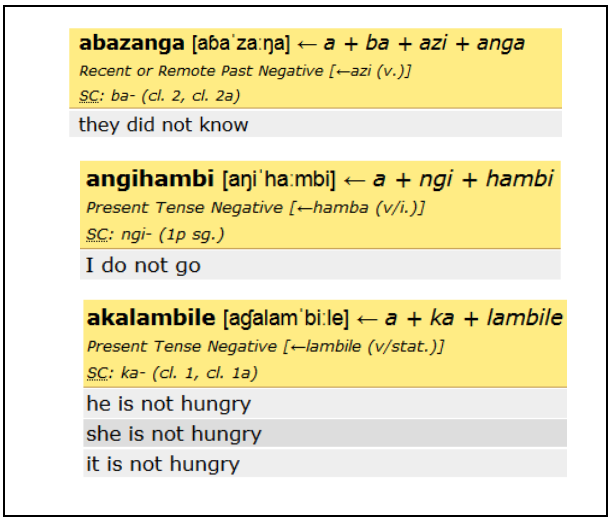
Figure 2: Example of textbox (De Schryver, 2010:431)

Although Nyembezi (1992) lists *musa* and *yeka* as (auxiliary) verb stems with the meanings 'do not; stop doing' there are no examples provided, and no description of any negation in the outer matter. The same applies to Mbatha's (2006) monolingual dictionary. In fact, syntactic negation in this dictionary is limited to the auxiliary conjunctive *phinde* 'never'.

isiZulu.net (2018) functions as a Zulu-English online dictionary that also offers morphological decomposition without the need of stem identification before a word is looked up. Prinsloo (2012:135) describes isiZulu.net as "probably the most sophisticated online dictionary for the Bantu languages." A fairly high amount of back matter is offered in the form of grammar and verb conjugation tables (which we used for developing the regular expressions in section 3). However, the only negative morphemes that occur in the tables are the first person singular subject concord *-ka-, and -zu-* the negative form of the future tense morpheme. isiZulu.net (2018) already offers a translation for negated verbs, e.g. *angihambi* is translated as 'I do not go'.

The individual analyses of lookups present automatic morphological decomposition, which in the case of negative verb forms decomposes the prefixes, i.e. the negative morpheme and subject concord, but the negative suffixes are only decomposed selectively, e.g. those of the past tenses. Nevertheless, learners of the language can use this information as a pattern for producing other negated verbs. Figure 3 shows three respective analyses by isiZulu.net (2018).

So far, we do not see a sufficient treatment of negation in Zulu in the major (paper) dictionaries, except maybe the Oxford learners' dictionary of De Schryver (2010). This dictionary, however, is rather small and addresses mainly learners. We are also not informed whether there are still newer editions of the existing printed dictionaries of Zulu planned. However, for such, we would suggest adding a number of textboxes which describe at least the negation forms of highly frequent verbs and rules for forming irregular (defective) forms. Syntactic negation should at least be mentioned with the respective auxiliaries adding examples of their use. Respective back matter information in the form of conjugation tables and/or mini-grammars should be added to all bilingual dictionaries.


**abazanga** [aɓaˈzaːŋa] ← *a + ba + azi + anga*
*Recent or Remote Past Negative [←azi (v.)]*
*SC: ba- (cl. 2, cl. 2a)*
they did not know

**angihambi** [aŋiˈhaːmbi] ← *a + ngi + hambi*
*Present Tense Negative [←hamba (v/i.)]*
*SC: ngi- (1p sg.)*
I do not go

**akalambile** [aɠalamˈbiːle] ← *a + ka + lambile*
*Present Tense Negative [←lambile (v/stat.)]*
*SC: ka- (cl. 1, cl. 1a)*
he is not hungry
she is not hungry
it is not hungry

Figure 3: isiZulu.net (2018) analyses of *abazanga, angihambi* and *akalambile*

## 5. Requirements for Improved (Electronic) Zulu Bilingual (Learners') Dictionaries

It is not known how negated verbs in isiZulu.net (2018) are analysed and we do not wish to speculate. In general, however, we do not think that changing the data model of the dictionary's database is a solution because morphological negation is a dynamic process of word formation. We rather see a query processor first checking whether the word queried by the user is contained in the database. If that is not the case, an analysis of the word must take place. In our view, there are two possible options for such an analysis tool when extending electronic dictionaries so that negated verb forms can be queried:

(a) Implementing a rule-based component on the basis of regular expressions as it was done for a few examples in the Zulu Learners' dictionary (Faaß and Bosch 2016); this method could be enhanced by utilizing a dictionary of affixal negations as suggested by van Son et al. (2016).

(b) Adding ZulMorph[5], the Finite State Morphological Analyser for Zulu as described in Bosch and Pretorius (2016:11) as a component of the dictionary.

The implementation of ad-hoc rules as described for option (a) would offer an opportunity to select and show the most probable analysis of a word form and to add

[5]A demo version of ZulMorph is accessible at http://gama.unisa.ac.za/demo/demo/zulmorph

didactic information for language learners, i.e. the users of the dictionary for instance by adding a link to adequate online lectures concerned with negation or by giving additional explanations on special cases. Such a component could be limited to the vocabulary and morphology addressed in the teaching materials as suggested by Antonsen (2013) for cases of morphologically complex indigenous languages that do not have morphological analysers. Instead of putting all the knowledge and the processing in one component, one could alternatively use the dictionary of affixal negation as proposed by van Son et al. (2016) as a model and compile a new dictionary of isiZulu affixal negations with data of the Zulu wordnet which is based on the English Princeton WordNet (cf. Bosch and Griesel, 2017). Another option would be to feed such a dictionary with data from the part-of-speech ontology implemented by Taljard et al. (2015). The result would become a knowledge base of which a processing component could make use of. Such an additional dictionary could also contain additional information on regular antonyms, again taken from wordnet data, e.g. *bonakala* 'appear' vs. *nyamalala* 'disappear'. However, implementing morphological rules to reproduce the natural processes of negation is an effort already performed with the existing finite state transducer (FST) machine and by adding such rules and extra data to a dictionary we would in a way re-invent the wheel. We hence rather look at ways and means to add the FST machine as a module to the dictionary. Here we are however facing the first challenge, namely that in the case of ambiguous words, the analyser returns multiple analyses: just for a rather simple verb like *abahambi* 'they do not walk', the FST offers five different analyses, *ungathi* 'you/it do(es) not say' even results in as many as 24 analyses. A solution for this problem could be an often-used and reliable method to reduce the number of analyses: the application of Optimality Theory (OT) (Archangeli and Langendoen, 1997) on this FST, i.e. by ranking its paths in order to find the most probable one. Such task would also be useful for instance for developing a parser or when making use of the FST for tagging, etc.

Another challenge is that of underspecification: When querying the verb form *ayibaleki* in ZulMorph, there are 12 analyses delivered[6]. For the verb root *-bal-* identified in (9) and (10), ZulMorph finds two valid analyses: The verb root *-bal-* means "count"; here it is extended with the neuter extension *-ek-* changing its meaning to the intransitive "be countable". No object concords occur in these analyses.

(9) a[NegPre]
i[SC][4]
bal[VRoot]ek[NeutExt]
i[VTNeg]
''they are not countable"

(10) a[NegPre]
i[SC][9]
bal[VRoot]ek[NeutExt]
i[VTNeg]
"he/she/it is not countable"

In (11) and (12), the intransitive root *-balek-* 'run away' is identified. Again, no object concord is identified; the analyses are therefore both valid.

(11) a[NegPre]
i[SC][4]
balek[VRoot]
i[VTNeg]
''they do not run away'

(12) a[NegPre]
i[SC][9]
balek[VRoot]
i[VTNeg]
'he/she/it does not run away'

Analyses (13) to (20) can be ignored because the identified base verb root *-al-* 'deny; refuse; reject' contains an object concord together with the neuter extension *-ek-* which in each case, changes the verb's valency[7].

(13) a[NegPre]
i[SC][4]ba[OC][2]
al[VRoot]ek[NeutExt]
i[VTNeg]*

(14) a[NegPre]
i[SC][9]ba[OC][2]
al[VRoot]ek[NeutExt]
i[VTNeg]*

(15) a[NegPre]
i[SC][4]bu[OC][14]
al[VRoot]ek[NeutExt]
i[VTNeg]*

(16) a[NegPre]
i[SC][9]bu[OC][14]
al[VRoot]ek[NeutExt]
i[VTNeg]*

(17) a[NegPre]
i[SC][9]bu[OC][14]
alek[VRoot]
i[VTNeg]*

(18) a[NegPre]
i[SC][4]ba[OC][2]
alek[VRoot]
i[VTNeg]*

(19) a[NegPre]
i[SC][4]bu[OC][14]
alek[VRoot]
i[VTNeg]*

[6] The results of ZulMorph were sorted here by the verbal roots identified; numbers and carriage returns were inserted for a better overview.

[7] So far, ZulMorph is not informed about the valencies of verbs.

(20) a[NegPre]
i[SC][9]ba[OC][2]
alek[VRoot]
i[VTNeg]*

When examining the valid analyses for *ayibaleki*, we find that ZulMorph identifies the following two verb roots as shown in (21) and (22):

(21) *-bal-* 'count; calculate'
(22) *-balek-* 'run away; escape; flee'

Using the Oxford Bilingual School Dictionary (De Schryver, 2010) as guideline with regard to corpus frequencies of verb stems, the most likely verb root in the above list is *-balek-* (two stars - the second group of most frequently used headwords) followed by *-bal-* (one star - the third group of most frequently used headwords). In the corpora consulted, as described in section 3, we investigated the present tense forms (short, long and negative form plus the forms of participial, and subjunctive mood) of *-balek-* and found the occurrences shown in Table 5. We do not know which corpus was used to generate the frequency lists for the Oxford School Dictionary, however our data differs slightly from that of the Oxford School Dictionary.

| ***Verb root*** | ***UKZN*** | ***LC*** | ***NCHLT*** | ***UK*** |
|---|---|---|---|---|
| *-balek-* | 1,919 | 166 | 0 | 8 |
| *-bal-* | 2,500 | 233 | 1 | 0 |

Table 5: Frequencies of occurrences of the present tense forms of *-balek-* and *-bal-*

Methodologically, a script working with respective regular expressions (described in the constraints above) which are informed about verb frequencies could determine that *ayibaleki* is a negated verb form with the roots *-balek-* or *-bal-*, of which the more frequent one is the preferred one and should be shown first. We are fully aware of the fact that *ayibaleki* might be a straightforward case, however, we see such a "picking" of the relevant repetitive parts of the analyses as a feasible option when connecting the finite state transducer to an electronic dictionary.

For syntactic negation, that is - from a technical perspective - for analysing and translating word sequences showing negation elements like *musa* or *yeka*, a word-based dictionary will most probably not be capable of offering the correct translation. For translating sequences, we are in need of a parser and/or a machine translation tool.

## 6. Conclusion and Future Work

In this paper, we examined the linguistic phenomena of morphological and syntactic verbal negation in Zulu. These are not very prominently discussed in printed dictionaries though they are difficult to (de-)construct for learners. In the only existing electronic dictionary providing a good coverage, isiZulu.net (2018), such negation is handled appropriately, however, as no publications exist, we can only speculate on how this implementation was done.

Verbal negation occurs frequently in the existing corpora of the language, we may thus assume that learners are confronted with verbal negatives quite frequently, especially in reception (for example in newspaper texts that were collected in the UKZN corpus). We hence provide suggestions on enhancing presentations in printed dictionaries, for example by making more extensive use of textboxes illustrating the linguistic phenomena in question (cf. Gouws and Prinsloo, 2014).

As Prinsloo et al. (2012) rightly state: "there are numerous complex situations where users need more detailed support than currently available in e-dictionaries, to make valid and correct choices". The proposal of Kovarikova et al. (2012) to interconnect affirmative and negative forms individually via referencing tools in e-dictionaries is a valid proposal too. We thus offer suggestions (including the incorporation of existing data and software) on how to enhance electronic dictionaries as language information tools so that they can handle at least the morphological negation phenomena appearing in Zulu and its related languages.

Although we only pay attention to negation in Zulu in this paper, this approach may lay the foundation for the lexicographic treatment of further complex constructions in Zulu, as well as negation in electronic dictionaries for the other four Nguni languages that are closely related to Zulu.

## 7. Acknowledgements

The various phases of research activities related to this paper were supported by the Scientific eLexicography for Africa project; the South African Centre for Digital Language Resources (SADiLaR); and the South African National Research Foundation. We also acknowledge the useful feedback from the anonymous reviewers.

## 8. Bibliographical References

Antonsen, L. 2013. Why ICALL for indigenous languages? https://en.uit.no/tavla/artikkel/sub?sub_id=354114&p_document_id=320146#antonsen (accessed on 15/01/2018)

Archangeli, D. and Langendoen, D.T. (eds.). 1997. *Optimality Theory: An Overview*. Oxford: Blackwell Publishing.

Bosch, S. and Griesel, M. 2017. Strategies for building wordnets for under-resourced languages: the case of African languages. *Literator* 38(1), a1351. https://doi.org/10.4102/lit.v38i1.1351

Bosch, S.E. and Pretorius, L. 2017. A computational approach to Zulu verb morphology within the context of lexical semantics. *Lexikos* 27:152-182.

Crystal, D. 1994. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell Publishers.

Dahl, Östen. 1979. Typology of sentence negation. *Linguistics* 17:79-106.

De Schryver, G.-M. (ed.). 2010. *Oxford Bilingual School Dictionary: Zulu and English.* Cape Town: Oxford University Press Southern Africa.

Dent, G.R. and Nyembezi, C.L.S. 1969. *Scholar's Zulu Dictionary.* Pietermaritzburg: Shuter and Shooter.

Doke, C.M., Malcolm, D.M., Sikakana, J.M.A. and Vilakazi, B.W. 2005. *English-Zulu, Zulu-English Dictionary*. Johannesburg: Witwatersrand University Press.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus WorkBench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 Conference*. University of Birmingham, UK.

Faaβ, G. and Bosch, S. 2016. An Integrated e-Dictionary Application - the case of an open educational trainer for Zulu. *International Journal of Lexicography* 29(3):296-310.

Gouws, R. and Prinsloo, D. J. 2014. Thinking out of the box. In: *Proceedings of the 16th EURALEX International Congress*, pages 501-511. Bolzano/Bozen, Italy

Gowlett, D. 2003. Zone S. In Derek Nurse and Gérard Philippson (eds) *The Bantu Languages.* Pp. 609-638. London and New York: Routledge.

isiZulu.net Zulu-English Dictionary. 2018. https://isizulu.net/ (accessed on 15/01/2018)

Kosch, I.M. 2006. *Topics in Morphology in the African Language Context*. Pretoria: Unisa Press.

Kovarikova, D., Chlumska, L. and Cvrcek, V. 2012. What belongs in a dictionary? The Example of Negation in Czech. In: Ruth Vatvedt Fjeld & Julie Matilde Torjusen (Eds.): *Proceedings of the 15th EURALEX International Congress*. 7-11 August 2012, Oslo, pp 822-827. Oslo: Reprosentralen, UiO. ISBN 978-82-303-2095-2. Accessed on 14 March 2016.

Mbatha, M.O. (ed.). 2006. *Isichazamazwi SesiZulu*. Pietermaritzburg: New Dawn Publishers.

Nyembezi, S. 1992. *Isichazimazwi sanamuhla nangomuso*. Pietermaritzburg: Reach Out Publishers.

Prinsloo, D. J. 2012. Electronic Lexicography for lesser-resourced languages. In Sylviane Granger & Magali Paquot (Eds): Electronic *Lexicography*. Pp. 119-144. Oxford: Oxford University Press.

Prinsloo, D.J. and Gouws, R.H. 1996. Formulating a new dictionary convention for the lemmatization of verbs in Northern Sotho, *South African Journal of African Languages* 16(3):100-107.

Prinsloo, D. J, Heid, U., Bothma, T. and Faaβ, G. 2012. Devices for information presentation in Electronic Dictionaries. *Lexikos* 22:290-320.

Quasthoff, U., Goldhahn, D. and Bosch, S. 2016. Morphology Learning for Zulu. Claudia Soria et al. (eds.) *Workshop CCURL 2016 - Collaboration and Computing for Under-Resourced Languages - 'Towards an Alliance for Digital Language Diversity'. 23 May 2016. 10th International Conference on Language Resources and Evaluation*, Portoroz, Slovenia. Pp.89-95.

Spiegler, S., van der Spuy, A. and Flach, P.A. 2010. Ukwabelana - an open-source morphological Zulu corpus. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, August 2010.

Taljard, E., Faaß, G., and Bosch, S.E. 2015. Implementation of a Part-of-Speech Ontology: Morphemic Units of Bantu languages. *Nordic Journal of African Studies* 24(2):146-168.

Van Son, C., van Miltenburg, E. and Morante, R. 2016. Building a Dictionary of Affixal Negations. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 49-56 (ExProM 2016). Osaka, Japan.

## 9. Language Resource References

NCHLT isiZulu Annotated Text Corpora. 2014. https://rma.nwu.ac.za/index.php/resource-catalogue/isizulu-nchlt-annotated-text-corpora.html. ISLRN 729-409-508-086-4.

UKWABELANA http://www.cs.bris.ac.uk/Publications/pub_master.jsp?id=2001224

University of KwaZulu-Natal. 2018. IsiZulu National Corpus. https://iznc.ukzn.ac.za/

Wortschatz Universität Leipzig. 2018. Corpus: Zulu (zul_mixed_2016). http://wortschatz.uni-leipzig.de/de

ZulMorph. n.d. Finite state morphology demo. http://gama.unisa.ac.za/demo/demo/zulmorph

## Annexure A

Frequencies of occurrences of negated verbs which might also be deverbative nouns (included in the corpus query results displayed in Table 3).

| ***Ambig.Deverb.N.*** | ***UKZN*** | ***ZULU*** | ***NCHLT*** | ***UK*** |
|---|---|---|---|---|
| *ababhali* | 346 | 54 | 0 | |
| *ababukeli* | 219 | 62 | 0 | |
| *ababulali* | 256 | 19 | 0 | |
| *abaculi* | 2,018 | 264 | 0 | |
| *abacwaningi* | 170 | 70 | 2 | |
| *abadayisi* | 136 | 27 | 0 | |
| *abadidiyeli* | 105 | 15 | 0 | |
| *abadlali* | 5,979 | 678 | 3 | 1 |
| *abafundi* | 7,143 | 1,743 | 59 | 1 |
| *abafundisi* | 189 | 55 | 3 | 1 |
| *abagadli* | 277 | 31 | 0 | |
| *abagibeli* | 499 | 197 | 0 | |
| *abagijimi* | 226 | 14 | 0 | |
| *abagqugquzeli* | 119 | 25 | 0 | |
| *abahlali* | 497 | 183 | 1 | |
| *abahlaseli* | 74 | 37 | 0 | |
| *abahleli* | 369 | 73 | 0 | |
| *abahloli* | 169 | 53 | 4 | |
| *abahluleli* | 40 | 9 | 0 | |
| *abahluzi* | 55 | 3 | 0 | |
| *abaholi* | 2,121 | 487 | 0 | |
| *abakaki* | 5 | 1 | 0 | |
| *abakhongi* | 296 | 10 | 0 | |
| *abalaleli* | 356 | 103 | 0 | |
| *abalandeli* | 2,692 | 443 | 0 | |
| *abalimi* | 491 | 583 | 3 | |

| | | | | |
|---|---|---|---|---|
| *abalingisi* | 370 | 57 | 0 | |
| *abalobi* | 62 | 13 | 0 | |
| *abalozi* | 44 | 4 | 0 | |
| *abameli* | 381 | 95 | 2 | |
| *abangani* | 1,125 | 159 | 0 | |
| *abanini* | 49 | 28 | 0 | |
| *abaphathi* | 1,194 | 313 | 2 | |
| *abaqashi* | 259 | 124 | 0 | |
| *abaqeqeshi* | 557 | 65 | 0 | |
| *abasakazi* | 525 | 157 | 0 | |
| *abasebenzi* | 2,791 | 1,546 | 25 | |
| *abaseshi* | 148 | 39 | 0 | |
| *abashayeli* | 657 | 235 | 0 | 1 |
| *abasiki* | 207 | 26 | 0 | |
| *abasizi* | 84 | 19 | 0 | |
| *abathakathi* | 237 | 9 | 0 | |
| *abathandi* | 627 | 92 | 0 | |
| *abathengi* | 412 | 159 | 0 | |
| *abaxhasi* | 287 | 67 | 0 | |
| *abazali* | 3,798 | 609 | 8 | 5 |
| *abefundisi* | 607 | 40 | 0 | 1 |
| *abelusi* | 95 | 6 | 0 | |
| Total | 39,363 | 9,101 | 112 | 10 |

# The Role of Conceptual Relations in the Drafting of Natural Language Definitions: an Example from the Biomedical Domain

**Sara Carvalho[123], Rute Costa[2], Christophe Roche[32]**

[1] ESTGA – School of Technology and Management – University of Aveiro
R. Comandante Pinho e Freitas, 28 3750-127 Águeda - Portugal
[2] NOVA CLUNL – Faculty of Social Sciences and Humanities – Universidade NOVA de Lisboa
Avenida de Berna, 26-C 1069-061 Lisboa – Portugal
[3] Condillac Research Group – LISTIC – Université de Savoie Mont Blanc
Campus Scientifique 73376 Le Bourget du Lac – France

E-mail: sara.carvalho@ua.pt, rute.costa@fcsh.unl.pt, christophe.roche@univ-savoie.fr

**Abstract**

Within the scope of the EndoTerm project, described in more detail in (Carvalho, Costa, & Roche, 2016; Carvalho, Roche, & Costa, 2015), this paper aims to explore Terminology's key role in supporting one of the fundamental forms of concept representation - the definition -, namely by assuming a double dimensional perspective in which the conceptual backbone supports the writing process. In particular, the article will focus on how conceptual information (i.e. the concept's position in the concept system, its characteristics, as well as the relationships linking it to other concepts) can be organised into a template-like format which would constitute the foundation of the natural language definition drafting process.

**Keywords:** conceptual relations, natural language definition, biomedicine

## 1. Introduction

In recent decades, the biomedical domain has undergone substantial changes: on the one hand, ageing population and the considerable decrease of the old-age support ratio have put more pressure on public health expenditure, raising concerns about the sustainability of social security systems and their role in health care; on the other hand, patients are playing an increasingly active and empowered role in their own healthcare; furthermore, technological innovation has been fostering an exponential growth in healthcare that is embodied not only in the widespread use of computerized examinations, procedures, prescriptions, and health records, but also in breakthroughs such as nanotechnology, 3D printing, robotic surgery, genomics, wearable technology, as well as the use of virtual, augmented and/or mixed reality.

At the core of this healthcare revolution are the current challenges regarding the creation, use, storage and dissemination of medical data, information, and knowledge. The ability to provide secure, reliable, efficient and cost-effective ways to process and exchange clinical information among the various stakeholders has become the foundation of eHealth action plans and programs worldwide, supported mainly by interoperability, i.e. "the ability of different information technology systems and software applications to communicate, exchange data, and use the information that has been exchanged" (HIMSS, 2013).

Therefore, this paper aims to explore Terminology's contribution to knowledge representation, knowledge organisation, and knowledge sharing in the biomedical domain. Anchored in a double dimensional approach to terminology work, the article will focus on how conceptual information can support the natural language definition drafting process. As regards its structure, the paper will be organised as follows: Section 2 presents an overview of the current biomedical terminological systems and their increasing need for natural language definitions, and Section 3 reviews the aforementioned double dimension perspective and its impact on the creation of natural language definitions. Section 4 is dedicated to the methodological approach underlying the EndoTerm resource, with a case study based around the concept of <Laparoendoscopic single-site surgery>[1] and encompassing both human- and machine-oriented formats, whereas section 5 provides examples of natural language definitions for <Laparoendoscopic single-site total hysterectomy[2]> and <Laparoendoscopic single-site ovarian cystectomy>. The final section summarises the main findings and outlines future lines of research.

## 2. Biomedical terminological resources and interoperability: is there still a place for natural language definitions?

As stated earlier, interoperability has become one of the 'hot topics' in healthcare, insofar as a successful implementation of interoperable solutions can contribute to enhance the quality and outcomes in the sector, while decreasing costs (Coiera, 2015). Yet, interoperability has also become one of the most challenging topics, due to the underlying complexity of delivering "the right information,

[1] A type of surgical procedure that is becoming increasingly prevalent in several medical specialties, including gynaecology. It is also known as LESS surgery.

[2] Throughout this paper, concepts will be capitalised and written between single chevrons, while terms will be presented in lower case and between double quotation marks (cf. (Roche, 2015)).

at the right time, to the right place" (Benson & Grieve, 2016). Thus, one of the key priorities in recent years has been to devise systems and applications that allow machines, rather than humans, to accurately communicate with each other (Sicilia & Balazote, 2013).

In this regard, the most recent versions of biomedical terminological systems (e.g. the Disease Ontology, the Unified Medical Language System (UMLS), or SNOMED CT) have been focusing predominantly on finding a solid conceptual foundation supported by formal (i.e. logic-based and computer-processable) concept definitions, as well as by Semantic Web standards, such as RDF and OWL, so as to enable inter-resource mapping. Within this framework, one might wonder whether there is still room in such resources for natural language definitions of concepts. It would appear so.

One of the short-term objectives of the Disease Ontology, for instance, is to expand the number of textual definitions until reaching full coverage (Kibbe et al., 2015). The 11th version of the International Classification of Diseases (ICD-11), to be released this year, will include "a short concise textual definition" for each entity, a feature that does not exist in the existing ICD-10 (WHO, 2011, p. 17). That will also be the case with the International Classification of Health Interventions (ICHI), currently awaiting its official release and where definitions will be used to "describe the intervention" and "assist the user in selecting the most appropriate [intervention] code" (ICHI, 2018). Moreover, and despite the fact that the current version of SNOMED CT lacks natural language definitions, it is also likely that this issue will be addressed soon. On the one hand, 63% of SNOMED CT users stated, in a 2010 survey, that textual definitions would be extremely relevant (Elhanan, Perl, & Geller, 2011). On the other hand, SNOMED CT's expected widespread use at an international level (e.g. it will fully replace the Read Codes in the UK National Health Service's Primary Care System by April 2018[3]) will presumably gather various stakeholders with different areas of expertise and subsequently raise particular needs, one of them probably being natural language definitions.

Notwithstanding this growing interest in textual definitions, no unequivocal guidance has been explicitly provided by the aforementioned biomedical terminological resources or their respective guidelines on how to draft such definitions. In ICD-11, for instance, contributors proposing a definition are advised to "describe the entity clearly and concisely" (WHO, 2011, p. 19), as well as to resort to existing definitions as much as possible. However, no further, more specific, drafting recommendations are outlined. The overall picture is not very different in the remaining biomedical terminological resources. In fact, one of the few - and pertinent - references to the governing principles of such definitions is to be found at the Draft ICHI Guidelines, which state that the definitions should "reflect the (...) axis categories from which the code is constructed[4]", thereby pointing towards the conceptual core structure of the classification as a useful starting point in the development of natural language definitions. Yet, once again, no additional information is given.

Bearing all of this in mind, it is believed that the current work can provide a contribution to systematising the natural language definition drafting process within this subject field, as will be further explored in the following sections.

## 3. Terminology: a matter of concepts and a matter of terms

At the heart of the work being carried out in this research project is the assumption that Terminology has a double dimension[5], linguistic and conceptual, in an approach that regards it as both a "science of objects and a science of terms" (Roche, 2015, p. 136). Therefore, terminology work needs to consider not only the analysis of discourses produced by experts but also the formal (or semi-formal) representations of the shared knowledge regarding their respective domains. For (Costa, 2013), the specificity of Terminology as an autonomous scientific subject lies precisely in these two dimensions and in studying the way they interrelate and become complementary. In short, the analysis of specialised texts, on the one hand, and the collaborative work with experts, on the other hand, play a key role in terminology work, supported by a theoretical and methodological framework that allows the terminologist to maximise the potential within each dimension and the synergies resulting from their interaction.

One of the areas of terminology work where the impact of this complementary approach can become more visible is precisely the definition, one of the core forms of concept representation and a topic that has been widely debated in Terminology for quite some time (de Bessé, 1997; Löckinger, Kockaert, & Budin, 2015; Rey, 1995; Sager, 1990, 2000; Sager & Ndi-Kimbi, 1995; Seppälä, 2007; Temmerman, 2000). According to the 1087-1 and 704 ISO standards (ISO, 2000, 2009), a terminological definition should allow a concept to be differentiated from other related concepts, either by stating its superordinate concept and the respective delimiting characteristics (intensional definition - regarded as preferential by ISO whenever possible) or by enumerating all its subordinate concepts under a given criterion of subdivision (extensional definition).

However, other approaches to Terminology (cf. (Meyer, Bowker, & Eck, 1992; Temmerman, 2000)) have highlighted the limitations and the lack of flexibility of such definitions, especially in more multi- or interdisciplinary subject fields, proposing, instead, a 'definitional template' that reflects the position that a given concept occupies in the conceptual system it belongs to. This has also been the case in Frame-based approaches to terminology work (Durán-Munoz, 2016; Faber, 2012, 2015) and to lexicography (Maks, 2006; Swanepoel, 2011), plus work by Fillmore (e.g. (Charles J. Fillmore, 2003; C. J. Fillmore & Atkins, 1994)).

Therefore, and within the scope of the EndoTerm project, examples will be provided in the following sections of how conceptual information (i.e. the concept's position in the

[3] https://digital.nhs.uk/SNOMED-CT-implementation-in-primary-care (20.12.2017)

[4] Cf. https://mitel.dimi.uniud.it/ichi/docs/#guidelines (15.01.2018).

[5] This approach has been described in more detail by (Costa, 2013; Roche et al., 2009; Roche, 2012, 2015; Santos & Costa, 2015).

concept system, its characteristics, as well as the relationships - both hierarchical and non-hierarchical - linking it to other concepts) can be organised into a template-like format which would constitute the foundation of the natural language definition drafting process.

## 4. EndoTerm: a double dimensional approach to terminology work within the biomedical field

The EndoTerm project[6] aims at the creation of a terminological resource focusing on medical terminology, namely on Endometriosis, a benign gynecologic condition affecting approximately 10% of women of reproductive age worldwide (Adamson, Kennedy, & Hummelshoj, 2010; Dunselman et al., 2014). Destined to future experts, experts of other, related domains, and also to expert patients, this research seeks to integrate both the linguistic and the conceptual dimensions in terminology work by relying on specialised corpus collection and analysis, as well as on a formal ontology, respectively. The latter constitutes the backbone of the aforementioned resource, combining hierarchical and non-hierarchical concept relations that allow a more accurate representation of the shared knowledge within this particular domain, as will be further explored in this section.

The development of EndoTerm led to the study of single-port surgery, a relatively recent type of surgical procedure that has been gaining significant ground regarding the treatment of gynecologic diseases, endometriosis being among them. A more detailed analysis of specialised resources from the subject field, including verbal, non-verbal, and multimedia content, pointed towards a lack of terminological consensus among the expert community, having identified more than 20 different terms in the literature (Carvalho, Costa, & Roche, 2016). In order to solve this terminological dispersion, the multidisciplinary Laparoendoscopic Single-Site Surgery Consortium for Assessment and Research (LESSCAR) issued a White Paper (Gill et al., 2010) that aimed to standardise the terminology in the field, proposing the term "laparoendoscopic single-site surgery" as the one that most accurately depicted this surgical procedure.

The analysis of the aforementioned sources, together with the feedback of senior expert gynaecologists who are also surgeons, helped ground the development of a micro-concept system concerning the main types of surgery performed in cases of endometriosis. As can be seen from Figure 1 below, this micro-concept system allows <Laparoendoscopic single-site surgery> to be positioned within the broader concept of <Surgical procedure> by resorting to a specific difference, Aristotelian-based approach. The figure depicts the initial stage of that conceptualisation process, i.e. a semi-formal concept representation developed with CMap Tools[7]. Moreover, three main axes of analysis were set up, thereby allowing the following specific differences to be outlined at each stage: i) degree of invasiveness[8]: /invasive[9]/ vs. /minimally invasive/; ii) existence of skin incision: /with skin incision/ vs. /without skin incision/; iii) number of skin incisions: /single skin incision/ vs. /multiple skin incisions/.

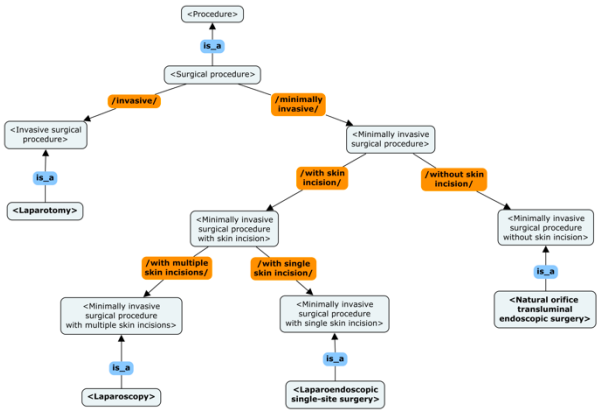



Figure 1: Types of endometriosis surgery.

Through this conceptual representation, it is possible to conclude that the existence of a single skin incision constitutes the essential characteristic (ISO, 2000) of <Laparoendoscopic single-site surgery>. Furthermore, it also allows a clearer distinction between different surgical approaches, i.e. the routes used to access the procedure site. In this case, <Laparotomy> is an example of an open or abdominal approach, <Laparoscopy> and <Laparoendoscopic single-site surgery> of a percutaneous endoscopic approach (either intraluminal or transluminal), whereas a procedure such as the <Natural orifice transluminal endoscopic surgery> (also known as NOTES) resorts to a per orifice transluminal approach[10].

As previously mentioned, this conceptual backbone can provide a valuable contribution to the development of natural language definitions, or to the enhancement of existing definitions. However, it is insufficient to distinguish between different surgical procedures that use the same surgical approach (e.g. **<Laparoendoscopic single-site hysterectomy>** is_a <Laparoendoscopic single-site surgery> is_a <Minimally invasive surgical procedure with single skin incision> *vs.* **<Laparoendoscopic single-site ovarian cystectomy>**

[6] Described in more detail in (Carvalho, Costa, & Roche, 2016; Carvalho, Roche, & Costa, 2015).

[7] A freely available software developed by the Florida Institute for Human and Machine Cognition (IHMC) and available at https://cmap.ihmc.us/cmaptools/.

[8] Following the existing lexicographic and terminological definitions, it has been assumed that all surgical procedures are, to some extent, invasive.

[9] As referred to earlier regarding the concepts and terms, the aforementioned differences also follow a typographical convention, being represented, in this case, between forward slashes.

[10] This results from a systematisation of the approaches listed on a set of current procedure classifications and other related biomedical terminological systems, such as SNOMED-CT, the IHCI, the ICD-10-PCS (Procedure Codes), used in the United States, the Canadian Classification of Health Interventions (CCI), and the French Classification Commune des Actes Médicaux (CCAM).

is_a <Laparoendoscopic single-site surgery> is_a <Minimally invasive surgical procedure with single skin incision>).

Therefore, and within the scope of the work that has been developed for EndoTerm, it is proposed that the preceding conceptualisation can be enhanced not only via hierarchical, but also non-hierarchical relationships[11], as well as a systematised categorial structure[12] for terminological systems of surgical procedures (ISO, 2012). The table below illustrates EndoTerm's conceptual framework regarding surgical procedures, in line with the ISO 1828: 2012, and includes the core top-level concepts, a set of is_a and non-hierarchical relationships and, lastly, the authorised Source Concept - Relationship - Target Concept combinations[13].

| SOURCE CONCEPT | RELATIONSHIP | TARGET CONCEPT |
|---|---|---|
| <Surgical procedure> | is_a | <Procedure> |
| <Surgical procedure> | has_method | <Surgical action> |
| <Surgical procedure> | has_procedure_site | <Human anatomy> |
| <Surgical procedure> | has_morphology | <Lesion> |
| <Surgical procedure> | has_surgical_approach | <Procedural approach> |
| <Surgical procedure> | uses_access_device | <Device> |
| <Lesion> | has_procedure_site | <Human anatomy> |
| <Device> | has_procedure_site | <Human anatomy> |
| <Device> | has_procedure_site | <Lesion> |

Table 1: EndoTerm's categorial structure.

The following micro-concept systems - built around the concepts of <Laparoendoscopic single-site total hysterectomy> (Figure 2) and <Laparoendoscopic single-site ovarian cystectomy> (Figure 3)[14], respectively - demonstrate how the template structure referred to above can help overcome the limitations of fully hierarchical concept representations, while providing a logical foundation that can prevent logical errors, especially at an initial, semi-formal stage where automatic reasoning may not be available.

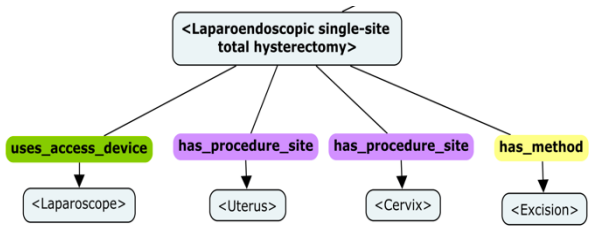



Figure 2: Micro-concept system for <Laparoendoscopic single-site total hysterectomy>.

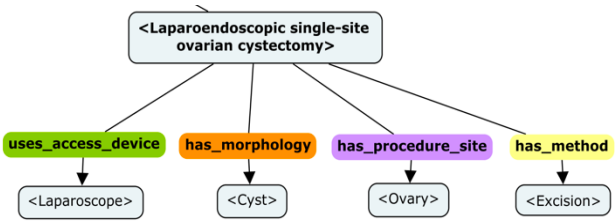



Figure 3: Micro-concept system for <Laparoendoscopic single-site ovarian cystectomy>.

To further substantiate the preceding approach, all of EndoTerm's micro-concept systems were then tested using TeDI (for OntoTerminology EDItor), a software environment created by C. Roche dedicated to the development of multilingual ontoterminologies[15]. In this case, and via TeDI, it was possible to validate EndoTerm's semi-formal concept systems and convert them into a formal ontology, also benefiting from the tool's built-in reasoner and from the subsequent logical verification that takes place during the ontology development process. The image below (Figure 4) shows a glimpse of TeDI's concept editor, namely from the concept <Laparoendoscopic single-site total hysterectomy>, its position in the hierarchy, the specific differences, as well as one of the non-hierarchical relationships (has_procedure_site).

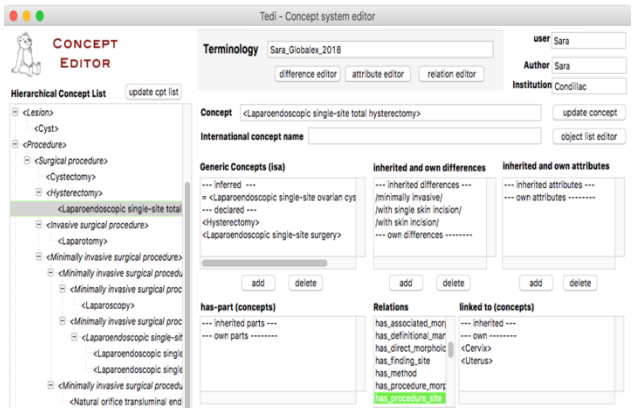



Figure 4: TeDI concept editor.

This formal concept definition can also be exported into W3C-compliant formats (RDF/XML), which can pave the way to a potential integration into existing biomedical, concept-oriented terminological resources.

[11] Despite their secondary role in the current ISO standards related to terminology and terminology work (ISO, 2000, 2009), non-hierarchical concept relationships are regarded as "equally important and more revealing about the nature of the concepts" (Sager, 1990, p. 34), as well as extremely relevant in the biomedical domain (cf. (McCray & Bodenreider, 2002; A. L. Rector et al., 1997; Smith et al., 2005)).

[12] i.e. a "minimal set of domain constraints for representing concept systems in a subject field" (ISO, 2007)).

[13] In Description Logic, the source and target concepts are also known as domain and range, respectively, and they are also subject to constraints (Baader, 2003; A. Rector & Rogers, 2006).

[14] Hysterectomy, often seen as a last resort in cases of severe endometriosis (Peter Rogers et al., 2016; Rogers et al., 2009), and ovarian cystectomy, i.e. the removal of ovarian endometriotic cysts or endometriomas (Working group of ESGE, ESHRE, and WES et al., 2017), are two common surgical procedures as regards the management and treatment of endometriosis.

[15] An ontoterminology is "a terminology whose conceptual system is a formal ontology" (C. Roche & Calberg-Challot, 2009). More information on the software can be found at http://christophe-roche.fr/tedi.

```
<owl:Class rdf:about="Laparoendoscopic_single-site_total_hysterectomy">
    <rdfs:label xml:lang="en">"laparoendoscopic_single-site_total_hysterectomy"</rdfs:label>
    <rdfs:label xml:lang="en">"LESS_total_hysterectomy"</rdfs:label>
    <rdfs:subClassOf rdf:resource="Laparoendoscopic_single-site_surgery"/>
    <rdfs:subClassOf rdf:resource="Hysterectomy"/>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="has_procedure_site"/>
            <owl:someValuesFrom rdf:resource="Cervix"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="has_procedure_site"/>
            <owl:someValuesFrom rdf:resource="Uterus"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="has_method"/>
            <owl:someValuesFrom rdf:resource="Excision"/>
        </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="uses_device"/>
            <owl:someValuesFrom rdf:resource="Laparoscope"/>
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

Figure 5: Formal definition of <Laparoendoscopic single-site total hysterectomy> in RDF/OWL.

## 5. Terminological definitions in EndoTerm: two examples

Based on the validated conceptualisation explored above, a template-based natural language definition can be put forward for each of the analysed LESS surgery concepts, with direct reference to both the specific difference approach and to the non-hierarchical relationships, supported by the categorial structure.

**Concept 1**: Type of <Surgical procedure> has_method <Surgical action> has_procedure_site <Human anatomy> uses_access_device <Device>

Hence, for the concept of <Laparoendoscopic single-site total hysterectomy>, the proposed definition is the following:

`<Minimally invasive surgical procedure> which consists of the <Excision> of the <Uterus> and <Cervix>, using a <Laparoscope> as an access <Device> via a /single skin incision/.`

**Concept 2**: Type of <Surgical procedure> has_method <Surgical action> has_morphology <Lesion> has_procedure_site <Human anatomy> uses_access_device <Device>

Regarding the concept of <Laparoendoscopic single-site ovarian cystectomy>, the proposal would read:

`<Minimally invasive surgical procedure> which consists of the <Excision> of a <Cyst> located in the <Ovary>, using a <Laparoscope> as an access <Device> via a /single skin incision/.`

Finally, it is believed that EndoTerm's knowledge organisation proposal, grounded by the outlined methodology and theoretical background, will enable an integration with some of the existing biomedical terminological resources dedicated to procedures, especially the ICHI and SNOMED CT. Despite the fact that these resources do not currently encompass any natural language definitions, nor any guidelines or drafting principles, as stated earlier, their solid concept orientation will undoubtedly constitute a valuable framework in that almost inevitable process. And when that happens, it is expected that EndoTerm can help to enhance the yet rather marginal presence of <Laparoendoscopic single-site surgery> - and other endometriosis-related concepts - in existing biomedical terminological resources.

## 6. Concluding remarks

This paper aimed to demonstrate that conceptual representations, in this case an ontology supported by a combination of the specific difference approach and a categorial structure for procedure concepts, can make a valuable contribution to the current lack of natural language definitions in most of the biomedical terminological resources. By providing an organised and clear framework of interrelated concepts, relationships, and domain constraints, these conceptualisations can become useful allies against the limitations of the so-called traditional terminological definitions.

The ongoing changes regarding the way medical information and knowledge are produced, used, stored and shared require efficient and reliable solutions, in a society that demands immediate and multi-platform access to all digital content. If one of the main postulates of terminology work is to provide tools and services that can respond to the concrete needs of a given target audience, at a certain moment in time, within a specific domain, and under particular circumstances, then terminological projects developed within the subject field of healthcare, especially those focusing on knowledge representation, knowledge organisation and knowledge sharing, must take the above-mentioned background into consideration.

## 7. Acknowledgements

This research has been financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade Nova de Lisboa – UID/LIN/03213/2013.

## 8. Bibliographical References

Adamson, D., Kennedy, S., & Hummelshoj, L. (2010). Creating solutions in endometriosis: global collaboration through the World Endometriosis Research Foundation. Journal of Endometriosis, 2(1), 1–46.

Baader, F. (2003). The Description Logic Handbook: Theory, Implementation and Applications. Cambridge: Cambridge University Press.

Benson, T., & Grieve, G. (2016). Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. London: Springer.

Carvalho, S., Costa, R., & Roche, C. (2016). LESS Can Indeed Be More: Linguistic and Conceptual Challenges in the Age of Interoperability. In H. Erdman Thomsen, A. Pareja-Lora, B. Nistrup Madsen, C. B. S. Cbs, Department of International Business Communication and Politics (Eds.), Term Bases and Linguistic Linked Open Data. Copenhagen: hal.archives-ouvertes.fr. Retrieved from http://openarchive.cbs.dk/handle/10398/9323

Carvalho, S., Roche, C., & Costa, R. (2015). Ontologies for terminological purposes: the EndoTerm project. In P. F. Thierry Poibeau (Ed.), Proceedings of the 11th International Conference on Terminology and Artificial Intelligence (pp. 17–27). Universidad de Granada, Granada, Spain, November

4-6, 2015.: Universidad de Granada.

Coiera, E. (2015). Guide to Health Informatics, Third Edition. Boca Raton, FL: CRC Press.

Costa, R. (2013). Terminology and Specialised Lexicography: two complementary domains. Lexicographica, 29(1). https://doi.org/10.1515/lexi-2013-0004

de Bessé, B. (1997). Terminological Definitions. In W. S. Budin (Ed.), Handbook of Terminology Management: Volume 1: Basic Aspects of Terminology Management (pp. 63–74). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Dunselman, G. A. J., Vermeulen, N., Becker, C., Calhaz-Jorge, C., D'Hooghe, T., De Bie, B., European Society of Human Reproduction and Embryology. (2014). ESHRE guideline: management of women with endometriosis. Human Reproduction , 29(3), 400–412.

Durán-Munoz, I. (2016). Producing frame-based definitions. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 22(2), 223–249.

Elhanan, G., Perl, Y., & Geller, J. (2011). A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. Journal of the American Medical Informatics Association: JAMIA, 18 Suppl 1, i36–i44.

Faber, P. (2012). A Cognitive Linguistics View of Terminology and Specialized Language. Walter de Gruyter.

Faber, P. (2015). Frames as a framework for terminology. In H. J. Kockaert & F. Steurs (Eds.), Handbook of Terminology, Volume 1 (pp. 14–33). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Fillmore, C. J. (2003). Double-Decker Definitions: The Role of Frames in Meaning Explanations. Sign Language Studies, 3(3), 263–295.

Fillmore, C. J., & Atkins, B. T. S. (1994). Starting where the dictionaries stop: The challenge for computational lexicography. In B. T. S. Atkins & A. Zampolli (Eds.), Computational Approaches to the Lexicon (pp. 349–393). Oxford: Oxford University Press.

Gill, I. S., Advincula, A. P., Aron, M., Caddedu, J., Canes, D., Curcillo, P. G., Teixeira, J. (2010). Consensus statement of the consortium for laparoendoscopic single-site surgery. Surgical Endoscopy, 24(4), 762–768.

HIMSS. (2013). Definition of Interoperability - Approved by the HIMSS Board of Directors. Healthcare Information and Management Systems Society. Retrieved from http://www.himss.org/sites/himssorg/files/FileDownloads/HIMSS%20Interoperability%20Definition%20FINAL.pdf

ISO. (2000). Terminology work -- Vocabulary -- Part 1: Theory and application (No. 1087-1:2000). Geneva: ISO.

ISO. (2007). Health informatics. Vocabulary for terminological systems (No. 17115:2007). Geneva: International Standardization Organization. https://doi.org/10.3403/30084386

ISO. (2009). Terminology work -- Principles and methods (No. 704). Geneva: International Standardization Organization.

ISO. (2012). Health informatics. Categorial structure for terminological systems of surgical procedures (No. 1828:2012). Geneva: International Standardization Organization. https://doi.org/10.3403/30208974

Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Schriml, L. M. (2015). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Research, 43(Database issue), D1071–D1078.

Löckinger, G., Kockaert, H., & Budin, G. (2015). Intensional definitions. In H. J. Kockaert & F. Steurs (Eds.), Handbook of Terminology - Volume 1 (pp. 60–81). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Maks, I. (2006). Frame-based definitions in a Learners' Dictionary for Dutch Business Language. In P. Ten Hacken (Ed.), Terminology, Computing and Translation (pp. 191–206). Tübingen: Narr.

McCray, A. T., & Bodenreider, O. (2002). A Conceptual Framework for the Biomedical Domain. In The Semantics of Relationships (pp. 181–198). Dordrecht: Springer.

Meyer, I., Bowker, L., & Eck, K. (1992). COGNITERM: An experiment in building a terminological knowledge base. In Proceedings, 5th EURALEX International Congress on Lexicography (pp. 159–172). Tampere, Finland.

Peter Rogers, David Adamson, Moamar Al-Jefout, Christian Becker, Thomas D'Hooghe, Gerard Dunselman, for the WES/WERF Consortium for Research Priorities in Endometriosis. (2016). Research Priorities for Endometriosis: Recommendations From a Global Consortium of Investigators in Endometriosis. Reproductive Sciences , 24(2), 202–226.

Rector, A. L., Bechhofer, S., Goble, C. A., Horrocks, I., Nowlan, W. A., & Solomon, W. D. (1997). The GRAIL concept modelling language for medical terminology. Artificial Intelligence in Medicine, 9(2), 139–171.

Rector, A., & Rogers, J. (2006). Ontological and practical issues in using a description logic to represent medical concept systems: Experience from GALEN. Reasoning Web 2006. Lecture Notes in Computer Science, 4126. Retrieved from http://link.springer.com/content/pdf/10.1007/11837787.pdf#page=207

Rey, A. (1995). Essays on Terminology. John Benjamins Publishing.

Roche, C. (2012). Should Terminology Principles be re-examined? Knowledge Engineering Conference (TKE), P., 17, 32.

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

Roche, C., & Calberg-Challot, M. (2009). Ontoterminology: A new paradigm for terminology. In Proceedings of the International Conference on Knowledge Engineering and Ontology Development. Funchal, Madeira: hal.archives-ouvertes.fr. Retrieved from https://hal.archives-ouvertes.fr/hal-00622132/

Rogers, P. A. W., D'Hooghe, T. M., Fazleabas, A., Gargett, C. E., Giudice, L. C., Montgomery, G. W., Zondervan, K. T. (2009). Priorities for endometriosis research: recommendations from an international consensus workshop. Reproductive Sciences 16(4), 335–346.

Sager, J. C. (1990). Practical Course in Terminology Processing. John Benjamins Publishing.

Sager, J. C. (2000). Essays on Definition (Vol. 4). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sager, J. C., & Ndi-Kimbi, A. (1995). The conceptual structure of terminological definitions and their linguistic realisations: A report on research in progress. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 2(1), 61–85.

Santos, C., & Costa, R. (2015). Domain specificity. In Handbook of Terminology (pp. 153–179).

Seppälä, S. (2007). La définition en terminologie: typologies et critères définitoires. Terminologie & Ontologies: Théories et Applications - Actes de la première conférence TOTh, 23–43.

Sicilia, M.-A., & Balazote, P. S. (2013). Interoperability in Healthcare Information Systems: Standards, Management, and Technology. IGI Global.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Rosse, C. (2005). Relations in biomedical ontologies. Genome Biology, 6(5), R46.

Swanepoel, P. (2011). Improving the Functionality of Dictionary Definitions for Lexical Sets: The Role of Definitional Templates, Definitional Consistency, Definitional Coherence and the Incorporation of Lexical Conceptual Models. Lexikos, 20(0). https://doi.org/10.5788/20-0-151

Temmerman, R. (2000). Towards New Ways of Terminology Description: The Sociocognitive-approach. John Benjamins Publishing.

WHO. (2011). Content Model Reference Guide - ICD-11 alpha (Version 11th revision). Geneva: World Health Organization.

Working group of ESGE, ESHRE, and WES, Saridogan, E., Becker, C. M., Feki, A., Grimbizis, G. F., Hummelshoj, L., De Wilde, R. L. (2017). Recommendations for the surgical treatment of endometriosis-part 1: ovarian endometrioma. Gynecological Surgery, 14(1), 27.

# ELEXIS - European Lexicographic Infrastructure: Contributions to and from the Linguistic Linked Open Data

**Thierry Declerck[1,2], John McCrae[3], Roberto Navigli[4], Ksenia Zaytseva[1], Tanja Wissik[1]**
[1]Austrian Centre for Digital Humanities at the Austrian Academy of Sciences
[2]DFKI GmbH, Multilingual Technologies Lab
[3]Insight Centre for Data Analytics at the National University of Ireland Galway, Ireland
[4]Sapienza University of Rome
[1]Sonnenfelsgasse 19,1010 Vienna, Austria
[2]Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
[3]IDA Business Park, Lower Dangan Galway, Ireland
[4]Via Regina Elena, 295 - 00161 Roma, Italy
[2]declerck@dfki.de, [1]{Ksenia.Zaytseva,Tanja.Wissik}@oeaw.ac.at,
[3]john.mccrae@insight-centre.org, [4]navigli@di.uniroma1.it

**Abstract**

In this paper we outline the interoperability aspects of the recently started European project ELEXIS (European Lexicographic Infrastructure). ELEXIS aims to integrate, extend and harmonise national and regional efforts in the field of lexicography, both modern and historical, with the goal of creating a sustainable infrastructure which will enable efficient access to high quality lexical data in the digital age, and bridge the gap between more advanced and lesser-supported lexicographic resources. For this, ELEXIS will make use of or establish common standards and solutions for the development of lexicographic resources and develop strategies and tools for extracting, structuring and linking lexicographic resources.

**Keywords:** eLexicography, Linguistic Linked Data cloud, BabelNet, OntolexLemon

## 1. Introduction

The field of lexicography has a long tradition of proposing as accurate as possible descriptions of languages. As stated in (Køhler Simonsen, 2017): "Lexicography is a four thousand year old discipline and dictionaries have been an integral part of commerce and human cultural history for centuries".

Since the 1980s, lexicographers have started to utilize computers and to apply computational methods. Online dictionaries are no longer only a reference work, but are also seen as platforms for supporting advanced search facilities. This emerging field of e-lexicography, nevertheless, is still not clearly shaped, and methods and workflows not yet fully agreed on. We see for example in a recent article (Rundell, 2015), in which the author describes the current situation of e-lexicography as being in a transitional phase, a quotation of Robert Lew stating that "It seems that the web community, while enthusiastically embracing the novelty of online collaboration, propagates the traditional model of lexicographic description"[1]. This transitional status is even more patent, when we consider the relations between the fields of lexicography and Natural Language Processing (NLP)[2], although both sides could greatly benefit from each other, as this was already pointed out in (Kilgarriff, 2000). Lexicographic work is also under-represented in the Linked Data (LD) cloud and in Semantic Web technologies.

In recent years, however, new developments have emerged in the field of e-lexicography, like the eLex conference series[3], which started in 2009, the Globalex initiative[4], which was established at eLex 2015 and which organized two workshops at LREC[5], thus directly addressing the Language Technology community, or the recently ended ENeL COST action[6], which is described below in more details.

In 2013, the European lexicographic community was brought together for the first time in the European Network of e-Lexicography (ENeL) COST action. This initiative was set up to improve the access for the general public to scholarly dictionaries and make them more widely known to a larger audience. In the context of this network, a clear need emerged for a broader and more systematic exchange of expertise, for the establishment of common standards and solutions for the development and integration of lexicographical resources, and for broadening the scope of application of these high quality resources to a larger community, including the Semantic Web, artificial intelligence, NLP and Digital Humanities. For describing such an integrative approach, the term "virtuous circle" re-emerged, as it characterizes very well the intended spiralling development of lexicographic data on the basis of a cross-disciplinary exchange of knowledge and the incremental contributions of the different methods and technologies to be involved.

We write "re-emerge", as the term was already coined in (Kilgarriff, 2000): "In the best of all possible worlds, computational enhancement and lexicographical upgrading would build upon each other in a virtuous circle that knew

[1]The quotation was taken from (Lew, 2014).

[2]In this paper, we will use the terms NLP or Language Technology (LT) interchangeably.

[3]https://elex.link/

[4]https://globalex.link/

[5]http://ailab.ijs.si/globalex/

[6]http://www.elexicography.eu/.

no bounds". The implementation of such a virtuous cycle for the generation of high-quality e-lexicographic resources is a central objective of the recently started ELEXIS project, described in section 2.

## 2. ELEXIS

ELEXIS (European Lexicographic Infrastructure) is fostering cooperation and information exchange among lexicographical research communities. The infrastructure is a newly granted project under the H2020-INFRAIA-2016-2017 call, with the topic "Integrating Activities for Starting Communities", and started in February 2018[7].

ELEXIS is building on infrastructures defined in other projects and initiatives, especially CLARIN[8] and DARIAH[9], which allow language or Digital Humanities resources (both tools and data) to be shared. In this, the partners of ELEXIS will get support for easily sharing their lexicographic resources, yet this does not necessarily lead to any interoperability of such resources. In order to support interoperability, ELEXIS enables stakeholders to encode their lexicographic data with common concepts and entities from models such as BabelNet[10], DBpedia[11] or Wikidata[12], which are accessible as nodes in the Linked Data cloud[13].

Moreover, to ensure that there is integration of lexical resources at even the most basic level, ELEXIS will define a minimal common data model capturing the core concepts of a lexicographic resource such as entries (single-word, multi-word), senses, syntactic and semantic frames, etymologies etc. and linguistic relationships such as synonymy/antonymy, translation, domain/region/register classification, relatedness, etc. that will be compatible with existing models used in the community, including TEI[14], Wikidata[15], LMF[16] and OntoLex-Lemon[17]. The data converted to this model will be available in RDF[18], facilitating linking and publishing on the Web as linked data.

A key goal of the ELEXIS project is thus to enable stakeholders to link their existing lexicographic resources, either as dictionaries or as standalone lexical descriptions encoded, and so to create a huge multilingual registry, a kind of "Matrix Dictionary" (see Section 3.) that connects lexicographic resources across common concepts. A possible infrastructure for hosting this registry is the Linguistic Linked (Open) Data cloud, which is described in section 4.. In this scenario, ELEXIS would also follow the W3C recommendations for "accessing, updating, creating and deleting resources from servers that expose their resources as Linked Data", as those are stated by the Linked Data Platform (LDP)[19].

## 3. A Matrix Dictionary for ELEXIS

A key goal of ELEXIS is the creation of a "Matrix Dictionary", that will be formed of links created between lexicographic resources in different languages, domains and forms. With this, ELEXIS will create a universal repository of linked senses, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in all types of existing lexicographic resources, monolingual, multilingual, modern, historical, etc. In order to reach this goal, ELEXIS will develop strategies, tools and standards for extracting, structuring and linking the high quality semantic data from lexicographic resources and make them available to the Linked (Open) Data family. Those processes are necessary, as current lexicographic resources, both modern and historical, have different levels of structure and are not equally suitable for applications in advanced NLP technologies, for which they should be disclosed to or from which they could benefit.

The project will also work on interlinking lexical content with other structured or unstructured data – corpora, multimodal resources, etc. – on any level of lexicographic description: semantic, syntactic, collocational, phraseological, etymological, translation equivalents, examples of usage, etc. By creating an integrated, linked and interlinked resource, a huge amount of high quality lexical data will not only become available to the linguistic, NLP and Semantic Web communities, it will also facilitate cutting-edge research in Digital Humanities.

This will be achieved by creating an infrastructure dedicated to automatic segmentation and structuring of content for dictionaries that are currently produced in digital environments but are typically encoded in their own custom data format. ELEXIS conversion and alignment tools will provide users of the infrastructure with the possibility to harmonise and convert their lexicographic resources to a uniform data format that allows their seamless integration in Linked Open Data.

This infrastructure is responding to one of the missions of ELEXIS consisting in enabling the integration of (big) data in different modalities into the lexicographic process, pre-

[7] See http://www.elex.is/.

[8] See https://www.clarin.eu/.

[9] See https://www.dariah.eu/.

[10] See http://babelnet.org/ and (Navigli and Ponzetto, 2012).

[11] See Seehttp://wiki.dbpedia.org/. See also (Unger et al., 2013) for a first study on how to publish a DBpedia based ontology lexicon as linked data.

[12] See Seehttps://www.wikidata.org/wiki/Wikidata:Main_Page.

[13] See http://linkeddata.org/ for more details.

[14] TEI stands for "Text Encoding Initiative". See http://www.tei-c.org/index.xml.

[15] See https://www.wikidata.org/wiki/Wikidata:Main_Page.

[16] LMF stands for "Lexical Markup Framework", an ISO standard. See http://www.lexicalmarkupframework.org/.

[17] OntoLex-Lemon is the result of a W3C Community Group, building on and extending LMF and an earlier version of *lemon* (lexicon model for ontologies, (McCrae et al., 2012)). See https://www.w3.org/2016/05/ontolex/ for the final W3C Community report and (McCrae et al., 2017) for the current status of OntoLEx-Lemon.

[18] RDF stands for "Resource Description Framework", a W3C standard model for interchanging data on the Web. It is a building stone for the realisation of the Linked Data cloud (see for this point http://linkeddata.org/faq).

[19] The source we quoted from LDP: https://www.w3.org/TR/ldp/.

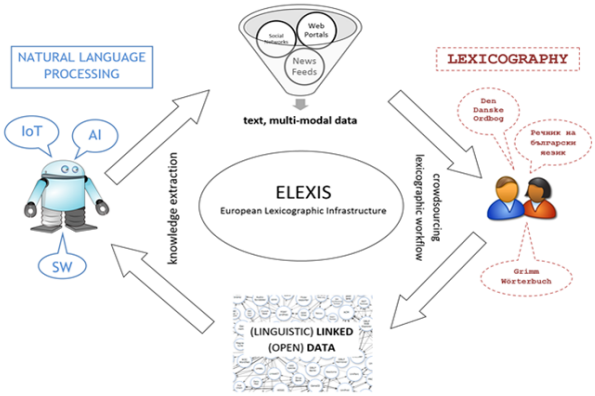



Figure 1: The virtuous cycle of e-lexicography

pared and visualised for human end users. Figure 1 is displaying this development, which is of cyclic nature.
The existence of common data models and standards that are produced bottom-up from within the lexicographic community fostered by ELEXIS is a necessary condition for the successful development of the whole platform. Standards will be developed and tested during the project on the data provided by the lexicographic partners and implemented in the newly-developed service.

## 4. Linguistic Linked Open Data

The Linguistic Linked Open Data (LLOD)[20] is an initiative started by the Open Linguistics Working Group (OLWG)[21] aims at breaking the data silos of linguistic data and thus encourage NLP applications that can use data from multiple languages, modalities (e.g., lexicon, corpora, etc.) and develop novel algorithms. Figure 2 gives a partial view of the current state of the LLOD cloud.

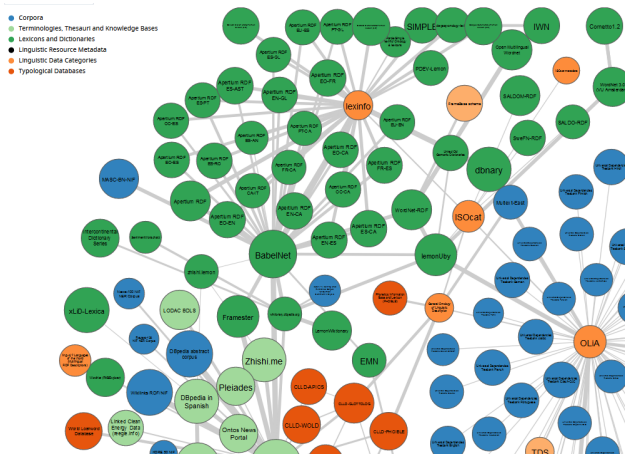



Figure 2: A (partial) view on the Linguistic Linked Open Data cloud, July 2017.

The rapid development of the LLOD cloud[22] was also supported by the European LIDER ("Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe") project[23]. LIDER has set up some basis for the further development of the Linguistic Linked Open Data and published a series of guidelines on how to publish linguistic data in the Linked Data framework. Those guidelines are used in relation to the task of making the LLOD actionable for language intensive use cases, with a focus on multilingual application. Those guidelines will be used and extended in the context of ELEXIS.
A cooperation established between LIDER and the aforementioned ENeL Cost Action, also in the form of short term exchanges of junior researchers and of the participation of ENeL members to a datathon organized by LIDER[24] has been in fact instrumental in the formulation of some of the central objectives of the ELEXIS project, which will also stress the need of community integration besides the technological one, whose description is the focus of this paper.
The successful development of the LLOD is also based and linked to the development of the Lexicon Model for Ontologies (*lemon*)[25] and its successor the OntoLex-Lemon model[26]. And although *lemon*, which stands for "LExicon Model for ONtologies", was originally developed in order to model language data used in ontologies, experience has shown that *lemon* or OntoLex-Lemon can indeed be used for modelling lexicographic data[27] or some specific lexical phenomena[28].

## 5. Further Developments of the LLOD and OntoLex-Lemon within ELEXIS

While looking in details at the current state of the Linguistic Linked Open Data (LLOD)[29], one can see that the data sets published in this cloud are classified along the lines of six categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

[20] See `http://linguistic-lod.org/`.
[21] See (Chiarcos et al., 2012) and (McCrae et al., 2016).
[22] The full LLOD cloud can be accessed at `http://linguistic-lod.org/llod-cloud`. There one can click on the various nodes and get more details about the data sets represented by the "bubbles".
[23] LIDER was an FP7 Coordination and Support Action from 2013-11-01 to 2015-12-31. See also `http://lider-project.eu/`.
[24] See `http://datathon.lider-project.eu/`.
[25] See (McCrae et al., 2012)
[26] `https://www.w3.org/2016/05/ontolex/`. See also for a kind of historical view on the development of *lemon* towards OntoLex-Lemon (McCrae et al., 2017).
[27] See (Declerck et al., 2017) or (Tiberius and Declerck, 2017).
[28] See (Declerck and Lendvai, 2016).
[29] See again `http://linguistic-lod.org/llod-cloud`.

However, as of today, the LLOD is not populated by many lexicographic resources, due to the lack of a dedicated infrastructure for resource interlinking and of effective ontology alignment algorithms, which depend on multilingual semantic similarity, entity linking and word sense disambiguation.
One goal of ELEXIS can be to have a specific lexicographic category containing its specific data sets and linking those to both data sets included into the other LLOD categories and to data sets included in the global Linked Data cloud[30]. This step is responding for example to insights described in (Gracia et al., 2017), in the abstract of which we can read: "[...] future dictionaries could be LD-native and, as such, graph-based. Their nodes are not dependent on any internal hierarchy and are uniquely identified at a Web scale". ELEXIS will address this view on the generation of linked data-native dictionaries and facilitate their publication in the LLOD cloud as lexicographic data sets. (Declerck, 2018) proposes a similar approach, but considering all types of lexical data, not only those included in a dictionary.
As the development of the LLOD cloud is closely related to OntoLex-Lemon and related vocabularies, a working group was built in order to study the representation of lexicographic data (sets) and to propose a lexicographic module to be added to Ontolex-Lemon[31], so that their linking to all types of lexical data covered by the Ontolex-Lemon Model is guaranteed. In this, ELEXIS partners are contributing to standardisation of the formalisation of lexicographic data.

## 6. Interoperability and Quality

To provide conceptual interoperability, services enabling linking of ELEXIS lexicographic resources will be developed and made available in the ELEXIS linking tools segment of the platform (see Figure1). This will provide the possibility to link lexical entries, senses and fundamental concepts in different lexical resources, using a semi-automatic approach. BabelNet[32], as an existing multilingual resource to provide cross-lingual linking, will be exploited for this purpose. Extensive linking of existing lexicographic resources by pivoting through BabelNet will enable the creation of what we call the ELEXIS matrix dictionary[33]. Data from this new resource will be available through ELEXIS matrix dictionary RESTful Web service as part of the platform.
This work will be achieved through four principle steps:

- **Common access and models**: We will define a set of common protocols, in the form of REST API calls that can allow dictionaries involved in the project to be accessed through a single interface. This model will be based on existing web standards and models including RDF, SPARQL and OntoLex-Lemon. Furthermore, the task will define common metadata and concept properties for use within the project. The outputs of this task will be technical documentation describing the formats and tools to allow resources in OntoLEx-Lemon RDF or TEI to be compliant with this protocol.
- **Semi-automatic dictionary linking**: Linking lexical resources is a challenge that requires impractical amounts of human efforts, but is still not easy to solve automatically. We will develop a semi-automatic system that will make the linking problem viable for large resources, by using state-of-the-art semantic and natural language processing techniques, especially deep learning methods such as LSTMs (Tai et al., 2015), with a human in the loop. Furthermore, we will apply constraint-based optimisation of the linking, which can quickly find the correct mapping in an active learning setting with only a small amount of human input. As such, we will develop a single tool where a user can upload two lexical resources and interactively link them. We will then evaluate this tool by developing gold standard mappings in the context of a shared task.
- **Cross-lingual linking through BabelNet**: In order to link lexical resources across languages, we will use one highly multilingual lexicon, BabelNet, as the basis for a cross-lingual mapping system. As such, we will extend our linking tools to cross language boundaries by pivoting through BabelNet. We will further allow for resources linked through BabelNet to be used to be submitted to BabelNet, so they can extend the resource in future releases.
- **Validation and quality assurance**: We will develop tools to automatically verify the quality of lexical resources at three levels: Firstly, the *technical quality*, which means ensuring that the resource maintains the validity of its output and does not make errors in encoding, this will be achieved by Web services that validate TEI and RDF data as provided by producers. Secondly, *operational quality* ensures that the lexical resources remain available and responsive as they are deployed on the Web, in particular, a service will measure uptime of each resource. Finally, *scientific quality* ensures that the results of the service are correct in the task they try to perform and will work by creating benchmarks for tasks in NLP, with Web services to automatically check resource performance against existing gold standards.

## 7. Lexicography for Natural Language Processing

To show the effectiveness of the interlinking across lexical resources, ELEXIS will study the impact of an enriched LLOD on several NLP tasks:

- **Multilingual Word Sense Disambiguation:** a long-standing issue of supervised Word Sense Disambigua-

[30] `http://lod-cloud.net/`. There the reader can observe that "Linguistics" is listed as a domain-specific sub-set of the cloud.

[31] See (Bosque-Gil et al., 2017) and for the current state of the discussions on the lexicographic module `https://www.w3.org/community/ontolex/wiki/Lexicography`.

[32] See again `http://babelnet.org/` and (Navigli and Ponzetto, 2012).

[33] The motivation behind the ELEXIS matrix dictionary has been described in 3.

tion (WSD) – the task of automatically determining the meaning of words occurring in context – is that huge amounts of sense-annotated sentences need to be manually created. This endeavour, which as of today, is incomplete even for English, needs to be repeated for each new domain and language, something that makes the task arduous to replicate in most European The ELEXIS lexicographic resources will be utilized to bootstrap large training datasets for WSD in dozens of languages.

- **Multilingual Semantic Parsing:** semantic parsing aims to map sentences to formal representations of their meaning. It has deeper relationships to syntactic parsing than WSD. However, most semantic parsing approaches in the literature either work in a supervised fashion with even higher annotation costs than those of WSD or require knowledge resources such as DBpedia or Wikidata which seem to work only in domain-restricted specific tasks such as question answering. In ELEXIS we will develop innovative algorithms that exploit the huge multilingual network of interlinked lexical knowledge to perform multilingual semantic parsing.

- **Word sense clustering:** where development of semi-automatic procedures to bring together subtle sense distinctions in clusters of meanings will be shown to improve the performance of tasks such as Word Sense Disambiguation;

- **Domain labelling of text:** where the aggregated information obtained from the lexicographic network of resources will be shown to improve automatic tagging of text with domain labels in arbitrary languages thanks to developing innovative neural techniques.

- **Study of the diachronic distribution of senses:** the use of the most frequent sense in NLP is a solid baseline used in WSD and other tasks. However, it is not systematic and it is useful only for the English language. We will develop novel techniques for aggregating the predominance information of senses a) from the multitude of resources b) considering evolution over time, so as to have important impact on disambiguation and corpus analysis.

## 8. The User Perspective

While the description of the foreseen ELEXIS platform can at first sight look like an academic exercise, it should be stressed that the project responds to the needs formulated by publishers and other professionals in the e-lexicographic field. Some of those needs were already articulated by industrial/commercial partners in the ENeL Cost action. The changing technological context calls not only for adaptations of the lexicographic workflows but also for the establishment of new business models, as this is for example expressed in (Køhler Simonsen, 2017). We quote from this eLex 2017 paper: "[...] the biggest problem of lexicography is that lexicographic products are no longer perceived as relevant for the vast majority of people. Most people in fact do not use dictionaries, and if they need to find help when communicating or when looking for data, they simply use the Internet instead. So dictionaries are in fact not being used as much as we want them to be. The most important question is: why do not people use online or mobile dictionaries? Obviously, there are a number of reasons, but I would argue that the most important reason is that most lexicographic resources are not tool-integrated and not specifically related to the user's job tasks".

We can see that ELEXIS is (at least partly) responding to this situation if we compare some of the technological goals of ELEXIS with the six theses that are formulated in (Køhler Simonsen, 2017) and which describe the components of what could and should be the ingredients of a viable business model for the modern e-elexicography. In those 6 theses, (Køhler Simonsen, 2017) requires among others that lexicographic products are moving to lexicographic services, the integration of lexicographic data in lexicographic platform and distribution, and to take increasingly into account the lexicographic users and their needs. Topics that are at the core of ELEXIS, as well as the move "from dictionary to lexicographic data in software [and] artificial intelligence".

## 9. Conclusion

We described in this paper the main technological challenges that the ELEXIS project will try to solve, based on existing initiatives, projects, infrastructures and standards. A focus will be on the interrelation between e-lexicography and the technologies used in the context of the Linguistic Linked Data, in order to generate high-quality lexicographic data that can be then immediately re-used in NLP and Semantic Web applications, which are themselves based on the LLOD. ELEXIS implements thus a "virtuous cycle" scenario to make sure that lexicographic resources and expertises are played a central role in high-quality language applications, also beyond the era of dictionary-based lexicographic products.

## Acknowledgements

This work was supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015 and by Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

## 10. Bibliographical References

Bosque-Gil, J., Gracia, J., and Montiel-Ponsoda, E. (2017). Towards a module for lexicography in ontolex. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017.*, pages 74–84.

Chiarcos, C., Hellmann, S., and Nordhoff, S., (2012). *Linking Linguistic Resources: Examples from the Open Linguistics Working Group*, pages 201–216. Springer Berlin Heidelberg, Berlin, Heidelberg.

Declerck, T. and Lendvai, P. (2016). Towards a formal representation of components of german compounds. In Micha Elsner et al., editors, *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Humboldt University, ACL, 8.

Declerck, T., Tiberius, C., and Wandl-Vogt, E. (2017). Encoding lexicographic data in lemon: Lessons learned. In John P. McCrae, et al., editors, *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*. CEURS, 8.

Declerck, T. (2018). Towards a linked lexical data cloud based on ontolex-lemon. In John P. McCrae, et al., editors, *Proceedings of the 6th Workshop on Linked Data in Linguistic (LDL-2018)*. ELRA, 5.

Gracia, J., Kernerman, I., and Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2017 conference*, pages 550–559. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.

Kilgarriff, A. (2000). Business models for dictionaries and nlp. *International Journal of Lexicography*, 13(2):107–118.

Køhler Simonsen, H. (2017). Lexicography: What is the business model? In Iztok Kosem, et al., editors, *Electronic Lexicography in the 21st Century*, pages 395–415. Lexical Computing CZ s.r.o.

Lew, R. (2014). User-generated content (ugc) in online english dictionaries. *OPAL - Online publizierte Arbeiten zur Linguistik*, 4:8–16.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., Garcia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719.

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A., and Pool, J. (2016). The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In *10th Language Resource and Evaluation Conference (LREC)*, pages 2435–2441.

McCrae, J. P., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In Iztok Kosem, et al., editors, *Proceedings of eLex 2017*, pages 587–597. INT, Trojína and Lexical Computing.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Rundell, M. (2015). From print to digital: Implications for dictionary policy and lexicographic conventions. *Lexikos*, 25(1).

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1556–1566.

Tiberius, C. and Declerck, T. (2017). A lemon model for the anw dictionary. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2017 conference*, pages 237–251. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.

Unger, C., McCrae, J. P., Walter, S., Winter, S., and Cimiano, P. (2013). A lemon lexicon for dbpedia. In *Proceedings of the NLP & DBpedia workshop co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 22, 2013.*

# The Diachronic Semantic Lexicon of Dutch as Linked Open Data

**Katrien Depuydt, Jesse de Does**
Instituut voor de Nederlandse Taal
Rapenburg 61, 2311GJ Leiden, The Netherlands
katrien.depuydt@ivdnt.org, jesse.dedoes@ivdnt.org

**Abstract**

This paper describes the Linked Open Data (LOD) model for the diachronic semantic lexicon DiaMaNT, currently under development at the Instituut voor de Nederlandse Taal (INT; Dutch Language Institute). The lexicon is part of a digital historical language infrastructure for Dutch at INT. This infrastructure, for which the core data is formed by the four major historical dictionaries of Dutch covering Dutch language from ca. 500 - ca 1976, currently consists of three modules: a dictionary portal, giving access to the historical dictionaries, a computational lexicon GiGaNT, providing information on words, their inflectional and spelling variation, and DiaMaNT, aimed at providing information on diachronic lexical variation (both semasiological and onomasiological). The DiaMaNT lexicon is built by adding a semantic layer to the word form lexicon GiGaNT, using the semantic information in the historical dictionaries. Ontolex-Lemon is a good point of departure for the LOD model, but we need extensions to be able to deal with the historical dictionary content incorporated in our lexicon.

**Keywords:** Linked Open Data, Ontolex, Diachronic Lexicon, Semantic Lexicon, Historical Lexicography, Language Resources

## 1. Background

Even though Dutch lexicography[1] can be dated back to the 13th century with the glossarium Bernense, a Latin-Middle Dutch word list, we had to wait until the 19th century for a more systematic and academic description of Dutch language. Two important dictionary projects were initiated by Matthias de Vries: a scholarly dictionary of Middle Dutch language, the *Middelnederlandsch Woordenboek* (MNW) (Dictionary of Middle Dutch) and the *Woordenboek der Nederlandsche Taal* (WNT) (Dictionary of the Dutch Language).

The MNW was compiled by E. Verwijs and J. Verdam and published between 1885 and 1929; a list of sources and a volume on dike building, water management and related terms by A. Beekman were added between 1927 and 1952. De Vries himself worked as editor-in-chief on the WNT, for which he made the design in 1852, until his demise in 1892. The first fascicle of the dictionary was published in 1864. The dictionary was finished in 1998, followed by three supplemental volumes in 2001.

Both dictionaries cover Dutch language from ca. 1250 until 1976. They were based on a corpus of quotations, written on slips of paper, and published in print. In 1995, the WNT was also released on CD-ROM, with a final release of the complete dictionary in 2003. The MNW was published on CD-ROM in 1998, accompanied by a collection of historical texts.

The former Instituut voor Nederlandse Lexicologie (Institute for Dutch Lexicology), founded in 1967 to host the WNT, decided to complete the description of historical Dutch by means of two separate projects, the *Vroegmiddelnederlands Woordenboek* (Early Middle Dutch Dictionary; 1988-1999), covering Dutch language from 1200-1300 and the *Oudnederlands Woordenboek* (Dictionary of Old Dutch, 1999-2009), covering the oldest Dutch language period from 500-1200. Both dictionaries were born digital, based on a closed corpus, in digital format. Having four scholarly dictionaries of Dutch in digital format opened up opportunities for further exploitation of the contents of these dictionaries.

### 1.1. Online Dictionary Portal (gtb.inl.nl)

The first step was to publish the dictionaries online in a dictionary portal (gtb.inl.nl)[2], which had its first release in 2007 (Depuydt and De Does, 2008). This application mainly supports semasiological search; most users use it to look up the meaning of a word. There is no dictionary of Dutch which describes the complete language period in the way the *Oxford English Dictionary* does for English, so combining all four dictionaries in a portal was the closest we could get to providing a diachronic lexicographic overview of Dutch language. A major challenge was to give the user optimal access to the dictionary information, without compromising the uniqueness of each individual dictionary. For this module, not only the dictionary software application was designed and built, but a lot of work went also into semi-automatic processing of the data to make the dictionary content suitable for searching. The data was converted into TEI XML. Easier access to the dictionary content was provided, among other things by adding a modern Dutch equivalent to each entry in the dictionaries. This does not only enable combined searching in several dictionaries by one single query, it also relieves users of the burden of having to search by one particular historical spelling of a lemma.

[1] For an elaborate description of the history of Dutch lexicography, see Mooijaart 2013.

[2] The first component is the online historical dictionary portal (gtb.ivdnt.org), of which the first module was released in 2007 by bringing the WNT online. In separate steps, the MNW, VMNW and ONW were processed and added and the data and application have had several updates.

### 1.2. GiGaNT: a Diachronic Morphosyntactic Lexicon

Also in 2007, work started on the design of the computational lexicon module GiGaNT[3] (Groot Geïntegreerd Lexicon van de Nederlandse Taal; large integrated lexicon of the Dutch language). A computational lexicon gives structured information on vocabulary and has to be suitable for use by computer software. GiGaNT provides information on words, their inflectional and spelling variation, and is aimed to cover Dutch language from the 6th century until present-day. The original aim of GiGaNT was to build a lexicon to support annotation of historical corpus material with part of speech (PoS) and lemma, so as to make these corpora better searchable. However, it can also be used to exploring new corpus material in order to harvest new material, not yet described in the available dictionaries. The lexicon has already been made available in a lexicon service, used for query expansion. A good example is the way a user gets suggested potential variants of a search word in the online historical material of the KB (Dutch Royal Library), in www.delpher.nl or in the Dutch national project (www.nederlab.nl) where a historical corpus is being compiled and put online. The lexicon is also used in Nederlab to establish the link between text material and the online historical dictionaries. Using the historical dictionaries as a primary resource for the GiGaNT lexicon was a logical thing to do. It is a very efficient way to build a historical computational lexicon. Each dictionary contains quotation material for which in each quotation, there is an occurrence of the dictionary entry in a particular form, so automatic detection of the correct word form belonging to the dictionary entry is comparatively easy.

### 1.3. DiaMaNT: a Diachronic Semantic Lexicon

The infrastructure as described above, offers users the means to find out the meaning of a historical word, and gives information on potential spelling and form variation, by means of which searching historical text is made easier. Having the option to search via a modern lemma form also simplifies searching in historical dictionaries and text. From the point of view of INT, another advantage is that it contributes to the structuring of the lexicographical description of the Dutch vocabulary. It gives a more systematic view on what is described, and allows easier detection of inconsistencies and gaps.

To take the infrastructure to the next level, however, would mean finding a solution to resolve one more aspect of the historical language barrier, which is not related to historical variation in form, but to historical variation in vocabulary and meaning. How can we give users the means to search in historical texts for a concept for which he or she only knows the modern Dutch term? In its most simple form, given a certain word, a user ought to get suggestions for potential synonyms of that word, combined with information on the time period in which a particular word was used. And it would even be better if we were able to allow users to look for words with a specific meaning. And can we offer historical linguists better means to study diachronic semantic variation in a systematic way?

This is why in 2015, work on the third module of the infrastructure was started, the diachronic semantic computational lexicon of Dutch (DiaMaNT, Diachroon seMantisch lexicon van de Nederlandse Taal). The main purpose of this lexicon is to enhance text accessibility and foster research in the development of concepts, by interrelating attested word forms and semantic units (concepts), and tracing semantic developments through time. In the lexicon, the diachronic onomasiology, i.e. the change in naming of concepts and the diachronic semasiology, i.e. the change in meaning of words, will be recorded in a way suitable for use by humans and computers. The onomasiological part of the lexicon is designed to enhance recall in text retrieval by providing different verbal expressions of a concept or related concepts (slager → beenhouwer, beenhakker, vleeshouwer (synonyms for 'butcher'); boer → landman ('synonyms for 'farmer'). The diachronic semasiological component (which charts semantic change), aims to enhance precision by enabling the user to take semantic change into account; the oldest meaning of apple for example is 'a fruit' (so apple is also used for pears, plums etc.). The lexicon is built by adding a semantic layer to the word form lexicon GiGaNT, using the semantic information in the historical dictionaries, i.e. the definitions from the dictionary articles from which the word form lexicon is built.

## 2. DiaMaNT as Linked Data

An important impulse for the deployment of DiaMaNT comes from the Dutch CLARIAH project[4]. One of the aims of the technical infrastructure of this project is to offer a generic linked open data graph, populated with entities relevant for the humanities like persons, locations and concepts, for network analysis, data annotation and linking purposes. The concept-entity graph has to provide the basis of a Dutch thesaurus for semantically related terms over time and DiaMaNT is the core of this graph. The lexicon will also be part of the CLARIAH infrastructure for linguistic resources, which enables federated search scenarios in which information from corpora, treebanks and lexica can be combined.

Publishing as Linked Open Data (LOD) facilitates this type of interoperability and integration of lexical resources (Chiarcos, 2003). The LOD paradigm provides a framework that facilitates information integration, and thus, interoperability, by ensuring that entities can be addressed in a globally unambiguous way using Unique Resource Identifiers (URIs), that entities can be accessed over HTTP, and that the descriptions of entities and links between them can be represented according to the W3C Resource Description Framework (RDF) standard (Berners-Lee, 2006).

The CLARIAH context was an important argument to go for a lexicon development strategy which would allow intermediate releases of the lexicon. So far, a project internal release has been done of the lexicon, containing synonym information extracted from the dictionary definitions. The

[3] The situation is now that two modules (based on MWN and WNT) have been released and work on the modules based on ONW and VMNW is scheduled for 2018.

[4] www.clariah.nl

basic LOD model of the lexicon has also been designed. And some exploratory research has been done into the potential distributional semantics offers for lexicon development and deployment.

## 2.1. DiaMaNT Source Data

The lexicon adds a semantic layer on top of the word form lexicon GiGaNT. Both DiaMaNT and GiGaNT have the historical dictionaries of Dutch as a base. The elements from the dictionaries used to create the computational lexica are: entry (historical form and modern Dutch equivalent), PoS, quotations and definitions. These elements are encoded in the TEI XML underlying the online dictionaries. The number of entries, quotations and definitions in the four dictionaries is given in table 1.

The core of the lexica is the corpus of quotations, present in the dictionaries. They illustrate the spelling, the morphological variation and the meaning of an entry as described by the lexicographers. Every quotation in the dictionaries has metadata, describing the provenance of the quotation. The quotations are dated and in all dictionaries but the WNT, also location information is provided. Each occurrence of the main structural elements in the TEI XML has its own persistent ID. In both GiGaNT and DiaMaNT, these persistent ID's are retained. In GiGaNT, the occurrences of an entry in each quotation have been detected and stored, together with the quotations and their metadata. Each word form has been given the correct analysis (lemma and main PoS). This means that in some cases, dictionary entries that in fact describe several lexical entries from the point of view of a computational lexicon, were thus split up.

Since DiaMaNT provides a semantic layer on top of GiGaNT, the word forms of GiGaNT are included in the lexicon in order to make the lexicon more suitable for text retrieval by query expansion. The aim is to develop a thesaurus (diachronic wordnet), where synonym clusters represent the concepts for which lexicalisations are described in the dictionaries. In the current prototype, a first semantic annotation layer on top of the entries and senses in the dictionaries consists of synonyms automatically extracted from the dictionary definitions from MNW and WNT. It is not yet a unified semantic resource, but both MNW and WNT entries are interlinked by a manually verified set of correspondences that go beyond the homograph level. The temporal information is provided by the metadata that comes with the quotations providing the lexicographical evidence for the definitions from which the synonyms are extracted.

## 2.2. Ontolex-Lemon

A standard for the representation of lexical data in RDF is Ontolex-Lemon[5], developed by the Ontology Lexicon (Ontolex) community group (Ciminiano et al., 2016). The model is designed to give linguistic grounding to ontologies, by linking the ontology to lexical entries with grammatical and/or semantic information. The Ontolex community group is currently working on a module dedicated to lexicographical data[6] (Bosque-Gil et al., 2017; McCrae et al., 2017). Even though DiaMaNT is not a mere conversion of historical dictionaries into RDF, there is enough traditional dictionary content in DiaMaNT to be confronted with similar issues, like how to deal with sense hierarchy, how to model diachrony, etc. (cf. Khan et al., 2016, 2017; Bosque-Gil et al., 2017).

We will not describe the complete LOD model for DiaMaNT. Instead, we want to focus on those components that are essential for our lexicon building approach, and for which we had to define extensions to the model.

The main objective for the implementation of the data model for the lexicon is to do justice to the character of the underlying scholarly lexicographical work. The core of our lexica is the corpus of attestations from the historical dictionaries. By analysing the corpus material, using their expert knowledge, lexicographers provided a careful description of the meanings of each word in the dictionary. According to Kilgarriff (1997) "the scientific study of language should not include word senses as objects in its ontology. Where 'word senses' have a role to play in a scientific vocabulary, they are to be construed as abstractions over clusters of word usages." For him "the basic units are occurrences of the word in context (operationalised as corpus citations)." The senses from the dictionaries we use in our DiaMaNT lexicon, and the ontological layer we add to it, remain an interpretation of historical language that came down to us via text. This motivates the extensions we propose to the Ontolex-Lemon model. Senses, lexical entries, lexical forms, and temporal information are linked to attestations. Keeping the complete description of the senses, including the hierarchy, of the lexical entry, is also motivated by the desire to contextualise. Likewise, provenance information concerning the data processing for the DiaMaNT lexicon is included in the lexicon.

We will now give a brief the description of how attestations, sense hierarchy and provenance are modeled for DiaMaNT.

## 2.3. Attestations

Figure 1 shows how we link evidence ("attestations") to lexical categories which we conceive as interpretations for which the dictionary quotations (or corpus references) provide evidence. The main elements of the lexical entry (*LexicalEntry* itself, *Form* and *LexicalSense*) are assigned to the superclass *LexicalPhenomenon* (the name is maybe not very elegant, *Observable* might be another option).

In this way, the dictionary quotations can be seen as a partially semantically tagged corpus.

Table 2 shows part of our efforts to "put the corpus into the dictionary" (Kilgarriff, 2005) by means of the standoff corpus annotation approach of NIF[7] ontology, and to define a suitable metadata model on top of Dublin Core[8]. Unprefixed class and property names are extensions we had to resort to. The extensive quotation metadata is the main ingredient for the temporal and spatial dimensions of the lexicon. In contrast with the lemonDIA model (Khan, 2016),

[5] www.w3.org/2016/05/ontolex

[6] www.w3.org/community/ontolex/wiki/Lexicography

[7] persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html

[8] dublincore.org/

| dictionary | lemmata | definitions | quotations | tokens |
|---|---|---|---|---|
| ONW | 9.268 | 12.619 | 30.025 | 1.056.926 |
| VMNW | 25.946 | 102.202 | 194.366 | 6.463.868 |
| MNW | 74.773 | 144.714 | 400.619 | 13.078.231 |
| WNT | 467.288 | 553.672 | 1.667.835 | 51.246.034 |
| *Total* | 577.275 | 813.207 | 2.292.845 | 71.845.059 |

Table 1: Content statistics of the historical dictionaries

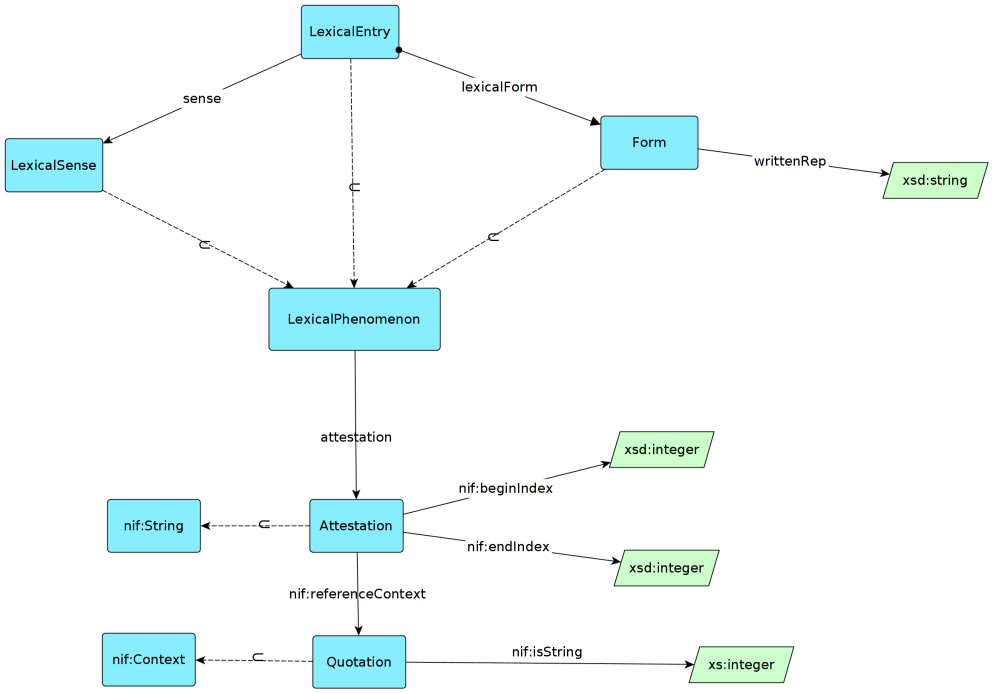



Figure 1: Attestations

the CLARIAH DICOLOD project[9] (Maks et al., 2016) and Cimiano et al., 2013, in which lexical senses are assigned to a time period, we aggregate this information from observed usage.

## 2.4. Senses, Subsenses and Definitions

Many scholarly dictionaries have a hierarchical subdivision of the sense section, mostly (but not exclusively - grammatical distinctions also play a role) based on semantic criteria. One might wonder whether it makes sense to model this subdivision in the more strictly structured semantic lexical infrastructure we work towards.

Despite the somewhat fuzzy semantic significance of the hierarchy, we think it makes sense to include it in the lexicon. Human perusal of, for instance, the result of a query over the data which presents an unstructured list of senses, immediately prompts the desire to know their position in the article hierarchy. Moreover, we have a usage and evidence-based view of meaning. A sense hierarchy implies a semantically motivated hierarchical subdivision of the evidence (set of quotations and their metadata in the entry). NLP applications like word sense disambiguation profit from the possibility of defining a coarse-grained division. Although the hierarchical information requires postprocessing to make it optimally suitable for this purpose, discarding it would entail unwarranted loss of information.

We briefly describe the sense-related part of the model. In agreement with the core Ontolex model, we use the *reference* property to refer from a lexical sense to a concept in an external ontology[12], and synsets are modeled by sharing Lexical Concepts. We encode the sense hierarchy by means of a (non-transitive) property *subsense* (in the Lemon namespace) and, like Bosque-Gil et al. 2017, an integer-valued data property *senseOrder* is attached to the sense nodes. Attaching the order information in this way implies that a sense cannot be shared among lexical entries, which is not a problem in our setting, as we

[9] github.com/cltl/clariah-vocab-conversion

[10] A terminus post quem is the earliest possible date something may have happened,

[11] A terminus ante quem is the latest possible date something may have happened.

[12] For instance, the Dutch National Species Register, www.nederlandsesoorten.nl/

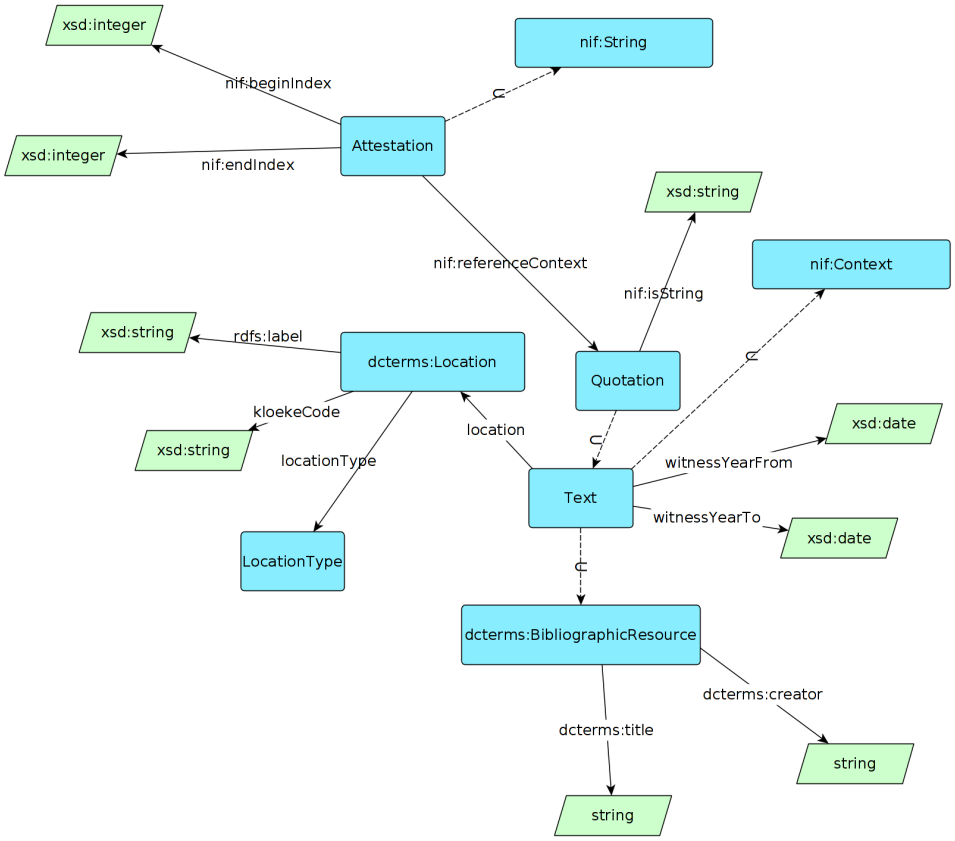



Figure 2: Attestation metadata

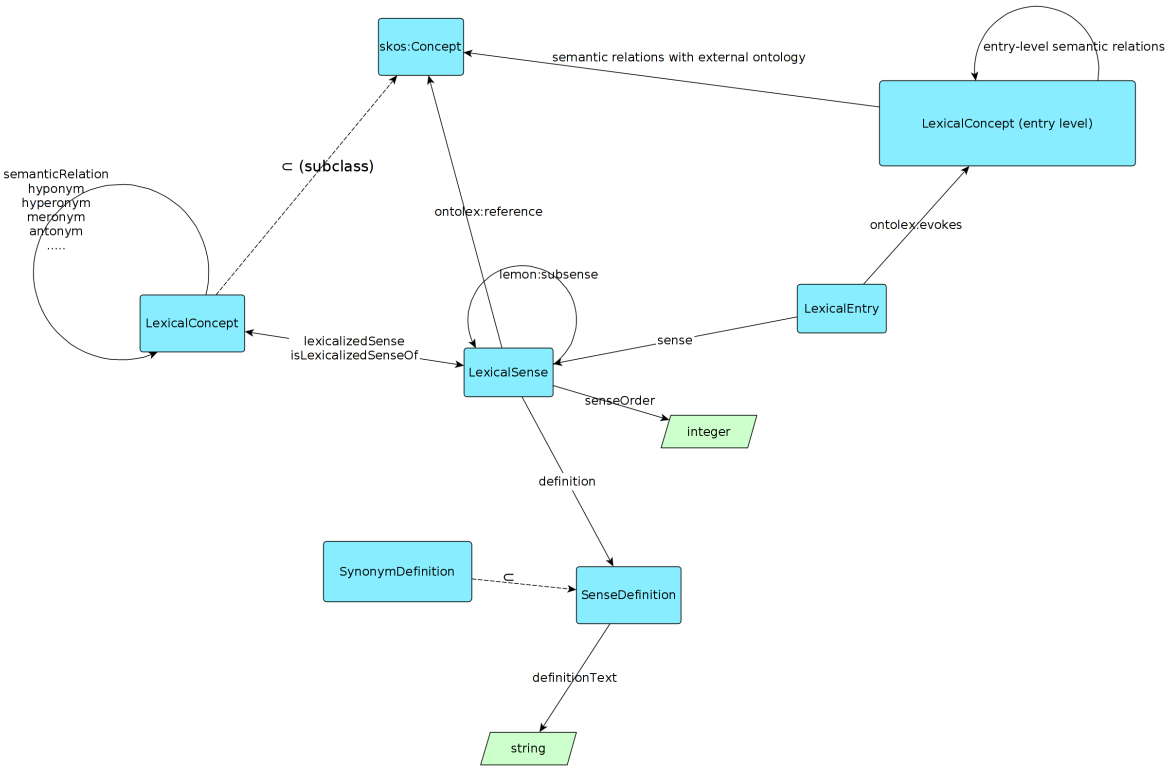



Figure 3: Sense and subsense structure

| Metadata property | description |
|---|---|
| witnessYearFrom | Terminus post quem[10] for the document from which the evidence is obtained |
| witnessYearTo | Terminus ante quem[11] |
| LocationType | Some element in an enumeration containing levels like Country, Province, City, etc |

Table 2: metadata properties

model (near)-equivalences between senses from different resources by links between the associated lexical concepts. The alternative options of modeling the hierarchy by means of RDF collections or containers, or the *senseSibling* property proposed by (Khan et al., 2016, 2017) generate a huge amount of extra triples, especially given the extensive hierarchy (maximum depth of 9 levels, with up to 760 "senses" per article[13]). We chose to re-reify definitions (current Ontolex dropped the *SenseDefinition* class of its predecessor Lemon and proposes skos:definition, which is a data property) in order to be able to attach provenance (and other information) to them. We further propose that the (Lemon) *SenseDefinition* class can be subclassed, according to different types of lexicographical definition. We are in the process of transforming automatically extracted synonym definitions into semantic links. We use the subclass *SynonymDefinition* to represent the synonym references extracted automatically from the dictionary definitions.

### 2.5. Provenance

Scholarly lexicography provides evidence for the assertions made. The user can assess the reliability of the interpretation on the basis of the evidence. When dealing with enriched data, equal standards should be adhered to. The PROV ontology[14] provides us with mechanisms to provide information about the provenance of the added layers of information. For the core lexicographical data, provenance is specified in a more succinct way by referring to the id's of data elements. For those enrichments which have been added automatically and only partially verified manually, it is important to distinguish the verified and the unverified instances. By restricting results to resources associated with agents from the subclass *Person*, a user can exclude the unverified part of the lexicon.

## 3. Conclusion and Future Work

The lexicon model has been tested by converting the dataset to a medium-size resource of about 40M triples and deploying it in a SPARQL endpoint using Jena TDB version 3.1.0[15]. In several realistic usage scenarios, both as a standalone resource and in combination with other resources (DBpedia, Open Dutch Wordnet, distributional thesauri), performance is quite acceptable for non-distributed queries (although the engine used is rather sensitive to the ordering of subqueries). Query formulation is not too cumbersome for users with some knowledge of SPARQL. The main remaining challenges (apart from the development of the lexicon content) are to improve performance on federated queries over several endpoints and to implement a user-friendly query interface for non-technical users.

[13] gtb.inl.nl/iWDB/search?actie=article&wdb=WNT&id=M089102
[14] www.w3.org/TR/prov-o/
[15] jena.apache.org/documentation/tdb/

## 4. Acknowledgements

This research was made possible by the Instituut voor de Nederlandse Taal and the CLARIAH-CORE project financed by NWO (www.clariah.nl).

## 5. Bibliographical References

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

Bosque-Gil, J., Montiel-Posoda, E. and Aguado-de-Cea, G. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: The K Dictionaries Case. *GLOBALEX 2016 Lexicographic Resources for Human Language Technology*, 65–72.

Bosque-Gil, J., Gracia, J. and Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in Ontolex. In *CEUR Workshop Proceedings* (Vol. 1899, pp. 74–84).

Cimiano, P., Mccrae, J., Buitelaar, P. and Montiel-Ponsoda, E. (2013). On the Role of Senses in Ontology-Lexica. In A. Oltramari, P. Vossen, L. Qin, & E. Hovy (Eds.), *New Trends of Research in Ontologies and Lexical Resources. Ideas, Projects, Systems* (pp. 43–62). Springer.

Cimiano, P., McCrae, J. P. and Buitelaar, P. (2016). *Lexicon Model for Ontologies: Community Report, 10 May 2016.* Retrieved from www.w3.org/2016/05/ontolex/.

Depuydt, K. and De Does, J. (2008). United in Diversity: Dutch Historical Dictionaries Online. In E. Bernal & J. De Cesaris (Eds.), *Proceedings of the XIII Euralex International Congress (Barcelona , 15-19 July 2008)* (pp. 1237–1241). Barcelona: Institut Universitari de Lingüística Aplicada Universitat Pompeu Fabra.

Hirst, G. (2009). Ontology and the Lexicon. *Handbook on Ontologies*, 269–292.

Khan, F., Díaz-Vera, J. E. and Monachini, M. (2016). Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web. In I. Draelants, C. Faron Zucker, A. Monnin, & A. Zucker (Eds.), *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage Co-located with 13th Extended Semantic Web Conference (ESWC 2016)* (Vol. 1595, pp. 37–46).

Khan, F., Bellandi, A., Boschetti, F. and Monachini, M. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using Ontolex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon.

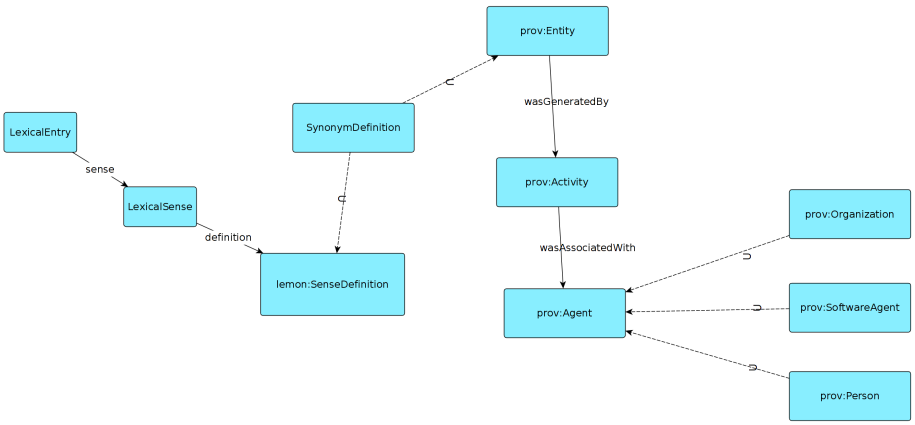



Figure 4: Provenance information attached to the automatically extracted synonym definitions

Kilgarriff, A. (1997). "I don't believe in word senses". In *Computer and the Humanities* 31 (1997), pp. 91-113.

Kilgarriff, A. (2005). Putting the Corpus into the Dictionary. *Proc. MEANING Workshop.* Trento, Italy. Retrieved from https://www.kilgarriff.co.uk/Publications/2005-K-Meaning-PCID.doc.

McCrae, J., Aguado-de-cea, G., Buitelaar, P., Cimiano, P., Declerck, Th., Gómez Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E, Spohr, D. and Wunner, T. (2010). *The Lemon Cookbook*. Retrieved from lemon-model.net/lemon-cookbook.pdf

McCrae, J. P., Bosque-gil, J., Gracia, J. and Buitelaar, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *Elex 2017 proceedings*.

Maks, E., van Erp, M. G. J., Vossen, P. T. J. M., Hoekstra, R. J. and van der Sijs, N. (2016). Integrating Diachronous Conceptual Lexicons through Linked Open Data (pp. 1–2).

Moerdijk, F. (1994). *Handleiding bij het Woordenboek der Nederlandsche Taal (WNT)*. 's-Gravenhage: Sdu Uitgeverij Koninginnegracht.

Mooijaart, M. (2013). A History of Dutch Lexicography. *Trefwoord, Tijdschrift Voor Lexicografie*, 1–34.

Moreau, L., Groth, P., Cheney, J., Lebo, T. and Miles, S. (2014). The Rationale of PROV. *Journal of Web Semantics*, *35*, 235–257.

# Attempts at Visualisation of Etymological Information

**Armin Hoenen**
Goethe University Frankfurt
Juridicum, Senckenberganlage 29,
hoenen@em.uni-frankfurt.de

*Abstract content*

## 1. Introduction

Reconstructing word histories constitutes an important part of lexicographers (especially etymologists) work. We would like to present a rather simple and then a more complex hypothesis for a word history exemplarily, both along with concurrent visualizations. The key question is how to derive useful visual representations of the histories of single words representing the content of articles from etymological lexica.

## 2. Study Object and Related Work

Our main objects of study are single words, the histories of which we would like to trace. It must be said, that in etymological print lexica, visualizations are no mainstream phenomenon. One reason may be that drawing and printing visualizations can be relatively cumbersome (in comparison to text) given the print medium. Furthermore, a textual representation was required in any case. With the advent of the digital and especially effective automatic extraction, conversion and visualization methods, the question of adding value by visualization comes into focus. The only work explicitly focusing on this issue known to the author is Dixit and Karrfelt (2016) who use Etymological WordNet by De Melo (2014) as basis for their visualization. While visualization for etymological relations seems understudied, in recent years with the large scale migration of content from print to digital representation and the emergence of primarily digital resources, data on etymology has been transported into the digital medium. A large lexical resource in this respect is the DWDS, see Klein and Geyken (2010), which comprises under more digitized versions of several large German lexica among which the "Etymologisches Wörterbuch" and the lexicon of the Grimm brothers which contain many etymologically relevant articles. For the simple visualization attempt, we will use data from this resource. Just as articles in the Wikipedia have been produced in a primary written form, such resources are mainly textual in content. Wikipedia quite soon has become the object of intense study in Computer Science and especially the Linked Open Data community has spent a lot of effort to extract information from the Wikipedia in a structured way and derive various knowledge bases from it, the most famous project being the DBPedia from Auer et al. (2007). The same has happened to a much smaller extent for etymological textual data. De Melo (2014) and Sagot (2017) use Wiktionary as their basis for the extraction of etymological patterns, whereas Chiarcos and Sukhareva (2014) and Abromeit et al. (2016) use more specialized data and explicitly etymological dictionaries such as the Turkic Etymological Dictionary. Consequently information can be interpreted as modelled as a graph, where words or morphemes typically form nodes and are connected by relations or typed edges. Typically those types carry labels such as "derived from", "cognate", "variant orthography" or "etymological origin". Can visualization of such graphs help grasp etymological relationships more effectively than when forced to read and evaluate longer textual representations?

### 2.1. Visualization as Added Value

All but sign languages have a very sequential character that is one word has to follow the other, see also Ong (2013). This implies that textual representations are at a loss when it comes to presenting multidimensional relations, a phenomenon remedied only partly by interlinked hypertext. Especially for the representation of etymological word histories, which are often complex involving many languages and alternative hypotheses, a good visualization could become a means of effectively transporting this information, more effectively so than pure text. If this effectiveness is achieved, then visualization can be used to save effort and time and increase comprehension.

## 3. A Simple (?) Case

As a real world example for a simple case, we use the German word "Schatulle" - casket (also small ornate box for valuables). Can we generate visualizations of such etymological relations on a larger scale and relatively easily add them to digital representations? Partly, this is being done. The informational foundation has already been laid, see De Melo (2014). For new data, the way of conduct would be an extraction of such patterns from the text, which as in the DBPedia may be tricky at times and may lead to some loss without further fine-tuning, compare Abromeit et al. (2016). A result could look like Figure 1.

### 3.1. A Complex Case

The history of Japanese SHA-KAI (社会) 'society' free after Yanabu, Akira (柳父 章 (Yanabu, 1991). Initial problem: Such a word does not exist in the early Meiji-era (starting 1868) in Japan, absence of a translation equivalent and missing awareness of any semantically equivalent entity in contemporary Japanese society. Society has 2 main extant senses, see the OED.[1] One with a local implication naming a group of people such as the National Geographic Society, the other relating to the larger context of all individuals of

[1] http://www.oed.com/view/Entry/183776?redirectedFrom=society

e.g. a state. Sketch of the word history, free summarization after (Yanabu, 1991) with additional references:

- the two constitutive ingredients:
  - SHA – originally, the Chinese character 社 (in its modern Japanese pronunciation SHA) was referring to the shrines of earth gods (as opposed to for instance air gods 神 KAMI and villagers ceremonial meetings around them as 社会 SHAKAI, see also Matsumura (2006)
  - KAI – more traditionally, this Chinese character 会 refers to meeting, but also to the fit of something, to some (harmonic) togetherness
- Early translations: friends (NAKAMA), meet (ATSUMARU), government (SEIFU) ... in the Ekaijiten dictionary (1847/48): KAI (Meeting), kessha (group; SHA expresses group of people sharing the same goals)
- Since sense 2 was largely missing for translations of Western works, some new terms were coined: NINGEN KOUSAI, from NINGEN (mankind) and KOUSAI (typically delimited human relationships - master and servant etc.)
- In the broad public meanwhile groups form which discuss Western cultural artifacts (and texts). They call themelves something-SHA. The probably most important SHA is the MeirokuSHA, which issues the journal Meirokuzasshi concerned with new Western phenomena.
- SHAKAI and KAISHA both are attested as generalizations when talking about the phenomenon of those groups. KAISHA: (any) -SHA: head = SHA, KAI = meet SHAKAI: rather the phenomenon of meeting in such SHAs, (attested: SHAKAI ENZETSU), head = KAI
- SHAKAI forms new bonds in the lexicon and becomes some antonym of SEKEN simple folk, ordinary people

At this point, translators presumably start using SHAKAI for society, sense 2, on a large scale and the word enters common vocabulary.
Alternatively, parallel to English, leaving *National Geographic* out, *Society* remains. For Japanese however, the single character SHA – because of a) it is mostly used as affix where a usage as a single word – not as an affix – would be perceived as unusual and because b) its singulary reading (many characters if isolated as a word have to be read in a different way than if in compounds) of YASHIRO which means temple/shrine is dominant – may have not been linguistically fit for this purpose. A loanword SOSAITI is attested, but would not manifest, probably because the contemporary need to coin a new generalized expression for the something-SHAs temporally (and in subject) coincides with the arrival of a new semantic concept, society sense 2, which needed to be named. Although the visualization in Figure 2 is by no means more than a clumsy attempt, for its generation a large variety of very different semantic and temporal relations had to be integrated. A simple graph based visualization may not be the best and most effective visualization to be found for complex word histories.

## 4. Discussion

It is evident that cases cannot be classified binarily into simple and complex very easily. The distinction is technically inspired and should refer or be mapped to a threshold of complexity (thus a binary decision boundary) where for cases which display more complexity, a simple visualization would be too error-prone. Guided by this principle, the threshold should be chosen rather too low than too high. There are many possible ways of potential measurement of this kind of complexity with factors such as article length, the number of matched relations in the article, the number of cross-referenced dictionary entries, the depth or breadth of the concurrent graph and so forth. An empirical study could provide better insights.

## 5. Conclusion

We think that for rather simple word histories, effective visualizations are possible and possibly extractable while for more complex cases many more layers of information and more complex visualization have to be considered so that at the current point in time not only machine readable complex data, but also structured models for their effective visualization are largely absent. In the presentation, I will therefore display attempts at the visualization of simple and complex cases and finally try to extend visualizations to attempts at the visualization of more complex etymological processes (for instance the introduction and discussion of many new terms in Japans Meiji era and the concurrent transformation of the word network) or processes whereby older words get used more and more pejoratively (comp. Dornseiff and Waag (1955)).

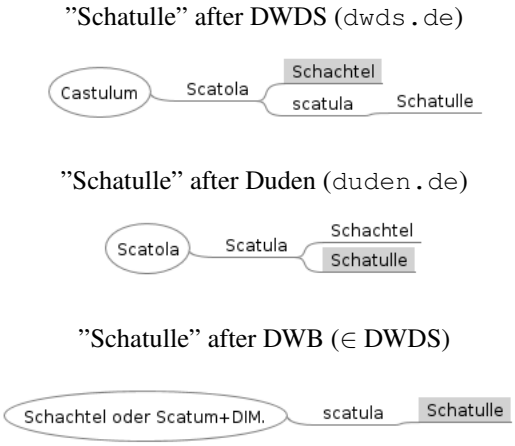



Figure 1: Visualizations of different hypotheses on the etymology of German *Schatulle*. Note, that the third visualization is displaying two alternative hypothesis represented in one node. The hierarchical tree structure, which is also common to many other etymological resources allows in this case the use of the mindmapping software Freemind.

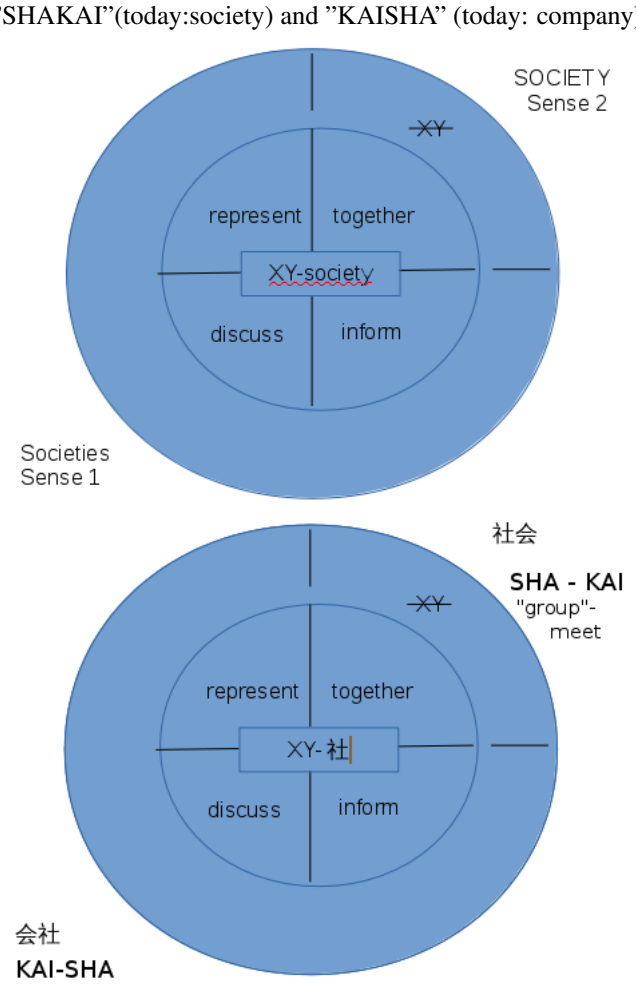



Figure 2: Visualizations attempt of the more complex word history of SHAKAI - society. The outer circle depicts generalization. Above: English, Below: Japanese.

## 6. Bibliographical References

Abromeit, F., Chiarcos, C., Fäth, C., and Ionov, M. (2016). Linking the tower of babel: modelling a massive set of etymological dictionaries as rdf. In *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources, Portoroz, Slovenia*, pages 11–19.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The semantic web*, pages 722–735.

Chiarcos, C. and Sukhareva, M. (2014). Linking Etymological Databases. A Case Study in Germanic. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, page 41.

De Melo, G. (2014). Etymological wordnet: Tracing the history of words. In *LREC*, pages 1148–1154.

Dixit, C. and Karrfelt, F. (2016). Visualizing etymology: A radial graph displaying derivations and origins.

Dornseiff, F. and Waag, A. (1955). *Bezeichnungswandel unseres Wortschatzes: ein Blick in das Seelenleben der Sprechenden*. Schauenburg.

Klein, W. and Geyken, A. (2010). Das digitale wörterbuch der deutschen sprache (dwds). In *Lexicographica: International annual for lexicography*, pages 79–96. De Gruyter.

Matsumura, A. (2006). 2006. *Super Daijirin*, 1.

Ong, W. J. (2013). *Orality and literacy*. Routledge.

Sagot, B. (2017). Extracting an etymological database from wiktionary. In *Electronic Lexicography in the 21st century (eLex 2017)*, pages 716–728.

Yanabu, A. (1991). *Modernisierung der Sprache: eine kulturhistorische Studie über westliche Begriffe im japanischen Wortschatz*. Iudicium-Verlag.

# WordNet for "Easy" Textual Inferences

**Aikaterini-Lida Kalouli, Livy Real and Valeria de Paiva**
University of Konstanz, University of São Paulo, Nuance Communications
aikaterini-lida.kalouli@uni-konstanz.de, livyreal@gmail.com, valeria.depaiva@gmail.com

**Abstract**

This paper presents a WordNet-based automatic approach for calculating "easy" inferences. We build a rule-based system which extracts the pairs of the SICK corpus whose sentences only differ by zero or one word and then identifies which inference relation (i.e. entailment, contradiction, neutrality) exists between these words, based on WordNet relations. Since the sentences of those pairs only differ by the words of the comparison, the inference relation found between the words is taken to apply to the whole sentences of the pair. For some cases not dealt by WordNet we use our own heuristics to label the inference type. With this approach we accomplish three goals: a) we manage to correct the annotations of a part of the SICK corpus and provide the corrected corpus, b) we evaluate the coverage and relation-completeness of WordNet and provide taxonomies of its strengths and weaknesses and c) we observe that "easy" inferences are a suitable evaluation technique for lexical resources and suggest that more such methods are used in the task. The outcome of our work can help improve the SICK corpus and the WordNet resource and it also introduces a new way of dealing with lexical resources evaluation tasks.

**Keywords:** WordNet, natural language inference, SICK corpus, evaluation of lexical resources

## 1. Introduction

"Understanding entailment and contradiction is fundamental to understanding natural language, and inference about entailment and contradiction is a valuable testing ground for the development of semantic representations" say Bowman, Angeli, Potts and Manning in their introduction of SNLI (Bowman et al., 2015), the Stanford Natural Language Inference corpus. We agree and also share their goal of providing semantic representations for sentences which can then be used to compute inference relations between them. To reach this goal we started by investigating SICK by Marelli et al. (2014b), an inference geared corpus that we would like to use as the golden standard for our inference system. This investigation led us to interesting observations on the logic of contradictions, shed light onto faulty corpus annotations and gave us insights for the task at hand, as we discuss in Kalouli et al. (2017b) and Kalouli et al. (2017a). In this previous work we attempted to correct some of those faulty annotations but we soon realized that we could not manually check and correct the whole SICK corpus in a reasonable amount of time. Therefore, we decided to find better ways of correcting sub-parts of the corpus, which led us to this work.

In this work we present our automatic approach for re-annotating and thus correcting a subset of the SICK corpus. The approach is strongly based on Princeton Wordnet (PWN) (Fellbaum, 1998). But the corrected sub-corpus we get as outcome of this work is only one of the three contributions of this paper. Additionally, the approach can be used as a good preliminary basis for identifying "easy" inferences, meaning inferences where syntax is ruled out as a "common denominator" and a sentential inference boils down to a lexical inference and to the one-to-one lexical semantic mappings of the words involved. In other words, what we call "easy" inferences here, are pairs of sentences that can be labelled for entailment/neutral/contradiction relations considering only lexical semantics or world-knowledge. Identifying which inferences are "easy" and how many of them can be achieved with existing lexical tools is important if we want to pursue our goal of computing complex inferences. We believe that complex inferences can be broken down to easy ones and that we need to know how to handle the easy ones first. We also believe that for a symbolic grounded inference system it is important to distinguish different phenomena that play a role in Natural Language Inference (NLI) tasks and then have different ways to deal with them, as pointed out by McCartney (2009). With this approach, we thus seek to evaluate the completeness of PWN as a lexical resource for inference and identify strengths and weaknesses of the lexicon which can be used to improve the resource. A successful evaluation will bring us to our last goal which is to propose that such "easy" inferences tasks are good evaluation methods for lexical resources and that such methods should be used more often as one type of evaluation of appropriate lexical resources. Evaluating lexical resources from a qualitative point of view, more than simply in terms of coverage numbers, is a well known and still open issue for the Lexical Resources community, as pointed out, e.g., by de Paiva et al. (2016).

In the following section we will briefly introduce the SICK corpus. In section 3. we will describe in detail the approach we developed and how it helps us to automatically correct a part of the corpus. In the section after we will evaluate our approach by offering the results of our manual investigation and providing a taxonomy of "easy" inferences found in SICK. In section 5. we will discuss in detail the threefold contribution of this approach and how it can be used further. In the last section we will offer some conclusions and plans for future work.

## 2. The SICK corpus

SICK (Sentences Involving Compositional Knowledge) by Marelli et al. (2014b) is an English corpus, created to provide a benchmark for compositional extensions of Distributional Semantic Models (DSMs). The data set consists of

English sentence pairs, generated from existing sets of captions of pictures. The authors of SICK selected a subset of the caption sources and applied a 3-step generation process to obtain their pairs. This data was then sent to Amazon Turkers who annotated them for semantic similarity and for inference relations, i.e. for entailment, contradiction and neutral stances. Since SICK was created from captions of pictures, it contains literal, non-abstract, common-sense concepts and is thus considered a simple corpus for inference. The corpus is *simplified* in aspects of language processing not fundamentally related to composionality: there are no named entities, the tenses have been simplified to the progressive only, there are few modifiers, etc. The curators of the corpus also made an effort to reduce the amount of encyclopedic world-knowledge needed to interpret the sentences.

The data set consists of 9840 sentence pairs, which have been annotated as 1424 pairs of contradictions ($AcBBcA$), 1300 pairs of double entailment ($AeBBeA$), 1513 pairs of entailment ($AeBBnA$) and 4992 pairs of neutrals ($AnBBnA$). The SICK corpus is a good dataset to test approaches to semantic representations and natural language inference, due to its intended, human-curated simplicity; the pairs talk about everyday, concrete actions and actors. The fact that the captions were produced by different humans, should provide us with near paraphrases or different ways of describing the same scene. The process of normalization added some of the inferences that the corpus was meant to capture, e.g. negations and modifier dropping inferences were added. The number of sentences pairs of the corpus may seem substantial (almost 10K of pairs), but there is much redundancy in the corpus. In total we have 6076 unique sentences and only around 2000 unique lemmas, which means a few more concepts, as assigned by PWN synsets.

## 3. The "one-word difference" approach

Our approach of automatically annotating and correcting the inference pairs is based on the observation that several SICK pairs only differ by none or one word. Differing by "one word" means that there is either one more word in the one sentence than in the other or that each of the sentences contains a word that is not found in the other one. (We say two sentences differ by "no word" when they differ only in their use of the determiners *the* and *a/an*.) They are thus the perfect ground for dealing with some "easy" inferences, as we would like to call them, because we can ignore the syntax involved and find the relation between the pairs solely based on the relation between the different words. A nice example of a "one-word difference" pair is *Kids in red shirts are playing in the leaves.* vs. *Children in red shirts are playing in the leaves*, where the only difference is *kids/children*. This approach can automatically correct and re-annotate some of the pairs without having to solve all the inference challenges associated with the meanings of the sentences first. The approach will become clearer in the following.

### 3.1. Processing SICK

We parsed the SICK corpus sentences with the Stanford Enhanced Dependencies (Schuster and Manning, 2016), which offer us a strong basis for further processing. Then, the sentences were run through the knowledge-based JIGSAW algorithm (Basile et al., 2007) which disambiguates each (noun, verb, adjective, adverb) word of the sentence by assigning it the sense with the highest probability. Briefly, JIGSAW exploits the WordNet senses and uses a different disambiguation strategy for each part of speech, taking into account the context of each word. It scores each WordNet sense of the word based on its probability to be correct in that context. The sense with the highest score is assigned to the word as the disambiguated sense. Using this PWN-based algorithm corresponds to using PWN as our basic ontology or knowledge graph for the approach implemented. Princeton WordNet is a basic ontology and we expect that many inferences will not be supported. However, it is surprising how much we can get from it, which shows that, for the task at hand, PWN has the coverage we need (a similar sort of phenomenon, where PWN worked better than a more traditional ontology, Cyc, was observed in de Paiva et al. (2007). However, on that setting, much more information was available from the syntax, which was based on the Xerox Language Engine (XLE) and Lexical Functional Grammar (LFG) f-structures.)

### 3.2. Finding the "one-word difference" pairs

Having done this shallow linguistic processing of the sentences, we now focus on the surface form of the sentences and extract the ones that differ by none or only one token (we will call these "words-apart" from now on). Since we started working on the surface level, one should note that e.g. *drum* and *drums* still count as different words at this point. We create a small module which takes as input each pair of SICK and checks if the sentences of the pair contain more than two different words. This works on the basis of the creation of sets of words out of the two sentences and the comparison of the sets. If the sets have more than two different words, then they are discarded; if they are different by none, one or two (one from each sentence) words, then the pair is written in a new file, along with the words by which the pair is different as well as which sentence each of the "words-apart" comes from (e.g. the pair *A= A person in a black jacket is doing tricks on a motorbike. B= A man in a black jacket is doing tricks on a motorbike* would be assigned the pair of words *A:person,B:man*).

Note that we choose to exclude some determiners from this comparison. As discussed in Kalouli et al. (2017a), we need to take the SICK pairs as referring to the same entities and events, no matter if the introducing determiners are definite or indefinite articles, to be able to compute contradictions. Since we assume co-reference no matter the definiteness of the articles in the sentences of the pair, we can also exclude them from the difference comparison so that they do not count as words by which the sentences could be different. Note that this approach does not exclude all determiners from the corpus, but only the determiners *the* and *a*. Other determiners that play a role in SICK relations, as well as quantifiers, are taken into account.

By running this module on all 9840 pairs of SICK, we end up with 2936 pairs being "one-word apart" [1], so almost 30% of the corpus.

### 3.3. Assigning relations to the pairs

The two previous processing steps are necessary for the step of automatically assigning inference relations to the "one-word difference" pairs. We create a second module that takes the "words-apart" of the previously extracted pairs and depending on the nature of those words it either runs some heuristics on them or feeds them to WordNet for further processing.

**Heuristics for non-lexical relations** If at least one of the "words-apart" is not a PWN word, i.e. a noun, a verb, an adjective or an adverb — in other words, if it is one of the word classes not handled by PWN — then the "words-apart" are fed into a heuristic engine that decides which label should be given to the pair.
We need such an engine to account at a very primitive level for the missing syntax and at the same time to not lose the precision of such pairs. Only the following cases are dealt with:

- one of the words is a form of the auxiliary *be* (the only one used in SICK) and the other one is the negated version of that auxiliary: the sentences contradict each other;
- one of the words is the negation particle *not* or *no*: the sentences contradict each other;
- there is only one different word and it is the quantifier *one* which is handled as a determiner and thus "ignored" (see section 3.2.): the sentences entail each other;
- the two words are opposing prepositions, e.g. *on-off*, *up-down*, *with-without*, *in-out*: the sentences contradict each other[2];
- both words are quantifiers or there is only one different word and it is a quantifier: depending on the quantifiers different heuristics apply; e.g. if the word of $A$ is the quantifier *many* and the word of $B$ the quantifier *few*, then the sentences contradict each other but if the word of $A$ is the quantifier *many* and the word of $B$ the quantifier *some*, then sentence $A$ entails sentence $B$ but sentence $B$ is neutral to $A$, etc;
- both words are one of the pronouns *someone, somebody* or one word is one of those and the other one is the word *person*. In both cases the sentences are taken to entail each other.[3]

This means that every pair that enters this engine is finally labelled with one of the inference relations $AeBBeA$, $AeBBnA$, $AcBBcA$, $AnBBnA$, $AnBBeA$ or "-", where $A$ stands for sentence $A$, $B$ for sentence $B$, $e$ for $entails$, $c$ for $contradicts$ and $n$ for $neutral$. We use the symbol "-" for cases the heuristics cannot deal with.

**WordNet for lexical relations** If none of the "words-apart" is one of the above cases, then the words are fed into our PWN-based mechanism. The mechanism retrieves from our local repository of PWN3.0 the synonyms, hypernyms, hyponyms and antonyms that correspond to the disambiguated sense of each word, as this was assigned during the step of processing SICK with JIGSAW (see Section 3.1.). The entries found for each lexical relation (i.e. synonymy, hypernymy, hyponymy, antonymy) of the one word are compared with the entries for each lexical relation of the other word. Depending on the ordering of the sentences within the pair, different monotonicity rules apply (Hoeksema, 1986). For example, if the word $A$ is one of the hyponyms of the word $B$, then there is upward monotonicity that implies that sentence $A$ will entail sentence $B$ but $B$ will be neutral to $A$. Similarly, if the word $A$ is one of the synonyms of the word $B$, then the two sentences entail each other. The mechanism takes into account all possible combinations between the lexical relations of the "words-apart" and gives to each pair one of the inference labels mentioned above. If no relation between the "words-apart" can be established, then the pair is left unlabelled. If one of the "words-apart" cannot be found within PWN altogether, then the pair is marked with the label "not found".
The senses contained in SICK are expected to be daily actions and common entities that a knowledge base like PWN should already have. (By contrast, in a more specialized corpus such as a biomedical one, we would expect to need to add to the standard English vocabulary, the specific biomedical vocabulary required by the application.) We expect, for example, that the lexical resource knows that a *dog* is an *animal*, an easy and obvious taxonomic inference. After comparing some of the words found in SICK as a whole with the ones contained in PWN3.0, we observe that some words or senses are still missing. For instance, PWN has no adjective *shirtless* nor the noun *footbag*, although they are established dictionary words. Concretely, we observed that some 15 nouns are missing from PWN3.0. For the 1100 unique nouns of SICK, lacking only so few shows that PWN has a large coverage of English concepts and can be used for a corpus like SICK. However, we must remember that SICK is simplified on purpose, it aims to not have multiword expressions (MWEs), named entities or compounds. This is an important characteristic of the corpus that provides us with good results in this task. It is well-known that WordNet misses many of the well established MWEs in English, which may mean that, if we want to deal with larger inference corpora, like SNLI, we should extend our

[1] Available under `https://github.com/kkalouli/SICK-processing/tree/master/word_difference/one_word_difference`

[2] Princeton WordNet contains no functional words, e.g. no prepositions nor pronouns, so it cannot deal with meanings that depend on them. Newer work from the ARK Lab in CMU provides meanings for prepositions, so we hope to investigate the use of their resources described in `http://www.cs.cmu.edu/~ark/` soon and perhaps integrate some more of them in this module.

[3] We use this heuristic because, since PWN has no pronouns, *someone, somebody* are not mapped to the concept of *person* as humans would naturally map them.

resources using perhaps Wiktionary and Wikipedia. Even for SICK processing, WordNet lacks some concepts; it does not have *jetski* nor *jet ski* or even *water ski*, for instance. It does not have nouns such as *motocross, wetsuit, corndog* or verb predicates like *rock climb, unstitch, wakeboard* (verb). Other concepts of SICK cannot be found in PWN because of tokenization issues. Wordnet lists *fistfight* instead of *fist fight*, and *ping-pong* instead of *ping pong*, for instance, but SICK uses the tab-separated notation so the concepts do not match. Although it might sound trivial, this inconsistency causes several mappings to fail. Additionally, despite trying to avoid compounds, SICK has 1129 of these, as counted based on the Stanford dependencies. These come down to 435 unique compounds. Out of the 435 unique compound nouns in the processing of SICK, only 84 are included in PWN. Of course, many might not deserve to be listed as compounds in PWN. The criteria to be used for dictionarizing a compound is a thorny subject. For instance, a *toy train* is a perfectly compositional compound that appears in Wikipedia. Lexicographers perhaps have no need to list these compositional compounds, but ontologists (especially the ones interested in massive processing of texts) need to do so.

Our PWN-based mechanism has the merit of precision. No matter if ten or a hundred Turkers say that a *man* and a *person* entail each other, PWN will tell us that men are persons, but there are other persons too. So the sentence *A man in a black jacket is doing tricks on a motorbike* entails the sentence *A person in a black jacket is doing tricks on a motorbike*, but not conversely. Similarly, PWN will also tell us that a *guitar* is a musical *instrument*, but not all instruments are guitars and thus it avoids the issue noted by Beltagy and described in `https://github.com/ibeltagy/rrr` [4], that makes guitars and flutes entail each other. Note that this theoretical precision can be broken if the tools on which our system is based, i.e. the Stanford Parser and the JIGSAW algorithm, deliver faulty output. For instance, a missrecognized part-of-speech will lead to a faulty disambiguation which might lead to the assignment of the wrong PWN label.

But in this paper we wish to examine concretely what is the coverage of WordNet for the "one-word difference" pairs and not for the whole SICK corpus. In the next section we will evaluate our approach and discover strengths and weaknesses of this approach and WordNet.

## 4. Evaluation of the approach

The "one-word difference" approach presented above was applied on all 2936 pairs that are "one-word" apart and it could automatically label 1651 of them. We manually looked at both the labelled and unlabelled pairs to see on the one hand if the labelled pairs have the right annotation and on the other hand which kinds of lexical inference can be accomplished by PWN and which senses or relations are still missing.

### 4.1. Evaluation of WordNet labelled pairs

Our manual investigation of all 1651 pairs showed us that our "one-word difference" approach is reliable and has an almost 100% accuracy as it will be shown shortly. Although not all pairs get a label, the 1651 that do, are assigned mostly the correct inference relation.

We could confirm 1100 contradictions with most of them coming from the non-lexical heuristics we defined and 200 coming from lexical antonyms. We additionally found 221 single-sided entailments which correspond to hypernymy and hyponymy relations, two of the main PWN relations. These are taxonomic subsumptions of the kind: a *dog* is an *animal*, the collection of *pianists* is contained in the collection of *persons* and a *man* is a *person*.

We also have 330 double entailments coming mostly from synonyms known to PWN, e.g *couch* and *sofa*, *clean* and *cleanse* or *carefully* and *cautiously* or from some of our heuristics, e.g. the quantifiers heuristics. There are 199 pairs out of these double entailments which belong to a third category, in which no different word is found within the pair, e.g. *A = The teenage girl is wearing beads that are red. B= A teenage girl is wearing beads that are red*. However, since the very basic processing we are doing only considers the surface forms of the sentences and cannot distinguish between agents and patients, 33 pairs out of the 199 are wrong because the order of the words is changed, causing the predicate arguments to be scrambled and thus the sentences to not entail each other, e.g. *A= A baby is licking a dog. B= A dog is licking a baby*. These 33 pairs (1,9%) out of the 1651 labels cost us the 100% accuracy.

Using the present approach, we could automatically correct pairs such as *A = A woman is combing her hair. B= A woman is arranging her hair* that was labelled as $AnBBnA$ in the original SICK and in our present version in annotated as $AeBBnA$. In this way, we can improve the human annotation.

### 4.2. Evaluation of unlabelled pairs

There were 1285 pairs that could not get a PWN label (cf. Table 1). Surprisingly, only a few of them were due to words missing altogether from PWN; the rest were due to missing relations between the terms. The words *debone, atv (all terrain vehicle), biker and kickboxing* for example are missing from PWN3.0 altogether. A few other failures are due to issues with the disambiguation. For example, for the pair *A = A woman is amalgamating eggs. B= A woman is mixing eggs*, PWN does have the verb *amalgamate* in the same synset as *mix*, but JIGSAW wrongly assigns *amalgamate* to the lemma *amalgam* and wrongly annotates it as an adjective and thus as such cannot find it within PWN.

We have 325 pairs that we annotated as antonyms or near antonyms. Knowing that the corpus was constructed aiming for a reasonable number of contradictions and assuming that sentences refer to the same events and entities, we believe pairs such as *Children in red shirts are playing in the leaves* and *Children in red shirts are sleeping in the leaves* need to be annotated as contradictions, although *sleep* and

[4][...] This is because of inconsistencies in the annotations of the SICK dataset (remember that most of the rules are automatically annotated using the gold standard annotation for the pair where the rule is extracted from). For example, the relation between "flute" and "guitar" could be Entail but in most cases it is Neutral.

*play* are not direct antonyms. The same children cannot be sleeping and playing at the same time. These intended contradictions account for a high number of the illogical annotations we have observed before in Kalouli et al. (2017b). This pair was annotated by Turkers as $AnBBcA$, instead of $AcBBcA$. But such an antonym relation is not present in PWN. It is world knowledge that people, even kids, cannot play and sleep, or sit and jump at the same time. Many of the 325 pairs can be accounted for by such world knowledge. It is an interesting, open question whether some of these relations should be included in PWN and if yes, under which category. Some other *near antonyms* bring us to the well-known difficult issues of deciding on the granularity of events: *A man is resting* is not contradictory with *A man is exercising*, but the same man at the same moment cannot be doing both, even if exercising requires some resting between exercises.

There are 299 pairs that we called 'intersective'. These correspond to a single word difference and this word, usually either an adjective or an adverb, provides an intersective subset of the predicate described. For example, in the pair *A skilled person is riding a bicycle on one wheel. A person is riding a bicycle on one wheel*, we only need to check that a *skilled person* is a *person*. Similarly for the example *Some fish are swimming quickly. Some fish are swimming* we only need to know that *swimming quickly* implies *swimming*. A few of these intersectives are actually compounds, like *swimming pool, cyclone fence, etc*. Such 'intersective' cases are not expected to be handled by PWN as they need a module for inference, even if just a basic one, to deal with them. This example confirms what we pointed out in the introduction: even such "easy" inferences pose challenges and are not as "easy" as one might expect and therefore we need to be able to do these first, if we really want to compute more complex inferences.

Moving on with our investigation, among the unlabelled pairs, we found 283 that belong mostly to the taxonomic relations we described before, i.e. hyponymy/hypernymy and synonymy, and would thus be single-side (259 pairs) or double entailments (24), respectively. On the one hand, this (positively) low number (24) of double-entailments, or synonyms, not labelled by PWN shows interesting weaknesses of PWN. For example, PWN has nine synsets for the verb *fire*, at least four of which (02002410, 01133825, 01135783 and 01134238) have to do with guns and weapons, but the verb *shoot* does not appear anywhere in these four synsets. Similarly, the noun *cord* has four synsets, only one (04108268) relevant to its similarity to *rope*, which also has four noun synsets, only one relevant to *cord* (03106110), but these two synsets are not connected at all. On the other hand, the higher number of single-side entailments left unlabelled can mainly be explained by more complex challenges than plain weaknesses of PWN. For example, *to perform* does not necessarily imply *to play*; one can perform mimes, act on plays, do performance art. But *A band is performing on a stage* does entail that *A band is playing on a stage* and conversely. So, again here, we have relations, that only work in the specific context of the other arguments provided, similarly to what we observed for the antonyms.

It is again worth discussing if and how such relations and information should be encoded in lexical resources such as PWN. For some of them, we are convinced that we will need to use the strengths of machine-learning and word embeddings, which could probably give us some of the intended relations; e.g. in the pair *The dog is catching a black frisbee. The dog is biting a black frisbee*, the words *catch* and *bite* describe pretty different actions but in the context of a dog, the words are to be treated as similar. We have also observed that such harder cases mostly involve verbs as their senses are more controversial than nouns.

The further categories discussed in what follows constitute smaller groups. Firstly, there are 27 pairs whose sentences involve meronymy relations and precisely what specific nouns are made of. A representative example is the sentence *A dog is running on the beach and chasing a ball* pairing to *A dog is running on the sand and chasing a ball*. Since our approach is not considering the meronymy relation of PWN, which would provide us with the information that a beach is made of sand, such cases remain unlabelled.

Secondly, there is a collection of pairs (112) that seem to us a misguided effort on the part of the corpus creators to paraphrase certain complex expressions.

The first case (27 pairs) is the one of removing adjective expressions from the sentences. Transforming the sentence *A man in a black jersey is standing in a gym* into *A man in a jersey which is black is standing in a gym* seems a confusing source of mistakes for annotators and parsers.

The second case (32 pairs) is doing a similar job of rewriting 'noun-noun' compounds, but without creating a relative clause. For example, the sentence *A soccer player is scoring a goal* was expanded to *A player of soccer is scoring a goal* but how often would we say *player of soccer* instead of *soccer player*? These pairs mostly use the prepositions *for, of, from*, as in *fishing rod, roof top, tap water*, respectively: *a rod for fishing, the top of the roof, water from a tap*. Lastly, there are several pairs (53) where the expansion tried to explain a compound, to provide a definition for the term. To make it clearer, we can look at an example. The sentence *The crowd is watching two racing cars that are leaving the starting line* was paired to *The crowd is watching two cars designed for racing that are leaving the starting line*, in which there is an attempt to explain *racing cars* as *cars designed for racing*. But many other, less complex, closer to real-world 'definitions' could have been provided instead.

Clearly some of this information is lexical and could be codified having more Wikipedia-style world knowledge in PWN, like saying *motocross bike* is a kind of motorcycle for racing on dirty roads or a *ceiling fan* is a fan usually attached to the ceiling. Other information is instead the kind of world knowledge that tends to be codified in a knowledge base such as SUMO (Niles and Pease, 2001), like the fact that a *a sewing machine* is a machine used for sewing fabric and could thus not have been labelled by PWN anyway. However, many of the pairs of this category explain colors of concrete nouns, such as *blue shirt, brown duck, black dog* described as *a shirt dyed blue, a duck with brown feathers, a dog with a black coat*, respectively, which should be neither in the lexicon nor in a knowledge base in any case and it is thus not surprising that they are not found by PWN.

| | |
|---|---|
| Near Antonyms | 325 |
| Intersective | 299 |
| Synonyms | 24 |
| Hypernyms/Hyponyms | 259 |
| Meronyms | 27 |
| Paraphrases | 112 |
| Dropping | 34 |
| Scramble | 55 |
| Similar | 36 |
| Others | 114 |
| **Total** | **1285** |

Table 1: Phenomena in non-labelled pairs

Thirdly, we have 34 pairs of dropping modifiers or dropping conjunctions, for instance *A man is playing a piano at a concert. A man is playing a piano* or *The man is singing and playing the guitar. A man is playing a guitar*. Although such pairs can be solved by simple logic, similar to the one presented for the 'intersective' pairs, the knowledge required to do so is not lexical and is thus not encoded in PWN. Again, here we would need a basic inference module to do such "easy" inferences.

Additionally, we have 'scrambled' pairs (55), as described in Section 4.1.. Pairs such as *The woman is drawing a man. A man is drawing* cannot be resolved by lexical knowledge alone but instead would need at least a notion of comparing relationships.

Furthermore, we have cases of lexical similarity that are not really logical. For example, consider the pair *A= A dog is licking a toddler. B= A dog is licking a baby*. Toddlers are not babies, the words are not synonyms, but they are similar enough that people will use them as if they were synonyms. These similarity cases are interesting, as they prompt the question of how this kind of information should be encoded, similar to the discussion about "context-dependent" antonyms and synonyms early on. State-of-the-art machine learning techniques might be able to give us more expanded or more context-specific semantics for certain words which might facilitate this task.

Last but not least, there are cases of unlabelled prepositions, quantifiers and inter alia. As explained earlier, we only have a few prepositions in our heuristics and thus there are 42 pairs, whose differences are prepositional but our approach does not handle. Expansion of the heuristics would decrease this number. There are some 20 pairs that differ by numbers or quantifiers (e.g. *Three women are dancing. A few women are dancing*), for which more than lexical knowledge is required and another 40 pairs that seem to us really neutral and no linguistic knowledge, lexical or otherwise, would help. Representative is the pair *A man is thinking. A man is dancing*. People can dance and think at the same time. We call this entire last category 'Others' in the table following.

Looking at Table 1 it is clear that lexical semantics can only help with some of the phenomena, as it was described in detail above.

## 5. Contributions of the approach

As it was mentioned in the introduction, with the work and approach in this paper we hope to achieve three goals. In the following, we will see how each of these goals is fulfilled. We should note that this approach is simple, yet wide enough to be used on other corpora than SICK and achieve similar goals. Any corpus containing pairs of sentences differing by two or less words can be used as an application platform of this approach. There is nothing SICK-specific in this approach which makes the method ideal for verifying the annotations of further corpora, further evaluating WordNet and further discovering "easy" inferences. We see that similar efforts like the one by Pavlick and Callison-Burch (2016) are also breaking down the task of inference to smaller parts and are concerned with doing such "easy" inferences that are however essential for NLI.

### 5.1. Correcting a sub-corpus of SICK

One of the strengths of our approach is its precision with respect to a given lexicon. If some entailment is in the lexicon, it will be annotated correctly and the evidence of the entailment can be provided. But how close really are annotators' intuitions to the ones of the linguists that built lexical resources like PWN? Do they agree that a *a chef is cooking a meal* implies *a chef is preparing a meal*? Do they think that *typing* is *writing with a keyboard*? It seems that there is much disagreement as we already discussed in Kalouli et al. (2017a) and we could see once more in this work, something very astonishing if we take into account that these are "easy" inferences. Note that out of the 1651 pairs that PWN could label, 336 got a different label by the SICK annotators and accepting PWN as the correct, golden standard for such definitions, we can claim that 20% of this sub-corpus of SICK was wrongly annotated. Such a percentage raises worries, especially considering the fact that these are classified as "easy" inferences. So, the first contribution of our approach is to provide another corrected sub-corpus of SICK as we did before (Kalouli et al., 2017b) but this time with less effort. The corrected sub-corpus is available under `https://github.com/kkalouli/SICK-processing/corrected`.

### 5.2. Evaluating WordNet

Everyone should agree that there is an easy inference from the sentence *A dog is barking at a ball* to the sentence *An animal is barking at a ball*. Similarly no one would disagree with the assertion that *The baby elephant is not eating a small tree* contradicts the statement *The baby elephant is eating a small tree*. These are the kinds of trivial, non-controversial inferences that SICK is expected to account for because its construction process was conceived exactly to add these kinds of inferences to sentences extracted from captions. But do our lexical resources support these trivial inferences? To what extent?

We were able to answer such questions by looking at our "one-word difference" approach and investigating which cases could be handled by WordNet and which ones are missing. We have provided the taxonomies in section 4. and these could be taken into account by lexicographers to improve PWN and other such resources. Some of the data

presented above bring up old but interesting questions for further discussion, e.g. what is part of the definition of a noun (cf. example of *sewing machine*) and what is a relation of the word? We believe that the task of inference can and should be broken down to "easy" inferences like these ones and that therefore it is of great importance to have trustworthy, high-coverage resources that can solve big parts of them. Of course, lexical resources will never cover everything but they should always be expanded and then further supported by other state-of-the-art techniques such as word embeddings. We should also note at this point that our approach allows us to relatively evaluate the quality of the other tools used apart from PWN, i.e. the Enhanced Stanford Dependencies and the JIGSAW algorithm (cf. *amalgamating* example). It allows us to identify cases where PWN could not give a label not because of its own weakness but because the wrong sense was given to a word and this sense was not somehow related to the sense of the other word or a wrong part-of-speech was assigned which of course led to the wrong sense and thus again to no match.

### 5.3. "Easy" inferences as an evaluation standard

Evaluating lexical resources is a time-consuming task, mainly because we need to find appropriate test data which should on the one hand efficiently test the coverage of the resources themselves and on the other hand originate from real NLP scenarios that bring to light the whole complexity of language and thus the challenging cases. Our simple yet effective approach shows that tasks of "easy" inferences where the inference relation boils down to lexical relations that such lexical tools should account for, are a good method for testing and evaluating lexical resources. On the one hand, they offer concrete testing of the coverage because they can point out not only whether something is missing altogether but also if a word is missing some inter-relations essential for any NLP task (e.g. several adverbs *amusedly, amazedly, athletically* have only the adjective counterparts in PWN.). On the other hand, "easy" inferences that are extracted from corpora like SICK offer a reasonable and real-world testing scenario because it is exactly these corpora that are used for development or training of further NLP applications and it is thus important to test that the coverage for these corpora is there. Taking this suggestion a step further, we think that there should be an organized attempt to collect such real testing data from corpora and other similar language resources. No matter if these resources are inference-geared or not, it is important that "easy" inferences can be extracted from them so that testing data can be created. This means that it is important to be able to extract "one-word difference" sentences that can be used as testsuite data for the lexical tools. The more the lexical resources improve and expand with this method, the more complex the inference test cases should become so that we can reach a point where lexical resources are almost-complete, mature tools to deal with the first heavy lifting of reasoning.

Finally, we should remark that the approach here is very different from the one taken in the SemEval 2014 competition (Marelli et al., 2014a), where SICK was used as a testing corpus. Out of more than 20 participating teams in SemEval 2014, the top four performing systems are systems that build statistical classifiers on shallow features such as word alignments, syntactic structures and distributional similarities. Thus, our approach is incomparable to these, as we build a rule-based system that does not employ a statistical classifier at all and we only deal with one third of the corpus. The comparison with logic or hybrid logic-statistical systems is also hampered by the use of different grammatical and logical formalisms. We can suggest, following (Martínez-Gómez et al., 2017), that while we eventually envisage a system of Natural Logic or First-Order Logic, for the time being we only use the logic of PWN relations, which correspond to synonymy and subsumption between synsets, as well as simple heuristics.

## 6. Conclusions

We presented our PWN-based automatic approach for doing what we called "easy" inferences. With this approach we attained three goals: a) we could provide a corrected sub-corpus of SICK, b) we could evaluate facets of PWN and provide taxonomies of "easy" inferences and of PWN strengths and weaknesses and c) we observed that our approach is a suitable evaluation standard for lexical resources like PWN. We hope that this work can positively contribute to the improvement of WordNet which we would like to use further for our system of computing inference. We also hope that the concrete commenting and classifying of the PWN-labelled resource we provide publicly can raise interesting discussion points in the community. Last but not least, we believe that the suggestions coming from this work can be integrated in the general discussion about evaluating lexicographic resources and can help in future tasks.

Continuing this work, we would like to expand and pursue further all our three goals. We would like to come up with additional ways of automatically correcting the SICK corpus or, at least, parts of it. Furthermore, we intend to try our method on other suitable inference corpora like SNLI in order to see if we can provide further PWN evaluation and additional "easy" inferences as test data for the evaluation of lexical resources. Finally, we would like to compare the inference relations and the taxonomies we rediscovered from WordNet and the ones suggested by the annotations, to other inference relations obtained by researchers interested in precision focused inference over SICK such as (Beltagy et al., 2015).

## 7. Bibliographical References

Basile, P., de Gemmis, M., Gentile, A. L., Lops, P., and Semeraro, G. (2007). UNIBA: JIGSAW algorithm for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 398–401, Prague, Czech Republic, June. Association for Computational Linguistics.

Beltagy, I., Roller, S., Cheng, P., Erk, K., and Mooney, R. J. (2015). Representing meaning with a combination of logical form and vectors. *arXiv preprint arXiv:1505.06816 [cs.CL]*.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural

language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

de Paiva, V., Bobrow, D. G., Condoravdi, C., Crouch, D., Nairn, R., and Zaenen, A. (2007). Textual inference logic: Take two. In *Proceedings of the International Workshop on Contexts and Ontologies: Representation and Reasoning (C&O:RR) Collocated with the 6th International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-2007), Roskilde, Denmark*.

de Paiva, V., Real, L., Oliveira, H. G., Rademaker, A., Freitas, C., and Simões, A. (2016). An overview of Portuguese WordNets. In *Global Wordnet Conference 2016*, Bucharest, Romenia, January.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Hoeksema, J. (1986). *Monotonicity phenomena in natural language*. Linguistic Analysis.

Kalouli, A.-L., Real, L., and de Paiva, V. (2017a). Correcting contradictions. In *Proceedings of Computing Natural Language Inference (CONLI) Workshop, 19 September 2017*.

Kalouli, A.-L., Real, L., and de Paiva, V. (2017b). Textual inference: getting logic from humans. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014a). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014b). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*.

Martínez-Gómez, P., Mineshima, K., Miyao, Y., and Bekki, D. (2017). On-demand injection of lexical knowledge for recognising textual entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 710–720.

McCartney, B. (2009). *Natural Language Inference*. Ph.D. thesis, Stanford University, Stanford, CA, USA. AAI3364139.

Niles, I. and Pease, A. (2001). Toward a Standard Upper Ontology. In Chris Welty et al., editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9.

Pavlick, E. and Callison-Burch, C. (2016). Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany, August. Association for Computational Linguistics.

Schuster, S. and Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

# A Methodology for Locating Translations of Specialized Collocations

Marie-Claude L'Homme, Nathalie Prévil and Benoît Robichaud
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec) H3C 3J7 CANADA
{mc.lhomme,nathalie.previl,benoit.robichaud}@umontreal.ca

**Abstract**: This paper presents a method for locating translations of specialized collocations for the purpose of balancing lists of collocations in specialized resources. The main steps of the method are: 1. Identifying collocations in a source language for which translations are missing in a target language using an encoding based on lexical functions (Mek'čuk 1996); 2. Locating possible translations of the collocates in the source language in a bilingual resource; 3. Validating equivalents of the target language equivalents in a specialized corpus. In this paper, we focus more specifically on English and French collocations in the domain of the environment. We tested the method manually using 26 English terms and the collocations in which these terms appear and sought to locate translations of these collocations in French. Results show that this strategy for finding translations of collocations is promising and can help terminologists locate and validate collocates in a given language more quickly. With some adaptations, the method could be automated, but human validation is required, especially during step 3.

**Keywords:** terminology, collocations, terminological resource, environment, translation

## 1. Introduction

It is now recognized that adding collocations to terminological resources is extremely useful for certain types of users (translators, technical writers, or any user wishing to know how to insert a term adequately in a specialized text or finding out more about specialized usage). However, there are still few terminological resources that contain large sets of collocations. Some printed dictionaries are available for specific fields of knowledge: stock exchange (Cohen 1986) and business (Binon et al. 2000). A few electronic resources are also available. The Canadian term bank Termium (2018) includes collocates in some term records. IATE contains different kinds of "phrases and formulaic expressions" (Fontenelle 2014: 35). EcoLexicon (2018) lists verbal collocates in some of its entries and classifies them semantically. A resource called ARTES encodes collocations linked to scientific language (Pecman 2012). In our own resources – the DiCoEnviro (2018) and the DiCoInfo (2018) – collocations are listed along with other paradigmatic lexical relations (synonyms, antonyms, morphologically related terms, etc.) in English, French and Spanish (a few Portuguese terms are also included in the DiCoEnviro). Collocations are encoded and the meaning of collocates explained using the system of lexical functions (Mel'čuk 1996).

Collecting collocations from corpora and encoding them in specialized resources is time consuming and this might partly explain why few specialized resources list them. Methods were developed over the years (e.g. Kilgarriff and Tugwell 2001; Kilgarriff et al. 2012) to identify relevant word combinations automatically in running text, but combinations extracted must still be validated by lexicographers or terminologists.

This paper investigates a method for finding translations of specialized collocations and help terminologists locate valid collocations more quickly. Furthermore, the method is designed to balance lists of collocations between languages in multilingual resources. Often (and it is the case with the resource that we are currently compiling, i.e. the DiCoEnviro), entries are written in each language separately. Hence, collocations can be listed in a first language but their translation might not be available in another language. This is unfortunate since tools such as lexical functions can be used to access and retrieve equivalent collocations in different languages.

In this paper, after a brief overview of how collocations are described in our resources (Section 2) we present our method (Section 3) along with an experiment to test its usefulness in the context that we just described. We first tested the method manually in order to verify its potential for automation (Section 4). We examined 26 English terms in the field of the environment. Our test took into account 82 collocations for the 26 English terms. The identification and validation of equivalent collocations was carried out for French. Results are commented in detail in Section 5.

## 2. Collocations in the DiCoEnviro

As was mentioned above, collocations are listed in our resources and encoded using the system of lexical functions, LFs (Mel'čuk 1996). LFs take into account the syntactic structure of the collocation, its general and abstract meaning and, finally, the relation between the collocation and the argument structure of the keyword. For instance, assuming that the term *habitat* has the following argument structure: a habitat: ~ used by X, the collocation *occupy a habitat* would be encoded as follows:[1]

$$\text{Real}_1(habitat) = occupy$$

[1] $\text{Real}_i$ represents collocates that denote the typical activity associated with the key word. In addition, $\text{Real}_i$ is used when the key word is first complement (other LFs denoting typical activities are used when the keyword has another syntactic functions). Finally, the subscript "1" refers to the argument of *habitat* since it realizes the subject of *occupy*.

From the point of view of encoding, LFs have several advantages. First, they allow us to take into account different properties of collocations (syntactic, semantic and argument structure) and thus classify collocations accordingly. Furthermore, they are language-independent. Hence collocations in different languages that have the same meaning are encoded with the same LF.

$Real_1$(*habitat*) = *occupy*, *inhabit*
$Real_1$(*habitat*) = *peupler*

This kind of encoding can be used to establish equivalent relations between collocations in different languages without having to translate them one by one. The DiCoEnviro (and the DiCoInfo, for that matter) allows users to retrieve translations of collocations when they are available in the resource (L'Homme et al. 2012).

However, LFs can be quite difficult to decipher for users who are not familiar with the system. Therefore, different proposals were made to make them more transparent. In the online interface of our resources, LFs are explained with paraphrases that are superimposed on LFs and are designed to translate them in natural language. Our paraphrases are adapted from the proposal made by Mel'čuk and Polguère (2007). Hence, although collocations are encoded by terminologists with LFs, users only view the associated paraphrases in the online textual version (Table 1).[2]

| **Collocation** | **LF** | **Explanation** |
|---|---|---|
| *occupy a habitat* | $Real_1$ | The species uses a h. |
| *inhabit a habitat* | $Real_1$ | The species uses a h. |
| *peupler un habitat* | $Real_1$ | L'espèce utilise un h. |
| *the habitat disappears* | $FinFunc_0$ | The h. ceases to exist |
| *disappearance of a habitat* | $S_0FinFunc_0$ | Noun for "The h. ceases to exist" |
| *loss of a habitat* | $S_0FinFunc_0$ | Noun for "The h. ceases to exist" |
| *rétablir un habitat* | $Caus_{@}De\_nouveauFunc_0$ | Qqn ou qqch. remet un h. dans son état antérieur |

Table 1: Encoding of collocations in the DiCoEnviro

## 3. The problem: imbalance between lists of collocations in different languages

When compiling a terminological resource, the different steps of the methodology are often carried out separately in different languages: specialized corpora are compiled for each language; terms are extracted and identified in each language; each corpus will be searched to retrieve relevant information for terms in that language, and so on.

This clear separation of the workflow in different languages is necessary to ensure that the information collected truly reflects usage in each language and not translation strategies. Furthermore, it prevents resorting to parallel corpora and thus translated texts for one of the languages.

It does, however, have a drawback. Indeed, corpora built may differ from one language to another. Hence, the content of these corpora might not be completely comparable leading to the addition of different kinds of information in a term record. This does not mean that the information given on terms is contradictory. However, the data recorded might not completely overlap when comparing entries in different languages. This problem can be observed in the lists of collocations compiled in the English and French versions of the DiCoEnviro (2018) as shown in Table 2 for the term pair *habitat* (En) and *habitat* (Fr).

| **habitat.1.en** | **habitat.1.fr** |
|---|---|
| *conserve* $_{1a}$ ~<br>*preserve* $_{1a}$ ~<br>*protect* $_{1a}$ ~ | *conserver* $_1$ *un* ~<br>*protéger* $_1$ *un* ~ |
| | *restaurer* $_1$ *un* ~<br>*rétablir* $_{1b}$ *un* ~ |
| *alter* $_{1a}$ ~<br>*degrade* $_{1a}$ ~ | *dégrader* $_{1b}$ *l'*~ |
| | *améliorer l'*~ |
| | *modifier l'*~ |
| *the* ~ *disappears* $_1$ | |
| *introduce* $_1$ *... into a* ~ | |
| | *détruire l'*~ |
| *inhabit* $_{1a}$ ...<br>*occupy* $_{1a}$ ... | *peupler* $_1$ *un* ~ |
| *conversation* $_1$ *of a* ~<br>*protection* $_1$ *of a* ~ | *conservation* $_1$ *d'un* ~<br>*protection d'un* ~ |
| ~ *regeneration* | *restauration* $_1$ *d'un* ~<br>*rétablissement* $_1$ *d'un* ~ |
| | *appauvrissement de l'*~<br>*dégradation de l'*~ |
| | *amélioration de l'*~ |
| | *modification de l'*~ |
| *degradation* $_1$ *of a* ~<br>*deterioration of a* ~ | |
| *disappearance* $_1$ *of a* ~<br>*loss* $_1$ *of* ~ | |
| *recession of a* ~ | |
| | *expansion de l'*~<br>*extension de l'*~ |
| *change in a* ~ | |
| *destruction of a* ~ | *destruction de l'*~ |

Table 2: Collocations recorded for the English term *habitat* and its French equivalent *habitat*

Besides the contents of the corpora, there might be other reasons for this imbalance. For instance, some lexical items might display a higher level of polysemy in one language than in another, leading to difficulties in locating relevant collocates for a specific term. The experience of terminologists might not be the same either and some of them might not spot relevant collocates as easily as others. All in all, we calculated the following discrepancies between English and French collocations in the DiCoEnviro (Table 3). We can see that most collocations do not have an equivalent one in the other language: between 66% and 77% depending on the language considered.

[2] Recently a new representation was added to the DiCoEnviro so users can visualize all lexical relations (including collocations) in the form of a graph (L'Homme et al. 2018). The graph shows both the explanation and the original lexical function.

| | English collocations | | French collocations | |
|---|---|---|---|---|
| With equivalent collocations | 302 | 34% | 315 | 23% |
| Without equivalent collocations | 596 | 66% | 1052 | 77% |
| Total | 898 | | 1367 | |

Table 3: Current imbalance between English and French collocations in the DiCoEnviro

## 4. Methodology

To identify and validate missing equivalent collocations in a target language we defined a method that comprises the following steps (we will illustrate them using examples taken from Table 2):

1. Locate a term in language A for which collocations are listed.

   e.g. *habitat* in English

2. Locate the equivalent in language B for this term in language A. Equivalents are stated explicitly in term records.

   e.g. *habitat* in English → *habitat* in French

3. Retrieve collocations of the term in language A.

   e.g. *habitat* in English:
   *occupy a ~,*
   *introduce … in a ~,*
   *conserve a ~*
   etc.

4. For each collocation, retrieve the lexical function used to describe it.

   e.g. *habitat* in English:
   *occupy a ~* ($Real_1$),
   *introduce … in a ~* (Labreal@$_1$),
   *conserve a ~* (Caus@ContPredVer)
   …

5. For each collocation in language A, locate a collocation in language B that has the same lexical function. This step leads to two different situations:

   Situation 1: An equivalent collocation is listed in language B.

   e.g. *occupy a ~* ($Real_1$) → *peupler un ~* ($Real_1$)

   Situation 2: No equivalent collocation is found in language B.

   e.g. *introduce … in a ~* (Labreal@$_1$) → ?

The remainder of the method applies to Situation 2.

6. For each collocation in language A, take the collocate and search for its equivalents in an online bilingual dictionary.

   e.g. *introduce*

7. Retrieve the equivalents of this collocate from the bilingual dictionary.

   e.g. *introduce* → *introduire, initier, présenter, faire connaître*

8. Search each equivalent in language B and the equivalent of the keyword in language B in a specialized corpus.

   e.g. *introduce* → *introduire + habitat*
   *initier + habitat*
   *présenter + habitat*
   *faire connaître + habitat*

9. When a threshold number of contexts contain a term and a translation of the collocate, this can be considered a candidate translation of the collocation in language A.

   e.g. *introduce … in a habitat* → *introduire + habitat*

10. Encode the Language B equivalent collocate in the entry using the same LF as in English.

    e.g. *introduire + habitat*:
    Labreal$_{@1}$(habitat) = *introduire*

## 5. Manual validation of the method

We tested our method on a sample of terms and carried out part of the steps manually to assess its potential automation.

### 5.1 List of terms

We selected our keywords from a list of general English environmental terms collected for another experiment that consisted in identifying general environmental terms as opposed to terms that are linked to a specific subfield of the domain (Drouin et al. 2018). Among these 126 terms, 56 had French equivalents and 26 had recorded English collocations without French equivalents. The resulting list contains 26 terms[3] shown in Table 4.

| | |
|---|---|
| animal.1.en → animal.1.fr | land.2.en → terre.4.fr |
| bird.1.en → oiseau.1.fr | oil.1.en → pétrole.1.fr |
| carbon.1.en → carbone.1.fr | plant.1.en → plante.1.fr |
| climate.1.en → climat.1.fr | polulation.2.en → population.2.fr |
| ecosystem.1.en → écosystème.1.fr | sea.1.en → mer.1.fr |
| effect.1.en → incidence.1.fr | species.1.en → espèce.1.fr |
| fish.1.en → poisson.1.fr | stratosphere.1.en → stratosphère.1.fr |
| forest.1.en → forêt.1.fr | temperature.1.en → température.1.fr |
| fuel.1.en → carburant.1.fr | threat.1.en → menace.1.fr |
| habitat.1.en → habitat.1.fr | tree.1.en → arbre.1.fr |
| impact.1.en → impact.1.fr | vehicle.1.en → véhicule.1.fr |
| land.1.en → terre.2.fr | waste.1.en → déchets.1.fr |
| ocean.1.en → océan.1.fr | water.1.en → eau.1.fr |

Table 4: Term sample used for the manual validation

### 5.2 Extraction of collocations in English and French

For each term, we extracted all the lexical relations that were encoded as collocations from the English version of the database along with their lexical function. We

[3] Note that some lexical items are polysemous. We extracted them and their associated collocations separately.

proceeded to identify equivalent collocations in French based on the lexical functions. We obtained a table similar to that presented in Table 2 for each term.

We thus obtained for the 26 English terms:

- 180 English collocations;
- 98 English collocations with one or more French translation;
- 82 collocations without a French translation.

### 5.3 Searching for translations of collocates

We selected all 82 English collocations that did not have an equivalent in French. We extracted the collocates and searched for French translations in a bilingual resource. In this experiment, the translations were those produced by Google Translate.[4]

Equivalents labeled in Google Translate as "frequent" and "less frequent" were extracted (rare equivalents were not retrieved). Hence we obtained from 0 to 8 French equivalents for each English collocate (for a total of 211 equivalents). Examples are given in Table 5. The only English collocate that did not produce an equivalent was the verb *to power* (indeed, the only two French equivalents suggested for the verb by the bilingual resource were labeled as "rare").

| **Lexical function** | **Collocation in English** | **Translations of collocate in French according to Google Translate** |
|---|---|---|
| **Habitat.en.1 → habitat.fr.1** | | |
| FinFunc$_0$ | *~ disappears* | *disparaître* |
| Labreal$_{@1}$ | *introduce … into a ~* | *introduire, déposer, presenter* |
| S$_0$Degrad | *degradation of a ~* | *Degradation* |
| S$_0$Degrad | *deterioration of a ~* | *détérioration, degradation* |
| **vehicle.en.1 → véhicule.fr.1** | | |
| Fact$_2$ | *the ~ runs on …* | *fonctionner, passer, gérer, diriger, courir, tourner, marcher, faire fonctionner* |

Table 5: Some equivalents suggested by Google Translate for collocates of *habitat* and *vehicle*

### 5.4 Validating translations of collocates

For each translation produced by the bilingual resource, we searched for contexts in a specialized corpus on the environment that contained both the French key words and the translations of the collocates.

The corpus was a large extract of the PANACEA corpus, an automatically compiled corpus that has a French component containing environmental texts (Prokopidis et al. 2012). The corpus is a compilation of web pages dealing with different topics related to the environment and covers various genres, i.e. official (governmental) reports, popularization, blogs, etc. (according to Bernier-Colborne 2014).

We searched for occurrence of both keywords and collocates using an in-house concordancer called *Intercorpus* (2018). The extract we used (approx. 231 Mb) represented about half the original corpus and was deemed sufficient to obtain representative results.

Contexts were searched using truncation for key words and collocates and a distance of 5 words or less was allowed between the two character strings. Contexts were considered relevant only if there was an actual link between the key word and the candidate collocate. For instance, the following context was considered relevant for *animal* and *vivre* (as a possible translation for *animal lives in …*):

> *C'est aussi parce que ces **animaux** **vivent** dans les forêts tropicales qu'il est important d'agir rapidement* (PANACEA/18159.txt)

However, the following two contexts were not considered:

> *Pour les plantes, il s'agit des conditions de sol et de microclimat propres à la station où elles **vivent**. Grâce à leur mobilité, les **animaux** peuvent utiliser divers types d'abris présents dans leur domaine vital.* (PANACEA/2051.tx)

> *L'ectofaune épizoaire, qui **vit** à la surface d'un **animal**, est une autre forme d'épifaune.* (PANACEA/ 41.txt)

## 6. Results

The number of valid contexts found in the reference corpus was recorded for each potential collocate as shown in Table 6.

| **habitat.en.1 → habitat.fr.1** | | |
|---|---|---|
| FinFunc$_0$ | *~ disappears* | disparaître (28)<br>--<br>-- |
| Labreal$_{@1}$ | *introduce … into a ~* | --<br>--<br>introduire (6), deposer (0), présenter (0) |
| S$_0$Degrad | *degradation of a ~* | dégradation (203)<br>--<br>-- |
| S$_0$IncepPred [MAN:différent] | *change in a ~* | modification (39)<br>évolution (11)<br>changement (8), variation (0) |
| **vehicle.en.1 → véhicule.fr.1** | | |
| Fact$_2$ | *the ~ runs on …* | fonctionner (35)<br>--<br>passer (0), gérer (0), diriger (0), courir (0). tourner (0), marcher (0), faire fonctionner (0) |

Table 6: Frequency of equivalents in the corpus (PANACEA)

[4] We first searched for equivalents in BabelNet (2018). However, for a small set of collocates no translation was available for French.

Three frequency categories were established: A) 20 and over occurrences; B) between 10 and 19 occurrences; C) up to 9 occurrences. It was assumed that valid collocates should appear with 20 occurrences and over in our reference corpus. It was also assumed that the last category would contain invalid translations. Results obtained in each category are detailed and commented below.

Among the results obtained, 53 French equivalents suggested by Google Translate were found in at least 20 contexts. These French equivalents were suggested for 40 source collocations (for a possibility of 82 that our test sample contained). In nearly all these cases, the translations were valid. This confirms our hypothesis according to which valid collocates would be found in that category.

For some source collocations, multiple valid translations were found although with varying frequencies of occurrence. For *conserve* (in *conserve an ecosystem*), the three French equivalents *protéger* (207), *préserver* (106), and *conserver* (20) were validated in the reference corpus and are all valid translations. The collocation *management of water* led to a slightly different situation. Three French equivalents suggested for *management* in the bilingual resource were found over 20 times in the reference corpus, i.e. *gestion*, *administration*, *direction*. However, the first one would qualify best as a valid equivalent and has by far the highest number of occurrences. Results obtained for *conserve an ecosystem* and *management of water* are reproduced in Table 7.

| **ecosystem.en.1 → écosytème.fr.1** | | |
|---|---|---|
| Caus@ContPredVer | *conserve an ~* | protéger (28), préserver (106), conserver (20) |
| | | -- |
| | | évoluer (0) |
| **water.en.1 → eau.fr.1** | | |
| PermFunc$_0$ | *management of ~* | gestion (369), administration (41), direction (35) |
| | | -- |
| | | management (0) |

Table 7: Frequency of equivalents for the collocations *conserve an ecosystem* and *management of water* in the corpus (PANACEA)

In the 10-19 category of results, 19 equivalents suggested by Google Translate were found in the reference corpus (for 18 source collocations). In some of these cases, a valid French equivalent was already recorded in the 20 and over category. For instance, *pose* (in *pose a threat*) can be translated with *poser* (48). Among the other equivalents suggested by Google Translate, *créer* was also found in the corpus, but with only 10 occurrences.

In other cases, there was no French equivalent with over 20 occurrences in the corpus. However, a less frequent suggestion could be a plausible translation. For example, the only French equivalent proposed by our bilingual resource for *accumulation* (in *accumulation of carbon*) was *accumulation*. It was found only in 11 contexts but still remains a valid translation. Finally, the 10-19 category did contain invalid translations. For *grow* (*the plant grows*), the bilingual resource proposed *devenir* (among other translations). *Devenir* appeared in 11 contexts, but was never a valid translation for *grow* considered from the point of view of the environment. On the other hand, *croître* that appears in the same category is the valid translation. Results obtained for *pose a threat*, *accumulation of carbon* and *plant grows* are reproduced in Table 8.

| **threat.en.1 → menace.fr.1** | | |
|---|---|---|
| Oper$_1$ | *pose a ~* | poser (48) |
| | | créer (10) |
| | | présenter (9) |
| **carbon.en.1 → carbone.fr.1** | | |
| S0IncepPredPlus@ [@:lieu] | *accumulation of ~* | -- |
| | | accumulation (11) |
| | | -- |
| **plant.en.1 → plante.fr.1** | | |
| Fact$_0$ | *plant ~* | produire (31) |
| | | croître (12), devenir (11) |
| | | grandir (0), devenir (0), augmenter (0) |

Table 8: Frequency of equivalents for the collocations *pose a threat*, *accumulation of carbon* and *plant grows* in the corpus (PANACEA)

The final category 0-9 contained 95 cases where no attestations of the equivalents suggested by our bilingual resource were found. In nearly all these cases, equivalents were invalid translations in the context of a collocation and could be discarded immediately. For example, *avilir* was suggested for a translation of the English verb *degrade (for degrade an ecosystem)*, but can certainly not be considered a valid translation in the context of *degrade an ecosystem*. Of course, it was never found in our environmental corpus.

In this category, 34 additional suggestions were made by the bilingual resource but were only found a few times in the reference corpus. Many of these suggestions were invalid translation, thus confirming our assumption about candidates with low frequencies. A few suggestions could correspond to valid translations, but did not occur very frequently (along with the key word) in our reference corpus. This was the case with two of the French equivalents suggested for *disturb* (in *disturb an ecosystem*), namely *déranger* and *troubler*. Results obtained for *degrade and disturb an ecosystem* are reproduced in Table 9.

| **ecosystem.en.1 → écosytème.fr.1** | | |
|---|---|---|
| Caus@Degrad | *degrade an ~* | degrader (26) |
| | | se degrader (11) |
| | | avilir (0) |
| Caus@NonFact$_0$ | *disturb an ~* | perturber (42) |
| | | -- |
| | | déranger (3), troubler (2), inquiéter (0) |

Table 9: Frequency of equivalents for the collocations *degrade an ecosystem* and *disturb an ecosystem* in the corpus (PANACEA)

## 7. Discussion

In addition to the quantitative results commented in the previous subsection, the method yielded some qualitative

results that we did not anticipate when we started this project.

First, the corpus clearly showed that the same English collocate could translate differently in French. For instance, Google Translate suggested four different equivalents for the verb *disturb*: namely *déranger*, *inquiéter*, *perturber* and *troubler*. *Déranger* is preferred when *animal* is the key word (23 occurrences); while *perturber* is preferred with *écosystème* (42 occurrences). This, combined with the fact that some French equivalents were never found in the corpus along with given terms, shows that a validation with a specialized corpus remains necessary and is a strong aspect of the method.

Secondly, the corpus could reveal a clear preference for a given equivalent in the context of a collocation. For instance, *conserve* (in *conserve an ecosystem*) translates into French (according to the corpus) as *protéger un écosystème* (207 occurrences). Other equivalents are possible, but less frequent: *preserver un écosystème* (106 occurrences) and *conserver un écosystème* (20 occurrences).

The two observations above show that even collocations if specialized corpora often have a compositional meaning, usage influences the choice of a collocate and must be taken into consideration.

Thirdly, a human validation of the occurrences found in the corpus is necessary. For instance, some English collocates are highly polysemous and lead to French equivalents that are not synonyms or not even remotely semantically related. Our bilingual resource suggested the following French equivalents for the verb *occupy* (*the animal occupies …*): *habiter*, *occuper*, *prendre*, *remplir*. In this case, only *habiter* and *occuper* would be accurate translations. However, in the corpus, *animal* was found in contexts with *prendre* and *remplir* as shown below:

> *Si l'animal prend la fuite à quatre reprises, il est en danger de mort.* (PANACEA/381.txt)

> […] *car il s'agit souvent de plantes et animaux non-autochtones, ne pouvant pas remplir les fonctions qu'ils rempliraient dans la nature, ni ne pouvant remplacer les écosystèmes locaux détruits ou dégradés par les activités humaines.* (PANACEA/2659.txt)

The method also has some limitations. We listed four below:

- In a few cases, accurate translations were unavailable in bilingual resource. For instance, for the English term *warming*, the French equivalents suggested by Google Translate were *chauffage* and *échauffement*. The correct translation is the field of the environment is *réchauffement*. In order to correct this limitation, we could consider using more than one bilingual resource as long as they can be accessed freely.

- Some equivalents were suggested by our bilingual resource, but could not be found in the corpus. For instance, three French equivalents were suggested for the verb *thrive* (*prospérer*, *se developer*, *réussir*). None could be found along with specific terms of our set in the corpus.

- We hypothesized that valid translations of collocates would be found with 20 occurrences and over in the reference corpus. Although we confirmed this hypothesis to a large extent, for many source collocations (half of our sample), no equivalent suggested by the bilingual resource could be found with a sufficient number of attestations. This limitation could perhaps be corrected by using different resources: using a different bilingual resources or more than one bilingual resource, and increasing the size of our reference corpus.

- There was a non-negligible number of cases for which only a few occurrences of both key word and collocate could be found in the specialized corpus. Even if the corpus used (PANACEA) is very large, it covers many different areas of the environment. Perhaps a more focused and specialized corpus would increase the number of occurrences of some collocations. We could also use some of the corpora we compiled manually to increase the number of occurrences of keywords and collocates.

## 8. Conclusion and future work

In our opinion, our method produced a sufficient number of valid translations for our English collocations and could be used with some adaptations to complete other missing translations of collocations. Some suggestions were made above to correct some of its limitations (use of another more focused and specialized corpus, use of other bilingual resources, etc.). Our next step would consist in automating the method step by step. However, it seems that human validation cannot be avoided for this kind of work.

One strength of our method that we did not anticipate when we embarked on this project is that it allowed us to identify some clear preferences for some translations of collocates. It would draw terminologists' attention to phenomena that would be missed otherwise.

There are few directions that we can take in the near future. We could also apply this method the other way around for finding English translations for French collocations. We could also validate its potential for populating versions in other languages for which we have few collocations (our resource also has Spanish and Portuguese components). Looking back on the method, we could also extend it to all collocates in a source language and not exclusively to collocations for which there are no equivalents. This could lead us to find and fill other gaps in descriptions in different languages.

## 9. Acknowledgements

This work was supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada. We would like to thank two anonymous reviewers for their useful suggestions on both the method and the paper.

## 10. Bibliographical References

Bernier-Colborne, G. (2014). Analyse distributionnelle de corpus spécialisés pour l'identification de relations sémantiques, In Actes de SemDis : enjeux pour la sémantique distributionnelle, Marseille, France, pp. 238-251.

Binon, J., S. Verlinde, van Dyck and J., Bertels, A. (2000). *Dictionnaire d'apprentissage du français des affaires*, Paris: Didier.

Cohen, B. (1986). *Lexique de cooccurrents. Bourse – Conjoncture*, Brossard (Québec) : Linguatech.

Drouin, P., L'Homme, M.C. and Robichaud, B. (2018). Lexical Profiling of Environmental Corpora. In Language Resources and Evaluation, LREC 2018, Myazaki, Japon.

Fontenelle, T. (2014). From Lexicography to Terminology: a Cline, not a Dichotomy. In Abel, A., Vettori, C. and Ralli, N. (Eds.). 16th Euralex Conference 2014. Proceedings, Bolzano, Italy, pp. 25-45.

Kilgarriff, A. and Tugwell, D. (2001). WORD SKETCH: Extraction, Combination and Display of Significant Collocations for Lexicography. In Proceedings of the Workshop on Collocations: Computational Extraction, Analysis and Exploitation, ACL-EACL 2001, Toulouse, pp. 32-38.

Kilgarriff, A., Rychly, P., Kovar, V. and Baisa, V. (2012). Finding Multiwords of More Than Two Words. In 15th Euralex Conference, Proceedings, Vatvedt Fjeld, R. and J. Matilde Torjuse (eds.). Oslo, Norway, 693-700.

L'Homme, M.C., Robichaud, B. and Leroyer, P. (2012). Encoding collocations in DiCoInfo: from formal to user-friendly representations. In S. Granger and Paquot, M. (Eds.). Electronic Lexicography. Oxford, Oxford University Press, pp. 211-236.

L'Homme, M.C. Robichaud, B. and Prévil, N. (2018). Browsing the Terminological Structure of a Specialized Domain: A Method Based on Lexical Functions and their Classification. In *Language Resources and Evaluation, LREC 2018*, Myazaki, Japon.

Mel'čuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In Wanner, L. (Ed.) Lexical Functions in Lexicography and Natural Language Processing, Amsterdam / Philadelphia: Benjamins, pp. 37-102.

Mel'čuk, I. and Polguère, A. (2007). *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Bruxelles: De Boeck.

Prokopidis, P, Papavassiliou, V. Toral, A., Poch, M. Frontini, F., Rubino, F. and Thurmair, G. (2012). *Final Report on the Corpus Acquisition & Annotation subsystem and its components* (https://repositori.upf.edu/handle/10230/22514). Accessed 23 February 2018.

Pecman, M. (2012). Étude lexicographique et discursive des collocations en vue de leur intégration dans une base de données terminologiques. *JoSTrans. The Journal of Specialised Translation* 18.

## 11. Language Resource References

*ARTES : Aide à la Rédaction de TExtes Scientifiques*. http://www.eila.univ-parisdiderot.fr/recherche/artes/ Accessed 11 January 2018.

*BabelNet*. http://babelnet.org/ Accessed 5 January 2018.

*DiCoEnviro* http://olst.ling.umontreal.ca/dicoenviro Accessed 11 January 2018.

*DiCoInfo*. http://olst.ling.umontreal.ca/dicoinfo Accessed 11 January 2018.

*EcoLexicon. Terminological knowledge base*. http://ecolexicon.ugr.es/en/index.htm Accessed 12 December 2017.

Google Translate. https://translate.google.com Accessed 28 February 2018.

IATE. *Interactive Terminology for Europe* http://iate.europa.eu/about_IATE.html Accessed 19 February 2018.

Polguère, A. & E. Chièze. *Intercorpus* http://olst.ling.umontreal.ca/intercorpus/ Accessed 11 January 2018.

PANACEA. http://panacea-lr.eu/en/info-for-researchers/ Accessed 11 January 2018.

Termium. 2018. http://www.btb.termiumplus.gc.ca Accessed 11 January 2018.

# Blockchain Lexicography: Prototyping the Collaborative, Participatory Post-dictionary

**Daniel McDonald, Eveline Wandl-Vogt**
Bitpanda GmbH; Austrian Academy of Sciences, Austrian Centre for Digital Humanities,
Austrian Academy of Sciences, Austrian Centre for Digital Humanities
Burggasse 116/2 1070 Vienna
Wohllebengasse 12-14/2 1040 Vienna
daniel.mcdonald@bitpanda.com
tEveline.Wandl-Vogt@oeaw.ac.at

**Abstract**

In this paper, the authors introduce blockchain lexicography, developed and prototyped within the framework of the open innovation exploration space (Research Group Methods and Innovation) at the Austrian Academy of Sciences. Blockchain lexicography exploits emerging technologies (e.g. the blockchain), social developments (do-it-yourself-science, crowd-innovation) and management methods and practices (open innovation), applying them to a case study of lexicography in order to create an accessible, constantly evolving linguistic resource. The authors deliver the design and a prototype of the system, as well as related data. The system, *wugsy*, asks users to provide natural-language texts for images, to score others' texts, or to assess the accuracy of tag clouds, with text types (i.e. stories, descriptions, etc.) tailored to match user profiles. Answers are recorded in a distributed database, which can be hosted, verified or queried by anyone. Responses to games are scored by consensus and rewarded proportionally with a cryptocurrency token. A simple API allows extraction and filtering of database contents.

**Keywords:** blockchain, consensus, trust, democratisation, post-dictionary

## 1. Context

The increasing pervasiveness of digital communication in everyday life facilitates the development and use of novel computational methods that allow better understanding language, society and culture. Lexicography is one area of research that stands to benefit from increasingly digitised life, in terms of *(a)* research presentation; *(b)* use of social media and digital news as corpora, *(c)* interlinking and harmonisation of linguistic data; and *(d)* opening up communicative channels between experts and volunteers (Chesbrough 2006). Digital lexicography, however, has so far rarely made use of the affordances of new media, making it difficult to imagine the future of lexicography; as Hanks explains, it is currently still 'too early, to say, which form innovative dictionaries of the future will take' (2012, p. 82). For this reason, exploration of emerging technologies for the purposes of uncovering new ways of building lexicographical resources is timely.

A parallel computational development is the blockchain (Wood 2014, Pilkington 2015)—a decentralised, trustless ledger that can accurately keep track of digital information. To date, the most common use-case for blockchain technology is as a currency or payment network (e.g. Bitcoin, Ethereum). Recently, however, a number of blockchain research projects have aimed to go beyond cryptocurrency applications, using blockchains as ways of providing proof of existence of documents, as well as tracking migration and medical histories. Blockchain-based systems permit the transfer of real or symbolic value in a way that is very resilient to system outages and malicious code. Meanwhile, blockchain-based databases are provably open-source, limiting researcher bias, increasing reproducibility, and promoting data re-use. For this reason, blockchains have a key potential use case within the open source, open science and open innovation movements, which aim to facilitate access to research tools, data and publications. While cryptocurrency systems have demonstrated the utility of blockchains as both a reward mechanism and store of value, still to be empirically tested is the suitability of blockchain protocols for research data collection.

Related to both the increased presence of digital communication and the rise of decentralised networks is crowdsourcing—the targeted collection of large amounts of data from a pool of online participants. Though some crowdsourcing work in linguistics has been criticised based on the accuracy of generated results, as well as issues of exploitation of labour, ethical crowdsourcing is a major component within the emerging framework of open innovation (Sloane 2007, Chesbrough 2006), due to the fact that crowdsourcing engages the public in science and research, promoting democratisation and the synergy of diverse sources and kinds of knowledge.

Blockchains provide a natural, but thus far underutilised, complement to crowdsourcing tasks. By storing data and rewards in a publicly accessible database that is very difficult to corrupt, it is possible to develop crowdsourcing systems that are provably fair, with results that are inherently publicly accessible. We therefore believe that the combination of blockchain technology and crowdsourcing methods can lead to systems for natural language data generation and collection that surpass current methods in terms of both utility and fairness.

## 2. Aim

In this paper, the authors describe the potential for emerging technologies to be put to use in the context of the post-dictionary phenomenon at the currently founded *exploration space @ ÖAW* (the Austrian Academy of Sci-

ences). They introduce the concept of blockchain lexicography and offer *wugsy*, an initial, open-source prototype of such a system, with the aim of furthering knowledge discovery in the context of linguistic, biological and cultural diversity. The open-source platform gives linguistic tasks to a crowd, and stores the results of these tasks within a blockchain. A separate, but related chain, distributes rewards to participants based on emerging consensus regarding the quality of their answers. Because the data accumulated by the system is free to access, its downstream applications are many. For our purposes, however, we aim to demonstrate that the system can generate insights that are novel and appropriate for inclusion within a dynamically generated post-dictionary.

## 3. Prototype

*wugsy* is human-centred, devised against a background of design thinking (Plattner, Meinel and Weinberg 2009) and agile development. Via a web platform (implemented in Python 3/Django), images are presented to actors, alongside one of a number of possible tasks. The user may variously be asked to:

1. Write a natural language text related to the image
2. Score/rank another user's existing text
3. Select relevant terms that appear within a visualised tag cloud generated through a simple NLP pipeline run over a text
4. Score/rank the accuracy of a tag cloud

The languages and text types requested from users can vary based on current gaps in the dataset and on users' stated language proficiencies, interests and areas of expertise. Tag clouds are generated by parsing texts with `spaCy`, and using POS tags and dependency positions and NER to identify likely tags. Results from these different games (i.e. natural language content, selected tags, rankings of others' stories and tag selections) are then sent to a decentralised database (McConaghy et al. 2016) hosted by those who wish to use the data for downstream tasks. As other actors score the accuracy of stories and selected tags, it becomes possible to determine answer quality by consensus. The degree of consensus for a given question dictates the size of the reward for an individual answer. Actor history can be used to further scale the size of the reward, and incentivise high-quality or high-effort answers (e.g. short composition or brainstorming tasks). Rewards are released to each user's account in the form of an Ethereum-based ERC20 token, which could be given an intrinsic, fluctuating value derived from, e.g., real-world investment in the infrastructure, through fees for API calls to nodes that host the database, or through fees paid in order to add new kinds of data and questions to the crowd. Such a structure incentivises not only participation in games, but also the addition of new data, which expands the explanatory potential of the project, and the hosting of nodes, which play an important role in the overall security and stability of the network.

## 4. Workflow

Taking lexicography as an aim, the workflow for the system is fairly simple. Europeana's historical multimedia collection (Haslhofer and Isaac 2011) is used as an initial image and caption dataset, with users asked to variously generate texts about images, score others' texts, or score the accuracy of tag clouds. These small, compartmentalised tasks are provided by a dynamic visualisations within a web front-end; the combined use of scores, currency rewards and high-quality visualisation of natural language text each gamify the process of data collection, motivating users to produce high-quality content. Participation in games can be anonymous, but participants are rewarded for adding user profiles, because the coupling of profile and answer data makes possible both targeted questioning, and, downstream, more nuanced insights into language use in different dialects, registers and demographics.

An open-source API allows querying the generated data, and dynamically presenting interesting insights online in real-time. The potential use of the API for lexicographic tasks is explored: searching information from the generated tag clouds gives us an insight into relationships between particular words, images and narratives; by restricting search results based on users' overall scores, we can see the differences between high and low-quality submissions, and consider their implications for the design of novel kinds of dictionary. Similarly, we explore how queries containing location filters can be used to uncover regional variations.

## 5. Design Parameters

The codebase is designed with five key design parameters in mind. Namely, the developed system is:

(a) inherently multilingual

(b) responsive to user-specific expertise

(c) self-improving

(d) adaptable to new kinds of language tasks

(e) sensitive to practices of open innovation and open science

Regarding parameters *(a)* and *(b)*, rewards are scaled by the current size of a given language's dataset, with profiles of crowdsourcing participants used to present language problems to participants in line with their stated interests and areas of expertise. Such a design means that languages and content areas with less accumulated information can be prioritised by a relative increase in reward sizes, and by putting more questions from less popular languages and content areas to the user base.

Regarding parameter *(c)*, the authors aim to use the incoming streams of crowdsourced answers continually to train algorithms responsible for selecting problems that are served to the crowd. For example, the algorithms that transform users' texts into tag clouds can be refined based on the kinds of tags that users mark as accurate, or by users' scoring of the tag clouds themselves.

Regarding parameter *(d)*, within the early prototype, lexicography acts as a test-case for a more abstract system that

is equally well-suited to other areas of research. By using different kinds of initial datasets, and by developing new kinds of language games, we expect the system to be able to collect data suitable for use in diverse kinds of research, including linguistic typology (in classifying languages and dialects), computational linguistics (i.e. in natural language generation and parsing), and the social, political and population sciences (in mapping language use to demographic details, or uncovering attitudes toward the data shown to participants).

Regarding parameter *(e)*, the prototype described here not only facilitates novel kinds of research, but, in doing so, also necessarily commits to core values of open science and innovation. *wugsy* guarantees open data and open-source development, connects problems with those best capable of solving them, and thus promotes the creation of knowledge that is provably accessible and diverse. Furthermore, *wugsy* empowers marginalised actors: because the proposed system is multilingual, and because rewards are scaled to incentivise answers for domains in which less data has accumulated, global participants can potentially receive fair compensation for their work.

## 6. Bibliographical References

Chesbrough, Henry W. (2006): Open Innovation. The New Imperative for Creating and Profiting from Technology. Boston.

Fellbaum, Christiane (2014): Large-scale Lexicography in the Digital Age. International Journal of Lexicography, Volume 27, Issue 4, 1 December 2014, Pages 378–395, https://doi.org/10.1093/ijl/ecu018 (accessed: 07.01.2018).

Hanks, Patrick (2012): Lexicography from earliest times to the present. In: Allan, K. (Ed.): The Oxford Handbook of the History of Linguistics. http://www.patrickhanks.com/uploads/5/1/4/9/5149363/2012dlexicography_from_earliest_times.pdf (accessed: 07.01.2018).

Haslhofer, B., and Isaac, A. (2011). *data.europeana.eu*: The europeana linked open data pilot. In International Conference on Dublin Core and Metadata Applications (pp. 94-104).

McConaghy, T., Marques, R., Müller, A., De Jonghe, D., McConaghy, T., McMullen, G., Henderson, R., Bellemare, S. and Granzotto, A., (2016). BigchainDB: a scalable blockchain database. white paper, BigChainDB. https://docs.bigchaindb.com/en/latest/

Pilkington, Marc (2015). Blockchain Technology: Principles and Applications. Research Handbook on Digital Transformations, edited by F. Xavier Olleros and Majlinda Zhegu. Edward Elgar, 2016. https://ssrn.com/abstract=2662660 (accessed: 07.01.2018).

Plattner, Hasso, Meinel, Christoph, and Weinberg Ulrich (2009): Design Thinking. München.

Sloane, Paul (Ed.; 2011): A Guide to Open Innovation and Crowdsourcing: practical tips, advice and examples from leading experts in the field. https://books.google.at/books?hl=de&lr=&id=mscjeFHY8NQC&oi=fnd&pg=PR4&dq=open+innovation+crowdsourcing&ots=2Nku9jDmRC&sig=L9CiT3ILk2plCtVmWn4YmGKca7w#v=onepage&q=open%20innovation%20crowdsourcing (accessed: 07.01.2018).

Wood, G., (2014). Ethereum: A secure decentralised generalised transaction ledger. Ethereum Project Yellow Paper, 151.

# Towards the Construction of a Lexical Data and Technology Ecosystem: The Experience of ILC-CNR

**Monica Monachini, Anas Fahad Khan**
Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche
Via Moruzzi 1 - Pisa,
{monica.monachini, fahad.khan}@ilc.cnr.it

**Abstract**
This paper describes the activities and projects being carried on at the "A. Zampolli" Institute for Computational Linguistics (ILC) at the crossroads between computational lexicography and e-lexicography and that are intended to assist in the creation of a queryable and interconnected ecosystem of standardised lexicographic datasets and technologies.

**Keywords:** e-lexicography, computational lexicography, lexical resources, standards, LOD

## 1. Introduction

Lexicography is traditionally recognised as the branch of applied linguistics that is concerned with the design and construction of practical resources for describing the lexicon of a language. In the last few decades or so the marriage of lexicography and digital technology has resulted in the creation of two new disciplines: e-lexicography, i.e., the compilation of digitally-born dictionaries for human users, and computational lexicography, a sub-branch of computational linguistics that deals with the use of lexicons in Natural Language Processing as well as with the use of computational techniques in building and enriching lexicons (for NLP purposes).

The use of language technology has had an important impact on the task of compiling dictionaries for human use. Not only do modern day technologies allow for the easier digitization of lexical resources, but current trends in language resources and data science make it possible to imagine the fulfilment, in the very near future, of one of the most important promises of e-lexicography - namely that of a large-scale interconnected ecosystem of of open, queryable and standardised lexicographic datasets and technologies. In fact it seems as if e-lexicography's moment may finally have arrived.

In the rest of this article we will describe some of the activities, past and present, in which ILC has been involved and/or still is involved and which we believe make a strong contribution towards this ultimate aim.

## 2. Lexical Resources, Standards and Infrastructures

ILC-CNR can boast of a long-standing involvement in computational lexicography dating back to the pioneering work of Antonio Zampolli and others[1]. These early activities eventually resulted in the creation of influential lexical resources such as PAROLE SIMPLE CLIPS (PSC[2]) and ItalWordnet (IWN[3]).

Aside from the creation of language resources, however, another important and salient aspect of the work carried out at ILC is the participation of its members, and in particular, those of the Language Resources and Infrastructures group (LaRI[4]) within the institute, in important standardisation projects and initiatives, such as LIRICS[5] and LMF[6] (Francopoulo 2013).

LMF for instance is an influential standard within the field of computational linguistics and language technologies; it is also important for lexicographic resources intended for human users. The LMF core model is currently being revised as a multipart standard. One of the other parts of the standard aims at a level of higher interoperability with TEI through the production of a TEI-XML serialisation of LMF[7]. In addition, a new module for etymology is being added to the new version of the LMF core[8].

Two important infrastructural projects in which the LaRI group is involved are PARTHENOS[9], and ELEXIS[10]. The former project includes the presence of various European infrastructures, such as DARIAH and CLARIN and has the goal of consolidating shared practices and data models among various domains within the humanities. A number of different standardisation initiatives are currently taking place within Parthenos with the aim of improving the interoperability of lexical resources including digitized dictionaries. The latter project -- ELEXIS -- begins in February 2018 and is an ambitious project within the domains of NLP and e-lexicography with the aim of creating a European wide lexicographic infrastructure. Several different standardisation efforts are likely to converge within the ambit of this project, in particular those carried out under the banner of the International Organization for Standardization (such as LMF) along with those newly emerging standards for Linked Open

[1] For an overview see (Calzolari, Monachini, and Soria 2013).
[2] http://hdl.handle.net/20.500.11752/ILC-88.
[3] http://hdl.handle.net/20.500.11752/ILC-65
[4] http://lari.ilc.cnr.it/
[5] http://lirics.loria.fr/
[6] LMF has been developed under the aegis of the ISO Committee TC37/SC4 (ISO-24613:2008)
[7] Here with a strong participation of ILC (one of the members of LaRI is the co-leader of the LMF working group).
[8] Here too with the participation of one of the members of LaRI who has a co-leader role.
[9] www.parthenos-project.eu
[10] ELEXIS is based on a previous Cost Action ENeL - aiming to establish a European network of lexicographers and a common approach to e-lexicography that forms the basis for a new type of lexicography (http://www.elexicography.eu/).

Data, developed under the banner of W3C (Ontolex-Lemon).

A strong impetus has been provided to the research directions mentioned above within ILC by the institute's official role as the leading Italian participant in the CLARIN-ERIC infrastructure[11]. Standardisation activities and the promotion of shared formats are crucial for CLARIN, and a Standard Committee is active within the infrastructure together with a task force dealing with Interoperability. While formats and best practices for corpora have been central till now, we foresee that the standardisation of lexical resources, and especially lexicographic resources, will become more and more important in the coming years. This is something that ILC, with its decades-long experience in standardisation, is well placed to make an significant contribution to.

## 3. Semantic Web Standards

In addition to activities described above ILC also has a strong commitment towards the adoption of semantic web technologies. In 2015, the semantic layer of PSC (Del Gratta et al. 2015) as well as ItalWordNet (Bartolini, Del Gratta, and Frontini 2013) were published as Linked Open Data (LOD), the former using the lemon model[12]. Other notable Semantic Web resources in whose creation ILC has been instrumental are the GeoDomain Wordnets (Frontini, Del Gratta, and Monachini 2016) – which connect the Geonames ontology with the ItalWordNet and Princeton Wordnets – and the sentiment lexicon for Italian (Maks et al. 2014).

Moreover ILC also participates in the W3C activities of Ontolex-lemon[13], with a particular focus on the modelling of dictionaries as well as the representation of etymology and language change. More broadly, ILC has carried out work on the creation of resources for historical languages (for instance the creation of Ancient Greek Wordnet[14] and the publication of the Intermediate Liddell Scott lexicon (1896) as LOD[15]). This interest for semantic web technologies extends towards other aspects such as the modelling of ontologies with OWL and the use of the semantic web rule language (SWRL).

With respect to the former the institute has published the OWL version of the SIMPLE ontology (Toral and Monachini 2007). As to the modelling of rules, an ongoing project aims at the translation of Italian inflexional morphology using SWRL (Khan et al. 2017).

[11] The Italian MIUR nominated the Department of Humanities and Social Sciences of CNR as the National Representative and gave ILC-CNR the role of building the national data center and the national repository (ILC4CLARIN, https:// ilc4clarin.ilc.cnr.it/). Monica Monachini was nominated National Coordinator of the CLARIN-IT Consortium.

[12] http://hdl.handle.net/20.500.11752/ILC-66, http//www.languagelibrary.eu/owl/simple/

[13] https://www.w3.org/community/ontolex

[14] See Bizzoni et al. (2014, 2015) and Del Gratta et al. (2015).

[15] Khan et al. (2016).

## 4. Towards an Ecosystem of Lexical Resources

These initiatives should be seen within the broader context of a new convergence of the once closely aligned but laterly somewhat estranged communities of language resources and digital humanities. In particular, we seem to be witnessing a new convergence between computational approaches to lexicography and the needs of e-lexicography. As more and more language technologists are collaborating on digital humanities projects the necessity of making the main formats (TEI, LMF, Ontolex-lemon) interoperable becomes more important and, at the same time, the encoding of levels of information that are of particular interest for DH - such as the representation of diachronic knowledge and language change - becomes essential.

The coexistence of various different competing standards is always a source of worry. Moreover, current lexicographic resources, both modern and historical, have different levels of structuring and are not equally suitable for application in other fields. However, we believe that current trends seem to be consistent with the idea of an ecosystem, where different standards can coexist and mutually enrich each other, with

i. TEI being a format for representing a digital edition of the lexical resource,
ii. LMF the basic tool for actionable lexicons within LT, as well as in contexts where an official ISO standard is required, and
iii. Ontolex-Lemon the standard format for interconnected lexical networks, in which individual datasets can refer back to TEI sources (when they exist).

This intuition underlies the current vision that ILC is promoting, i.e. developing strategies, tools and standards for extracting, structuring and linking lexicographic resources to unlock their full potential for LOD and the Semantic Web, as well as in the context of Digital Humanities.

We aim to create a unified platform for interlinked lexical resources with a focus on Italian and on classical languages, where language resources are distributed in different formats for different purposes and are:

- accessible by web based query interface for linguists, lexicographers, students and the general public;
- downloadable in various formats (via the ILC4CLARIN repository[16]);
- exposed as LOD, browsable through a SPARQL query interface (as a service of CLARIN-IT) for lexicographic linked open data.

[16] The list of ILC-CNR lexical conceptual resources (mentioned here) is available in the ILC4CLARIN repo: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/discover?filtertype=type&filter_relational_operator=equals&filter=lexicalConceptualResource

The Linguistic LOD paradigm provides a suitable approach for the development of such an ecosystem.

## 5. References

Bartolini, Roberto, Riccardo Del Gratta, and Francesca Frontini. 2013. "Towards the Establishment of a Linguistic Linked Data Network for Italian." In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and Other Language Data*, 76–81. Pisa, Italy: Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-5512.

Bizzoni, Yuri, Federico Boschetti, Riccardo Del Gratta, Harry Diakoff, Monica Monachini, and Gregory Crane. 2014. "The Making of Ancient Greek WordNet." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, 1140–47. Reykjavik, Iceland: ELRA. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1071_Paper.pdf.

Bizzoni, Yuri, Riccardo Del Gratta, Federico Boschetti, and Marianne Reboul. 2015. "Enhancing the Accuracy of Ancient Greek WordNet by Multilingual Distributional Semantics." In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-It 2015*, 47. Accademia University Press.

Calzolari, Nicoletta, Monica Monachini, and Claudia Soria. 2013. "LMF – Historical Context and Perspectives." In *LMF Lexical Markup Framework*, edited by Gil Francopoulo and Patrick Paroubek, 1–18. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118712696.ch1.

Del Gratta, Riccardo, Francesca Frontini, Fahad Khan, and Monica Monachini. 2015. "Converting the Parole Simple Clips Lexicon into Rdf with Lemon." *Semantic Web Journal* 6 (4):387–92.

Del Gratta, Riccardo, Federico Boschetti, Angelo Del Grosso, Fahad Khan, and Monica Monachini. 2015. "Cooperative Philology on the Way to Web Services: The Case of the CoPhiWordNet Platform." In *Worldwide Language Service Infrastructure*, edited by Yohei Murakami and Donghui Lin, 173–87. Lecture Notes in Computer Science 9442. Springer International Publishing. https://doi.org/10.1007/978-3-319-31468-6_13.

Francopoulo, Gil, ed. 2013. *LMF Lexical Markup Framework*. John Wiley & Sons.

Frontini, Francesca, Riccardo Del Gratta, and Monica Monachini. 2016. "GeoDomainWordNet: Linking the Geonames Ontology to WordNet." In *Human Language Technology. Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers*, edited by Zygmunt Vetulani, Hans Uszkoreit, and Marek Kubis, 9561:229–42. Lecture Notes in Computer Science (LNCS). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-43808-5_18.

Khan, Fahad, Andrea Bellandi, Francesca Frontini, and Monica Monachini. 2017. "Using SWRL Rules to Model Noun Behaviour in Italian." In *LDK 2017: Language, Data, and Knowledge*, 134–42. Lecture Notes in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-319-59888-8_11.

Khan, Fahad, Francesca Frontini, Federico Boschetti, and Monica Monachini. 2016. "Converting the Liddell Scott Greek-English Lexicon into Linked Open Data Using Lemon." In *Digital Humanities 2016: Conference Abstracts*, 593–96. Kraków: Jagiellonian University & Pedagogical University. http://dh2016.adho.org/abstracts/236.

Liddell, Henry George, and Robert Scott. 1896. *An Intermediate Greek-English Lexicon: Founded upon the Seventh Edition of Liddell and Scott's Greek-English Lexicon*. Harper & Brothers.

Maks, Isa, Ruben Izquierdo, Francesca Frontini, Rodrigo Agerri, Piek Vossen, and andoni Azpeitia. 2014. "Generating Polarity Lexicons with WordNet Propagation in 5 Languages." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA).

Toral, Antonio, and Monica Monachini. 2007. "Formalising and Bottom-up Enriching the Ontology of a Generative Lexicon." In *Proceedings of RANLP07-Recent Advances in Natural Language Processing*.

# Developing a Multi-functional e-Spelling-Dictionary – a South African Perspective

**Elsabé Taljard, D J Prinsloo**
Department of African Languages, University of Pretoria
University of Pretoria, Private Bag X20, Hatfield 0028.
elsabe.taljard@up.ac.za, danie.prinsloo@up.ac.za

**Keywords:** e-spelling-dictionary, online lexicographic support, lesser-resourced languages

## Abstract

For any language to attain or retain the status of a fully-fledged, official language of higher functions, the availability of sophisticated online lexicographic support tools is essential. In this paper we discuss the design and development of a multi-functional e-spelling-dictionary which is linked to the spelling rules of a particular language. The tool specifically caters for the on-the-fly user's spelling needs but also provides for cognitive information. It can be accessed departing from its main menu with clickable options to the different functions of the word list, e.g. illustration of a particular spelling rule, or as plug-in to an existing e-dictionary or word processor. For illustrative purposes, we refer to Afrikaans (AFR) and Northern Sotho (NSO), two of the official languages of South Africa. In terms of lexicographic e-resources, both languages can be described as lesser-resourced, furthermore, existing electronic resources catering specifically for spelling guidance can at best be described as rudimentary, not providing adequate provision for users' needs. Although spell checkers exist for both languages, these are stand-alone tools which provide no additional guidance with regard to broader spelling issues. The definitive spelling guide for Afrikaans, i.e. the *Afrikaanse Woordelys en Spelreëls* ('Afrikaans word list and spelling rules') is currently only available online for paying clients, which severely restricts its accessibility. A further shortcoming of the online spelling guide for Afrikaans is that it provides no additional cognitive support to the user; there is for example no clickable link guiding the user from the look-up to the actual spelling rule that determines the spelling of a particular word. For Northern Sotho, a spelling guide is available in paper format, but due to external factors, this guide is not generally available to the Northern Sotho linguistic community. It also does not address the issue of standardized versus non-standardized forms. No online spelling resource except the spell checker referred to above, is available. The spell checker for Northern Sotho is only available on CD (https://spel.co.za/en/product/african_spelling_checkers/), which renders it obsolete for many potential users, since the latest generation of PCs and laptops are no longer fitted with a CD-ROM drive. This spell checker is furthermore a static tool, which does not make provision for regular updates. Consequently, having been developed in 2009, the wordlist against which spelling is checked is already outdated. One of the advantages of having an e-tool is the fact that it can be updated in real time. Northern Sotho furthermore carries the additional burden of grappling with standardization issues, leading to a proliferation of spelling variants – an aspect which needs to be considered in the design of the tool.

The design of the proposed tool is approached from the perspective of the user, one of the prevalent approaches in modern day metalexicography. This perspective compels compilers of any lexicographic product to compile their dictionaries according to the information needs and research skills of a well-defined target user group.

"The user-perspective, so prevalent in modern-day metalexicography, compels lexicographers to compile their dictionaries according to the needs and research skills of well-defined target user groups. The dominant role of the user has had a definite effect on the compilation of dictionaries as well as on the evaluation of their quality. Good dictionaries do not only display a linguistically sound treatment of a specific selection of lexical items.Good dictionaries are products that can be used as linguistic instruments by their respective target user groups. The better they can be used, the better dictionaries they are." (Gouws and Prinsloo 2005:39).

Designing and developing a lexicographic tool of high quality therefore implies an analysis of the target users' information needs. According to Tarp (2009), analysis of log files of online dictionaries can give a good indication of users' needs. Having access to the log files of an earlier online version of the Afrikaans word list enabled us to identify a number of pertinent user's needs. The most obvious advantage of log file analysis is the identification of gaps in the lemma selection of the word list. Secondly, it provided us with valuable information on frequently misspelled and frequently confused words. Lastly, the analysis clearly indicated the need for additional cognitive information – users would often look for information regarding a specific orthographic issue, e.g. the use of circumflexes or hyphens. This implies that users need direct access to the rule that determines the use of these orthographical signs.

Based on the analysis of users' needs, the proposed tool has four functions, i.e. confirmation of correct spelling, direct spelling guidance, provision of cognitive information and guidance on frequently confused words. These functions far supersede the sophistication of currently available spelling checkers for these two languages.

Should users type in a correctly spelled search word (e.g. *galesome* 'ten times' (NSO), *lêer* 'file' (AFR)) and find it in the word list without any additional comments, it serves as **confirmation** that (a) the search word is correctly spelled and (b) the search item is regarded as part of the standard language. In the case of Northern Sotho a further indication as to the formally standardized status (or not) of

the search word is to be provided. An optional link to additional cognitive information is envisaged, e.g. to the relevant spelling rule pertaining to a specific category e.g. the spelling of derived adverbs (NSO), or the use of the circumflex (AFR). Users are also able to search for specific categories within the e-spelling rules, e.g. spelling of derived adverbs (NSO), or spelling of words with circumflexes (AFR).

Should the user type in a search word which is incorrectly spelled, (e.g. *ga lesome* 'ten times' (NSO), *leêr* 'file' (AFR)) **guidance on the correct spelling** is provided in the form of "Did you mean (*galesome* (NSO), *lêer* (AFR))?", or by means of the typical spellchecker function of a pop-up box containing a list of related correct spellings. Again, a clickable option giving access to further cognitive information in the form of the relevant spelling rule is provided.

Lastly, the tool also provides direct guidance with regard to words the spelling of which is **often confused**. Should users type in a search word which is correctly spelled (*wyer* 'wider' (AFR)), but which has shown itself to be frequently confused with another word (*weier* 'refuse'), they are alerted to the possibility that they may have spelled the search word correctly, but that the spelling and the intended meaning may not match.

Having access to (free) online spelling resources will not only provide adequate support to users, but will also contribute to strengthening the status of these languages as languages of higher functions.

The tool is still in the design phase, but most of the required components have been built, such as corpora and frequency lists for the two languages, pop-up boxes, cognitive information screens, spelling rules, typically misspelt words and alternative spell checkers for both languages.

## Bibliographical References

Tarp, S. (2009). Reflections on lexicographic user research. *Lexikos*, 19: pp. 257 – 296.
Gouws, R.H and Prinsloo, D.J. (2005). *Principles and practice of South African lexicography*. Stellenbosch: African Sun Media.

## Acknowledgements

Financial support by South African Centre for Digital Language Resources (SADiLaR) is hereby acknowledged.

# Historical Lexicography of Old French and Linked Open Data: Transforming the resources of the *Dictionnaire étymologique de l'ancien français* with OntoLex-Lemon

**Sabine Tittel**[*], **Christian Chiarcos**[◇]
[*]Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany
[◇] Goethe University Frankfurt, Frankfurt am Main, Germany
sabine.tittel@urz.uni-heidelberg.de, chiarcos@informatik.uni-frankfurt.de

**Abstract**

The adaptation of novel techniques and standards in computational lexicography is taking place at an accelerating pace, as manifested by recent extensions beyond the traditional XML-based paradigm of electronic publication. One important area of activity in this regard is the transformation of lexicographic resources into (Linguistic) Linked Open Data ([L]LOD), and the application of the OntoLex-Lemon vocabulary to electronic editions of dictionaries. At the moment, however, these activities focus on machine-readable dictionaries, natural language processing and modern languages and found only limited resonance in philology in general and in historical language stages in particular. This paper presents an endeavor to transform the resources of a comprehensive dictionary of Old French into LOD using OntoLex-Lemon and it sketches the difficulties of modeling particular aspects that are due to the medieval stage of the language.

**Keywords:** Linked Open Data, Ontolex-Lemon, Lexicography, Old French

## 1. Introduction

### 1.1. The Lexical Resource

The *Dictionnaire étymologique de l'ancien français* – DEAF (Baldinger, since 1971) is a longstanding dictionary compiled in Heidelberg under the aegis of the Heidelberg Academy of Sciences and Humanities. Its aim is to document and study the Old French language from its first resource 842 AD until ca. 1350 AD. To date, the publication channel of the outcome of the editorial process is twofold: The dictionary is traditionally published as a series of printed books (via LaTeX) and, since 2010, also as a versatile electronic dictionary (DEAF*él*) with on-line dictionary entries and elaborate research functions based on the XML and XHTML data exported from a MySQL database.[1]

However, DEAF*él* constitutes a data silo. The information stored can be accessed either by reading the articles or by using the research functions offered by the publication. This has the following shortcomings: Regardless of the high standard of the on-line publication, the accessibility and usability of the dictionary is to be improved. Using the dictionary may require a considerable knowledge of Old French in general and about the internal structure of the dictionary in particular. This is not necessarily given. To answer a research question (say, about the concepts of health and illness in medieval society based on Old French literature) is not an easy task for someone who is not familiar with the Old French terminology for the respective domain (here, medicine).

Also, the internal data format of such a data silo is proprietary and its publicly accessible serialization focuses solely on human consumption. It does not allow for queries that have not been foreseen a priori. Most importantly, the data format is not well suited for automatic processing.

Thus, by transforming the data into RDF and Linked Open Data (LOD), we want to emancipate the valuable dictionary outcome from the limits of such a data silo.

### 1.2. Facilitating Resource Interoperability with the Resource Description Framework

Following the emergence of the internet, the Resource Description Framework (Klyne et al., 2004, RDF) was developed as a standard to represent metadata, and to express relations between and statements about web resources as well as offline resources. The aim is to facilitate processability and interpretability of metadata entries, but, subsequently, also of web resources themselves. Beyond its original use case, RDF thus rose to importance as a cornerstone of the emerging Semantic Web and even beyond classical Semantic Web applications that involve reasoning, inference and formal knowledge bases. RDF established itself as a generic representation formalism for data on the web and, in particular, for the *integration* of data on the web. In this role, a rich technological ecosystem evolved and ultimately lead to the emergence of Linked Data and its adaptation in various fields, e.g., as Linguistic Linked Open Data (LLOD) in linguistics and natural language processing. Our objective here is to facilitate the usability, queriability and interpretability of DEAF data for *automated* consumption and transformation. On the basis of such automated processes, more advanced functionalities for the end user can then be developed, e.g., improved means of querying, exploring or integrating other lexical or textual data sets. Such services are our ultimate goal, and we address first steps towards the development of (L)LOD-based methodology and infrastructure for historical philologies.

RDF implements a (multi-)graph model, where nodes are connected via edges that point from a source node ('subject') to a target node ('object') and that have a particular semantic type ('property'). Source nodes, target nodes and properties are identified with URIs, e.g., objects accessible

[1] https://deaf-server.adw.uni-heidelberg.de/ [accessed 12-12-2017].

via HTTP. RDF is thus naturally suited to describe structured data on the web. In particular, this includes lexical data, as the (directed multi-)graph is generally recognized to be a generic formalism for the representation of dictionaries and machine-readable lexical resources. As such, already the Lexical Markup Framework (Francopoulo et al., 2006, LMF) built on *feature structures* (largely equivalent to directed multi-graphs, but serialized in XML), and the increasing popularity of OntoLex-Lemon (and RDF) for lexical resources mostly reflects a transition from traditional XML-based representations to RDF-based representations of the same underlying data structure (Gracia et al., 2018). In opposition to XML which provides validation on a syntactic level only, the RDF data model allows to formalize the semantics independently from constraints on their order of representation. It is thus more suitable to establish interpretability and semantic processability of the data by its subsequent users and downstream applications.

On a format level, RDF can be serialized in different ways. A common text-based representation is the Turtle format that allows to express statements in the form of *triples*, including the subject URI, the property URI and the object URI (or, alternatively, a literal value), followed by a dot. (Various shorthands are possible.) The W3C-standardized query language SPARQL basically follows the same notation for graph fragments to be retrieved but extends it with variables. In the examples below, we employ a Turtle serialization of RDF data because it is particularly well-suited for subsequent querying.

For transforming the DEAF into RDF, we implemented the following workflow: We firstly selected one exemplary dictionary article as data for a proof of concept implementation. Using this data, we defined an application profile for the dictionary entries. Secondly, we transformed the XML data of the selected article into LOD with RDF/Turtle and the OntoLex-Lemon vocabulary in line with the application profile. This step was performed manually. Thirdly, we developed a set of XSLT scripts to automatically perform this transformation step and we evaluated problematic issues within this step. We then tested the scripts with the data of the respective article and also with the data of further dictionary entries. Finally, directions for future work have been identified.

## 1.3. Linked Data for Lexical Resources

Linked Data has emerged as a paradigm for publishing and interlinking datasets about ten years ago. It has been a success story, leading to many datasets being published following the four Linked Data principles (Bizer et al., 2009):

- Use URIs as (unique) names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using Web standards such as RDF, and SPARQL.
- Include links to other URIs, so that they can discover more things.

Applying Linked Data principles to modeling lexical data comes with important advantages (Chiarcos et al., 2013), most notably *structural interoperability* (same format, same query language), *conceptual interoperability* (shared vocabularies), *accessibility* (uniform access, data can be accessed using standard Web protocols without additional software, etc.), *resource integration* (linking resources) and *federation* (cross-resource access).

Most important for our use case is interoperability: By resorting to RDF as data model, one achieves structural interoperability as language resources following the Linked Data paradigm are provided according to a uniform data model (in different, equivalent and convertible serializations). Conceptual interoperability, i.e., the use of shared vocabularies, is encouraged in Linked Data since its nature encourages the reuse of existing vocabularies across datasets. Following this practice thus leads to more and more datasets using the same vocabulary to describe data. Hence, it facilitates to establish interoperability on both the syntactic (format / access) and the semantic (conceptual) level.

One vocabulary that rose to particular importance for lexical resources is the *Lexicon Model for Ontologies* (Lemon). The Lemon model has originally been developed in the Monnet project to augment ontologies with rich linguistic information in order to facilitate their automated rendering in natural language (Declerck et al., 2010). Since 2012, the Ontology-Lexicon W3C Community Group has been further developing this model towards a generic data model for lexical resources, and its application to the historical lexicography of a medieval language variety is the main contribution of our paper.

Despite the growing popularity of the Linked Data paradigm in application to lexicographic resources (Witte et al., 2011; Bouda and Cysouw, 2012; Declerck et al., 2015), and in particular, adaptations of Lemon (Borin et al., 2014; Klimek and Brümmer, 2015; Bosque-Gil et al., 2016; Gracia et al., 2018), the focus of current activities in this direction lies on the modern stages of the languages. Notable exceptions in this context include etymological dictionaries, e.g., on Germanic languages (Chiarcos and Sukhareva, 2014), and dictionaries of classical languages, e.g., on Ancient Greek (Khan et al., 2017). To our best knowledge, however, these approaches take a technological focus in that they aim to demonstrate the applicability of digital methods in the humanities, rather than being grounded in philological research or traditions. This gap in research is being addressed in this paper: We present an endeavor to transform the resources of a comprehensive dictionary of Old French into LOD using OntoLex-Lemon and we evaluate the difficulties of modeling particular aspects that are due the medieval stage of the language.

## 1.4. The OntoLex-Lemon Data Model

In its published version from May 2016, the OntoLex-Lemon model[2] is divided into five modules: The *OntoLex core model* describes the elements that are necessary for all instantiations of the model, including lexical entries, forms

[2] `https://www.w3.org/2016/05/ontolex/` [accessed 12-12-2017]

and senses of a word. The *syntax and semantics* module describes in more detail the interaction of the syntax of words and their interpretation in an ontology. The *decomposition module* is used to describe the composition of multi-word terms and compound words. The *variation and translation module* supports the description of relationships between words and senses including translation and cross-lingual links. Finally, the *metadata module* allows for high-level descriptions of a lexicon and the number of links between elements.

The primary class in the OntoLex model is the *lexical entry*, which represents a head word in the lexicon. The lexical entry groups all forms of a word together into a single element, e.g., it includes inflected forms. For example, the entry for the Old French verb *jogler* "to ridicule someone" (< Latin JOCULĀRE v.) would include inflected forms such as *joglant*, *joglot*, *joglé*. However, the Old French noun *jogler* m. "juggler" (< Latin JOCULĀRIS adj. "funny") with a different part of speech and a different etymology would logically represent a separate lexical entry. Lexical entries are further grouped into three classes: (single) words, multiword expressions and affixes (such as *anti-*).

A lexical entry is composed of a set of lexical forms, each of which can be represented in different scripts by means of a string; one of the forms can be defined as the canonical form (i.e., the lemma). Thus, the simplest form of a lexical entry (e.g., Old French *flamesche* f. "spark") is as follows:

```
PREFIX ontolex:
       <http://www.w3.org/ns/lemon/ontolex#>

<flamesche>
  a ontolex:LexicalEntry, ontolex:Word ;
  ontolex:canonicalForm
    <flamesche#singular_form> ;
  ontolex:otherForm
    <flamesche#plural_form> .

<flamesche#singular_form> a ontolex:Form ;
  ontolex:writtenRep "flamesche"@fro .

<flamesche#plural_form> a ontolex:Form ;
  ontolex:writtenRep "flamesches"@fro .
```

The semantics of a lexical entry can be given by indicating that it *ontolex:denotes* an element in the ontology. The element in the ontology can be a class, a property or an individual. In many cases, this link to the ontology may need to be described in more detail. For this purpose, the model provides the class *lexical sense*, representing the connection between a lexical entry and its meaning in an ontology or knowledge graph. Unlike such 'semantic' entities provided by an external resource, lexical senses are specific to one particular lexical entry.

As a rule of best practice, lexical entries should be linked to ontologies via their respective lexical senses whenever an explicit definition or gloss is provided in the original dictionary. In this way, it is always possible to inspect their original definition regardless of possible (subsequent) updates of the definition (or usage patterns) of the ontological entity they refer to (Wang et al., 2011). Accordingly, lexical resources become more robust and verifiable in the face of concept drift in the Semantic Web. A simple example (extending *flamesche*) is the following:

```
PREFIX dbpedia:
         <http://www.dbpedia.org/resource/>

<flamesche> ontolex:sense
                      <flamesche#sense1> .

<flamesche#sense1> a ontolex:LexicalSense ;
   ontolex:reference dbpedia:Spark_(fire) .
```

As lexical senses are specific to individual lexical entries, lexical concepts have been added to the model to express groups of lexical senses that can be lexicalized in different ways. The exact definition of such *lexical concepts* is resource-specific, but one possiblity is to use them to represent sets of synonyms.[3] In particular, lexical concepts can be used for lexical entries that are defined with reference to (the definition of) another lexical entry, e.g., using conventional expressions such as *see also*, *cf.*, etc. However, in this case, also the definition of the referred lexical entry must be reflected as a lexical concept:

```
<flamesche> ontolex:sense
                      <flamesche#sense1> ;
  ontolex:evokes
            <flamesche#sense1_lexConcept> .

<flamesche#sense1_lexConcept>
  a ontolex:LexicalConcept ;
  ontolex:isConceptOf
                    dbpedia:Spark_(fire) ;
  ontolex:definition "petite parcelle ...
       ..., flammèche, braise légère"@fr ;
  ontolex:lexicalizedSense
                      <flamesche#sense1> .
```

## 2. Resource Modeling

To illustrate the modeling of a complete dictionary entry, we chose the Old French word *fiel* m. for it has an average complexity in terms of both its orthographic challenges and its semantic structure: *fiel* is the standard graphical representation of the Old French word (and is thus defined as the lemma of the entry) and it shows six more graphical realisations within the Old French literature, i.e., *fel*, *feel*, *fele*, *feil*, *fieil* and *fius*. Its semantic scope includes three main senses, i.e., "bile", "gall bladder" and, figuratively, "bitterness". The editor of the dictionary entry identified 13 sub-senses altogether, among which are collocations and metaphors (see the entry in its collapsed version in Fig. 1). Also, some of the lexical units (i.e., the entity of the lexeme *fiel* plus exactly one of its senses) are elements of the medical or the botanical terminology (e.g. in Fig. 2).

Following the core model of OntoLex-Lemon we defined the application profile for the DEAF entries. We visualize this in Fig. 3 (*fiel* with main sense n°1 "bile" [medical term]) and Fig. 4 (*fiel de la terre* "plant of the family of the common centaury, Centaurium erythraea Rafn." [botanical term], modeled as a multi-word term).

[3] This practice is not required by the model, and broader definitions are possible. In particular, a lexical concept cannot always be interpreted as a synset in the sense of WordNet (Miller, 1995).


Figure 1: DEAF*él* entry 'fiel', collapsed version.

1° t. de méd. "liquide verdâtre et amer qui est contenu dans la vésicule biliaire, bile" (dep. ca.1160, EneasS$^2$ 8221 [*el cors m'as mis une amartume Peor que suie ne que fiel*]; GautArrErR 3688; QSignesK 250 [*descendra dou ciel la cengle Que vos apelez arc ou ciel. Couleur avra semblant a fiel*]; TrotulaTrinH 212; HArciPèresO 442; SongeDan$^6$H 304; ConsBoëceTroyS II 139; Fevres f°25r°(3); f°97v°10; f°165v°2; etc.; etc.etc.; GlParR 3253; [AalmaR 3983; etc.etc.], TL 3,1819; ANDEl; GdfC 9,616b; DMF; TLF 8,844b ['vieilli']; FEW 3,445a)

Figure 2: DEAF*él* entry 'fiel', main sense n°1, partly expanded version.

Beyond the OntoLex-Lemon core vocabulary we used classes and properties of the following ontologies: the OntoLex *decomposition module* (`decomp`[4]) to model the components of multi-word terms (`ontolex:MultiwordExpression` with `decomp:subterm`), and the OntoLex *variation and translation module* (`vartrans`[5]) to model their relations (`lexicalRel`). To model the part-of-speech categories we used the LexInfo ontology (`lexinfo`[6]), and to expressing linguistic features beyond LexInfo (e.g., referencing language registers with `TechnicalRegister`), we used OLiA (`olia`[7]). As for metadata, FOAF properties define the name and website of the editor (`name`, `homepage`), DublinCore properties refer to the extralinguistic reality (`subject`) and also facilitate non-linguistic annotation (`creator`, `publisher`, `license`, `date`). Also, we defined new classes and properties to meet particular requirements of our use case: `deaf:TechReg` (technical register) defines specialized terminology and `deaf:idem` models the case where a sub-sense 'B' of a main sense 'A' inherits A's definition (and then specifies it in a certain way). The entity `deaf:TechReg` is defined as an instance of the OLiA class `olia:TechnicalRegister`; for `deaf:idem`, we found no existing vocabulary to be applicable.[8]

For the modeling process, we prioritized the lexical information, that is, the Old French lexemes including their written representations and their senses. However, this is a first step and the modeling currently ignores other relevant information such as the information given in the etymological discussion of each DEAF entry (etymon and corresponding words in other Romance and non-Romance languages), the dating of each lexical unit, the quotations taken from the Old French texts, and more. We thus identified the modeling of the hitherto excluded data as future work.

## 3. Converting DEAF to RDF

### 3.1. Manual Transformation

Preparing the transformation, we identified the following issue: The original XML data of a DEAF entry includes information that is not modeled by the application profile. We therefore isolated the data that is relevant for the transformation into RDF. The result is as follows (extract with only two graphical forms and one sense):

```
<?xml version="1.0"?>
<xsd:schema
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:m="http://www.deaf-page.de/ns/markup"
targetNamespace="http://www.deaf-
 page.de/ns/markup">

<article author="Sabine Tittel">
<title><lemma developed="false"
 language="afr.">fiel</lemma>
 <pos>m.</pos></title>

<variant type="standard">fiel</variant>
<variant>fel</variant>
[...]

<sense><description>
 <m:terminology type="medecine">
 t. de m&#xE9;d.</m:terminology>
 <m:definition>liquide verd&#xE2;tre et
 amer qui est contenu dans la
 v&#xE9;sicule biliaire,
 bile</m:definition></description>
</sense>
```

We then manually transformed the data of the entry *fiel* into RDF/Turtle. Finally, we reviewed the data using standard validation tools.

### 3.2. Automated Conversion

The application profile and the RDF data of *fiel* then served as a model for the creation of a set of XSLT scripts. In

[4] `http://www.w3.org/ns/lemon/decomp`.

[5] `http://www.w3.org/ns/lemon/vartrans`.

[6] `http://www.lexinfo.net/ontology/2.0/lexinfo`, Cimiano et al. (2011).

[7] `http://purl.org/olia/olia.owl`, Chiarcos and Sukhareva (2015).

[8] In particular, the `skos:broader` property of the Simple Knowledge Organization Scheme (Miles and Bechhofer, 2009) does not seem to be applicable as it should hold between SKOS concepts rather than between individuals. Accordingly, the former use of `skos:broader` within Monnet-Lemon has been considered deprecated and removed from the Ontolex-Lemon community report.

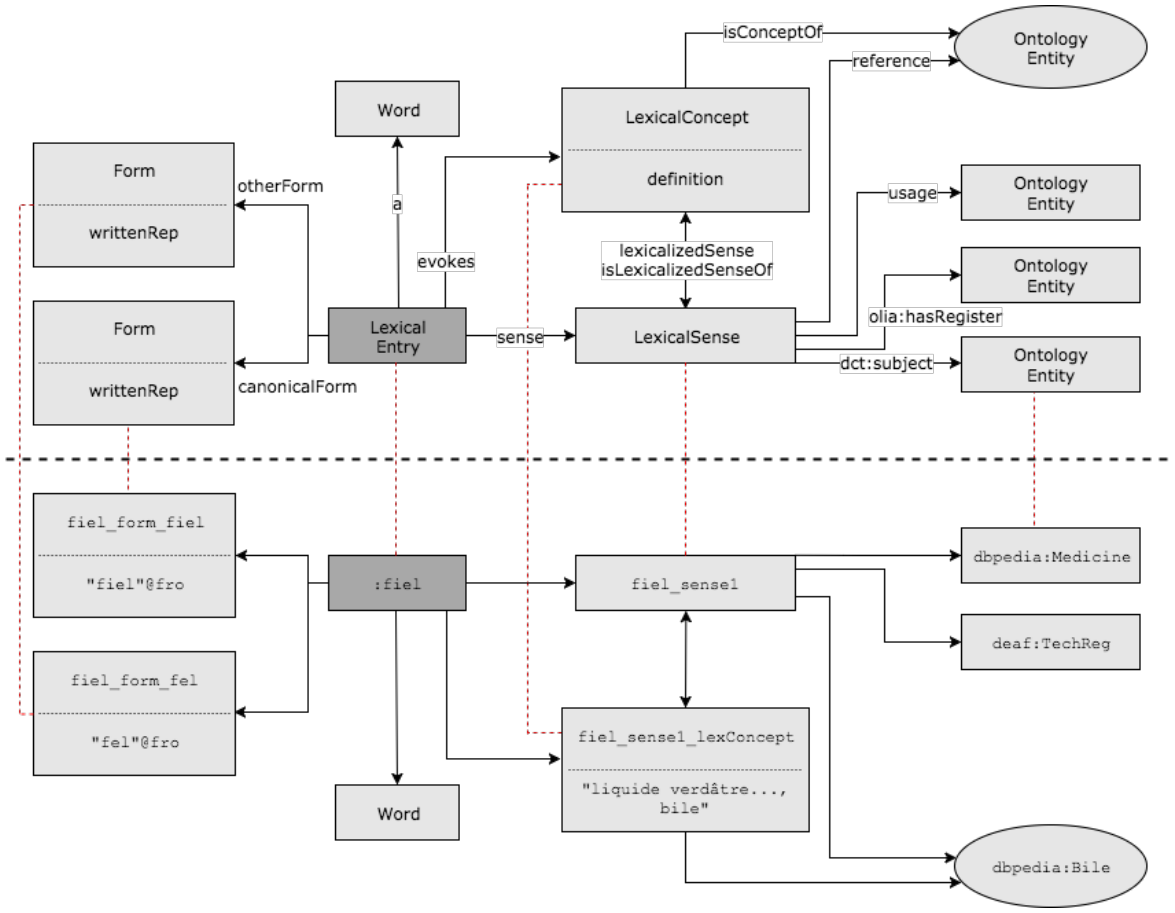



Figure 3: Model of DEAF entry 'fiel' with main sense n°1.

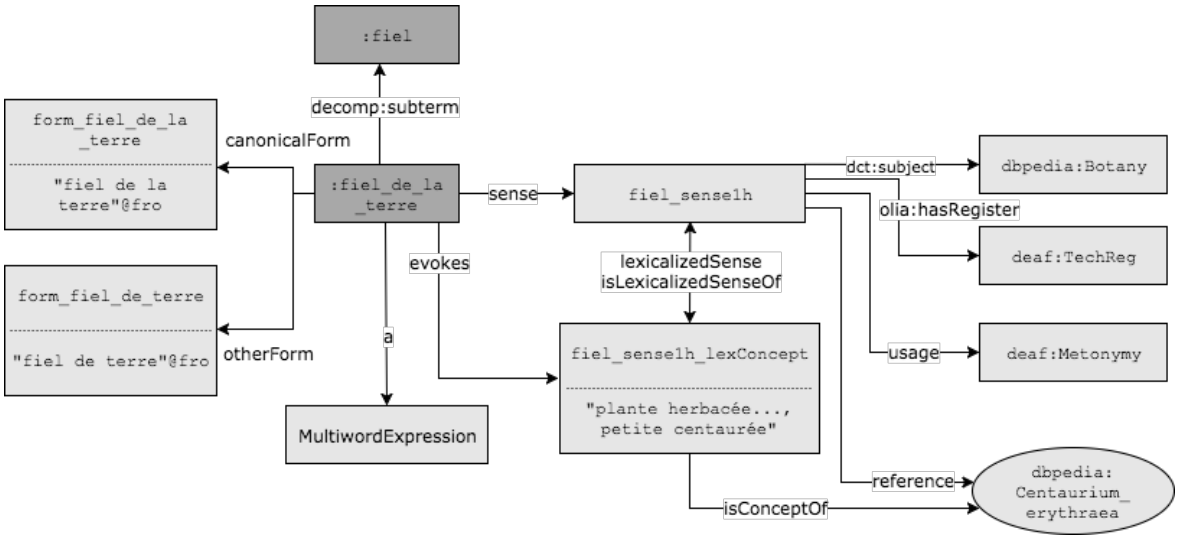



Figure 4: Model of multi-word term 'fiel de la terre'.

order to be able to eventually convert the total of approx. 83,000 dictionary entries, these scripts not only cover the specific use cases provided by our proof of concept article *fiel* but also all valid XML elements with their attributes and values defined by the XML schema of the dictionary. For example, we implemented a specific template for the automatic conversion of a given list of technical domains like medical, astronomical, musical terminology, etc. This template inserts links to the respective entity of DBPedia to define the type of terminology (using `dct:subject`, `olia:hasRegister`, `deaf:TechReg`, also literals). A fragment of the conversion template is illustrated in Fig. 5. An example of the outcome is:

```
:fiel_sense1
  a ontolex:LexicalSense ;
  dct:subject dbpedia:Medicine ,
    "medicine"@eng ;
  olia:hasRegister deaf:TechReg ,
    "t. de méd."@fr .
```

It should be noted that this representation aims for a middle ground between human and machine interpretability: We provide both the original information from DEAF (as a string value) and its semantic interpretation (with a URI that references external terminology repositories and knowledge bases), and in order to preserve their association, both are assigned as objects to the same property.

While this representation is lossless and allows to trace entity links in a (relatively) user-friendly, compact and unambiguous fashion that particularly facilitates their debugging, this is semantically valid only in the context of the general RDF data model. Nevertheless, it should be avoided in more strictly formalized Semantic Web languages such as OWL. However, a subsequent SPARQL Update script can easily eliminate literal values for OWL object properties such as `olia:hasRegister`, thereby deriving a more compact and semantically valid representation of DEAF that is suitable for consumption by downstream applications and users.

After conversion, we evaluated the validity of the XSLT scripts against the manually produced RDF data of *fiel*. In addition, a random sample of five further DEAF entries, resp. their Linked Open data conversion has been manually inspected and verified, indicating the applicability of our converter to DEAF data structures. However, this conclusion must be taken with a grain of salt when it comes to linking with external resources.

## 4. Linking DEAF Data

While data *structures* can be seamlessly converted to RDF, the generated outcome cannot always easily be put into relation with external knowledge bases. In particular, we find that the *Historical Semantic Gap* prohibits an unreflected and fully automated transition of philological resources of historical language stages with concept stores developed for present day applications and data: The mapping of a lexical unit to the correct entity in an ontology is a difficult task that cannot be automated for the Old French lexis. The reason for this is the historical dimension and the semantic gap lying therein: The extralinguistic concept of medieval reality denoted by a word in Old French oftentimes differs from the extralinguistic concept of modern reality denoted by the same word in modern French, e.g., because certain medical coherences were not yet known: For a 13$^{th}$ century doctor, *function of the brain* does not mean the same as for a 21$^{st}$ century one.

To overcome this problem, we implemented a semi-automatic process: This includes an automatic pre-processing as a time-saving preparation for a manual post-processing. The XSLT scripts place a wildcard (a simple `XXX`) where the entity of an ontology then needs to be specified by a linguist specialized in Old French lexical semantics. His expertise assures the correct mapping.

We believe it is possible to further enhance the automatic part of the procedure. For example, the sense definition of a botanical term is by default given in modern French but also includes the scientific Latin term of the plant. This term is usually taken from the *Systema naturae* by Carl von Linné (abbr. 'L.') or, less commonly, from the taxonomy by Carl Gottlob Rafn (abbr. 'Rafn', see above for *fiel de la terre*). We foresee an automatic mapping of these definitions to the entity in, e.g., DBPedia based on the scientific Latin term.

## 5. Discussion and Outlook

So far, we described the application of the OntoLex-Lemon model to modeling a reference resource for Old French lexicography as RDF, resp. its automated conversion to Linked Data – as well as limitations of a fully automated approach. To our best knowledge, this is the first broad-scale application of the Linked (Open) Data paradigm to a standard resource for medieval lexicography. We are aware of related activities on lexicographic resources for other language families, but we understand that these operate on the level of pilot studies, at the moment. Notable related work on medieval French beyond lexicography includes the Syntactic Reference Corpus on Medieval French (SRCMF[9]) use an RDF database as a backend for annotation graphs, albeit as an internal representation only, and without links to LOD resources. In fact, the actual data of the SRCMF is disseminated in a conventional XML format (Brants et al., 2004).[10]

The development of a LOD edition for the DEAF is conducted with the more general aim to transform the dictionary data into a sustainable and more easily re-usable format. The publication of the RDF edition of the full DEAF under an *open* license is foreseen by the first author, yet, it requires clarification about possible restrictions on use, dissemination and licensing – for these aspects, legal confirmation has been requested but is pending. The solution proposed is to model the role of the Heidelberg Academy of Sciences and Humanities using `dct:rightsHolder`.

With the LOD edition, we pave the way for the DEAF to become a part of the LLOD cloud in general and as a potential center within a net of linguistic resources of medieval French in particular. Beyond providing a novel set of philological lexical data in compliance with Linked Data principles, we also used this data to enrich a digitally published scholarly text edition of a medical treatise written in medieval French with references to the DEAF dictionary, as further described in Tittel et al. (accepted). This emphasizes the role of the DEAF as a standard reference also for other scholarly editions of Old French and Middle French texts. The conversion of the dictionary data into RDF and its publication within the LLOD cloud shows great capability of promoting the DEAF's role as a focal point of historical French text philology.

Apart from the afore-mentioned modeling of hitherto excluded data we identified two major issues that are yet to be

[9] `http://srcmf.org`, Mazziotta (2010).

[10] The distributed SRCMF RDF data is defective in the sense that '[t]he RDF file can be used to correct the annotation in NotaBene, but you need to pair it with the XML text source file.' (`http://srcmf.org` [accessed 03-02-2018]).

```
<xsl:template name="terminology_extern">
  <!-- the subject URI has been spelled out before -->
  <xsl:choose>
    <!-- when medicine -->
    <xsl:when test="./description/m:terminology/@type='medicine' or
      ./description/m:idem/m:terminology/@type='medicine'">
      dct:subject
        dbpedia:Medicine ,
        "<xsl:value-of select="./description/m:terminology/@type"/>
         <xsl:value-of select="./description/m:idem/m:terminology/@type"/>"@eng ;
      olia:hasRegister
        deaf:TechReg ,
        "<xsl:value-of select="./description/m:terminology"/>
         <xsl:value-of select="./description/m:idem/m:terminology"/>"@fr ;
    </xsl:when>
    <!-- when astronomy -->
    <xsl:when test="./description/m:terminology/@type='astronomy' or
      ./description/m:idem/m:terminology/@type='astronomy'">
      dct:subject
        dbpedia:Astronomy ,
        "<xsl:value-of select="./description/m:terminology/@type"/>
         <xsl:value-of select="./description/m:idem/m:terminology/@type"/>"@eng ;
      olia:hasRegister
        deaf:TechReg ,
        "<xsl:value-of select="./description/m:terminology"/>
         <xsl:value-of select="./description/m:idem/m:terminology"/>"@fr ;
    </xsl:when>
    <!-- etc. -->
  </xsl:choose>
</xsl:template>
```

Figure 5: XLST fragment for automated DEAF conversion.

addressed: language identification and sense hierarchies.

**Language identification:** The first issue concerns the modeling of the lemma and the (ortho)graphical variants of the respective word. We identify the Old French language in line with the International Standard for Language Codes ISO 639, i.e. with the ISO 639 code 'fro'.[11] We thus modeled the lemma and the variants using the OntoLex-Lemon vocabulary in the following way (*fiel* is the lemma = `canonicalForm`, *fel* is one variant = `otherForm`):

```
:fiel ontolex:canonicalForm
  :fiel_form_fiel .
:fiel_form_fiel a ontolex:Form ;
  ontolex:writtenRep "fiel"@fro .
:fiel ontolex:otherForm :fiel_form_fel .
:fiel_form_fel a ontolex:Form ;
  ontolex:writtenRep "fel"@fro .
```

However, it must be noted that – similar to the medieval stage of other Romance languages – Old French does not have a consistent orthographic norm. Each scribe of a manuscript realized the sound of a word in his own fashion, influenced by random circumstances but also by his dialect that could differ significantly from what we now consider the standard Old French language. As a consequence, we find a great variety of spellings for the same word.[12] Whenever a graphical variant is characteristic of a particular Old French scripta (i.e., the written form of a spoken dialect), the editor of the dictionary entry explicitly annotates it within the XML data of the entry. As a result, e.g., the entry *faisse*, designating a sort of ribbon or strap, lists Lorraine *faixe*, Anglo-Norman *fees*, and Picard *fasse* among the graphical variants.[13] Unfortunately, ISO 639 does not provide codes for Old French dialects,[14] and therefore, we provisionally identified all Old French dialects as standard 'fro'. But this is an intermediate solution because it ignores information that is very valuable for the research of Old

[11] `https://www.iso.org/iso-639-language-codes.html` [accessed 12-12-2017].

[12] The word with the highest number of attested variant spellings to date is the Old French adverb *iluec* "there" with more than 120 variants, see `https://deaf-server.adw.uni-heidelberg.de/lemme/iluec` [accessed 12-12-2017].

[13] `https://deaf-server.adw.uni-heidelberg.de/lemme/faisse` [accessed 12-12-2017].

[14] Varieties of historical language variants have been within the focus of ISO 639-6, which was, however, withdrawn as a standard in 2014, cf. `https://www.iso.org/standard/43380.html`. One possible alternative would be Glottolog `http://glottolog.org`, which does, however, take a focus on language documentation and is not appropriate for the needs of philologists. As an example, it conflates diachronic and dialectal criteria within a single hierarchy: The Romance language family is considered a subclass of Imperial Latin (as is, for example, Classical Latin), where – in fact – it evolved from it. Yet, this defective kind of modeling is not systematic, as Old Latin is a cousin of Imperial Latin rather than its ancestor/superclass.

French dialects. This information is given in the XML data but is lost in the LOD version. The solution to this shortcoming of the ISO 639 standard is to define the code 'fro' as a macrolanguage and to register the Old French dialects as varieties associated to 'fro'. A valid list of dialects is provided by the XML schema of the DEAF.

**Sense relations:** The second issue concerns the complex semantic relations between main senses and associated subsenses within the sense tree of a DEAF article. The hierarchical structure and the order of the sub-senses mirrors the semantic change the lexeme has undergone: It considers all figures of speech, e.g., metaphor, metonymy, irony, image, hyperbole, allegory, euphemism, etc. For each lexical unit of the respective lexeme the semantic relationship is explicitly expresses by, e.g., 'par métaph.', 'par méton.', 'par ironie'. This information is of great value for the study of semantic shift. We therefore attempt to model the semantic relationships expressed in the semantic tree. However, the properties of established vocabularies seem insufficient to do so. SKOS[15], for example, only offers two properties to model sense restriction and sense enlargement respectively: `narrower` and `broader`. In default of a more detailed range of properties we modeled the sense relations using the information contained in the XML data: 'par métaph.', etc. We implemented a template that automatically reads this information and transforms it into the respective RDF data using the OntoLex-Lemon property `usage` and a link to DBPedia. In the following we present an extract of this template:

```
<xsl:template name="usage_extern">
<xsl:choose>
<xsl:when test="./description/m:usage/
  @type='metaphor' or
  ./description/m:idem/m:usage/
  @type='metaphor'">
  ontolex:usage dbpedia:Metaphor ,
  "<xsl:value-of select="./description/
  m:usage"/>
  <xsl:value-of select="./description/
  m:idem/m:usage"/>"@fr ;
</xsl:when>

<xsl:when test="./description/m:usage/
  @type='irony' or
  ./description/m:idem/m:usage/
  @type='irony'">
  ontolex:usage dbpedia:Irony ,
  "<xsl:value-of select="./description/
  m:usage"/>
  <xsl:value-of select="./description/
  m:idem/m:usage"/>"@fr ;
</xsl:when>
</xsl:template>
```

An example of the outcome is:

```
:fiel_sense1.d a ontolex:LexicalSense ;
  ontolex:usage dbpedia:Metaphor ,
  "métaph."@fr .
```

[15] `http://www.w3.org/TR/skos-reference/#semantic-relations` [accessed 12-12-2017].

Aside from addressing the aforementioned shortcomings of established community standards, one direction of future research is to improve the linking with other lexical resources. We have to note, however, that the philological perspective entails that first-class citizens for such a linking would be dictionaries of historically or linguistically related language varieties. Such a linking requires also historical resources to become increasingly available within the LLOD cloud. Our own research represents a step in this direction, and, by demonstrating the feasibility, we hope to encourage others to work in this direction as well. In particular, we expect similar challenges to arise on other datasets from historical philologies, so that in the immediate future, a focus should be laid on developing rules of best practice and specifications for this particular community before we can expect a greater degree of convergence.

A linking with language resources for modern varieties, on the other hand, would be technologically more feasible, but the theoretical and philological implications of such a linking requires a theoretical reflection in order to avoid mislinkings and incorrect interpretations arising from the Historical Semantic Gap.

We intend to publish the converted dictionary under an open license. However, we have to admit that legal clearance is still underway. Unfortunately, this situation is symptomatic for many valuable resources in the historical philologies, which are characterized by massive collaboration, long-term projects, often involving several institutions and complicated publication agreements for the underlying print edition.

## 6. Acknowledgements

Sabine Tittel is a full time redactor of the dictionary DEAF, Heidelberg Academy of Sciences and Humanities. The contribution of the second author was supported by the project "Linked Open Dictionaries" (LiODi), an Early Carreer Research Group funded by the eHumanities programme of the German Federal Ministry for Education and Research (BMBF).

The OntoLex-lemon edition of the data was supported by the organizers and participants of the 2nd Summer Datathon on Linguistic Linked Open Data (SD-LLOD 2017), June 2017, Cercedilla, Spain, where the linking of DEAF data was explored in a collaborative effort. This paper elaborates and builds on these experiments. In particular, we would like to thank Yifat Ben-Moshe (K Dictionaries, Tel Aviv), Helena Bermúdez-Sabel (Universidad Nacional de Educación a Distancia, Madrid), Mariana Curado Malta (Polytechnic Unicersity of Porto, Portugal), Frances Gillis-Webber (University of Cape Town) and Maxim Ionov (Goethe-University Frankfurt, Germany) for their input and contributions.

Furthermore, we would like to thank the anonymous reviewers for helpful comments and insightful feedback.

## 7. Bibliographical References

Baldinger, K. (since 1971). *Dictionnaire étymologique de l'ancien français – DEAF*. Presses de L'Université Laval / Niemeyer / De Gruyter, Québec, Canada /

Tübingen / Berlin, Germany. [Kurt Baldinger (founder), continued by Frankwalt Möhren, published under the direction of Thomas Städtler; electronic version DEAF*él*: `https://deaf-server.adw.uni-heidelberg.de`].

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

Borin, L., Dannells, D., Forsberg, M., and McCrae, J. (2014). Representing Swedish Lexical Resources in RDF with Lemon. In Matthew Horridge, et al., editors, *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, pages 329–332, Aachen, Germany. CEUR-WS.org.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado de Cea, G. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries Case. In Ilan Kernerman, et al., editors, *Proceedings of GLOBALEX'16 Workshop at LREC'15, Portoroz, Slovenia*, pages 65–72, Aachen, Germany. European Language Resources Association.

Bouda, P. and Cysouw, M. (2012). Treating Dictionaries as a Linked-Data Corpus. In Christian Chiarcos, editor, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 15–23. Springer, Berlin / Heidelberg, Germany.

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Chiarcos, C. and Sukhareva, M. (2014). Linking Etymological Databases. A Case Study in Germanic. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, page 41.

Chiarcos, C. and Sukhareva, M. (2015). OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 518:379–386.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards Open Data for Linguistics: Lexical Linked Data. In Alessandro Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, pages 7–25. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

Declerck, T., Buitelaar, P., Wunner, T., McCrae, J., Montiel-Ponsoda, E., and de Cea, A. (2010). lemon: An ontology-lexicon model for the Multilingual Semantic Web. In *W3C Workshop: The Multilingual Web - Where Are We?*, Madrid, Spain, Oct.

Declerck, T., Wandl-Vogt, E., and Mörth, K. (2015). Towards a Pan European Lexicography by Means of Linked (Open) Data. In Iztok Kosem, et al., editors, *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 Conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, pages 342–355, Ljubljana/Brighton, 8. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical Markup Framework (LMF). In *5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 233–236, Genoa, Italy.

Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. (2018). The Apertium Bilingual Dictionaries on the Web of Data. *SWJ (Semantic Web Journal)*, 9(2):1–10.

Khan, F., Bellandi, A., Boschetti, F., and Monachini, M. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using OntoLex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017)*, pages 43–50, Galway, Ireland.

Klimek, B. and Brümmer, M. (2015). Enhancing Lexicography with Semantic Language Databases. *Kernerman DICTIONARY News*, 23:5–10.

Klyne, G., Carroll, J., and McBride, B. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C Recommendation.

Mazziotta, N. (2010). Building the Syntactic Reference Corpus of Medieval French using NotaBene RDF annotation tool. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW-2010)*, pages 142–146, Uppsala, Sweden, August. Association for Computational Linguistics.

Miles, A. and Bechhofer, S. (2009). SKOS Simple Knowledge Organization System reference. W3C Recommendation. Technical report, World Wide Web Consortium.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November.

Tittel, S., Bermúdez-Sabel, H., and Chiarcos, C. (accepted). Using RDFa to link text and dictionary data for Medieval French. In *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2018)*, Miyazaki, Japan, May.

Wang, S., Schlobach, S., and Klein, M. (2011). Concept drift and how to identify it. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):247–265.

Witte, R., Kappler, T., Krestel, R., and Lockemann, P. C. (2011). Integrating Wiki Systems, Natural Language Processing, and Semantic Technologies for Cultural Heritage Data Management. In Caroline Sporleder, et al., editors, *Language Technology for Cultural Heritage*, pages 213–230. Springer, Berlin / Heidelberg, Germany.

# Integrating Prepositions in Wordnets: Relations, Glosses and Visual Description

**Raquel Amaro**
NOVA CLUNL – Faculty of Social Sciences and Humanities – Universidade NOVA de Lisboa
& CLUL – Faculty of Arts – Universidade de Lisboa
Avenida de Berna 26-C 1069-061 Lisbon, Alameda da Universidade 1600-214 Lisbon
raquelamaro@fcsh.unl.pt

**Abstract**

Lexicology and lexicon models are necessarily concerned with content words, being grammatical and functional categories often set aside. Currently, however, lexicographers work for real needs and, in a NLP perspective, the nature of computational lexicons reflects a necessary match between what we know about the mental lexicon and what we need to encode about the set of words of a given language. Prepositions, in many languages, combine these two prototypes of words – lexical and functional –, as they can have full meaning or serve solely as structural aids. Based on the analysis of Portuguese prepositions related to the expression of movement, this paper describes how the integration of prepositions in wordnets is possible and quite easy, requiring mainly the linguistic adaptation of the tests and conditions that mediate the establishment of the relations of synonymy, antonymy, hyperonymy and cause. In what concerns lexicographic strategies, the integration of prepositions shows the difficulty of establishing equivalences between the concepts denoted by prepositions in different languages, as well as the difficulty of using glosses in natural language to describe their meaning. The use of visual information may obviate this issue, while posing issues on implementation.

**Keywords:** prepositions, wordnets, visual information

## 1. Introduction

One of the first appointed differences between the theoretical study of the lexicon – lexicology – and the crafts of making lexical resources – lexicography – is the set of words that is considered relevant for the first and that has necessarily to be described for the second (Crystal, 1995). Lexicology and models of the mental lexicon are essentially concerned with so-called content words, being grammatical or functional categories often set aside (Klein, 2001). Currently, however, lexicography works considering real users' needs, and often focuses its strategies for NLP purposes (Gouws, 2004). The nature of modern computational lexicons can thus be described as the perfect or necessary match between what we know and figure about the mental lexicon, considering conceptual and semantic properties, and what we need to encode about *all* the words of a given language in order to make a lexicon useful, whether this information is functional or not.

Prepositions, in many languages, perfectly combine these two prototypes of words, as they aggregate items that undoubtedly have meaning and items that serve solely as structure markers (Hernández-Pastor & Periñán-Pascual, 2016). In fact, accounting for prepositions in wordnets is listed as one of the existing challenges for the development and application of wordnets (Workshop Challenges for Wordnets, Bond & Piasecki, 2017) and prepositions are included at least in Bulgarian WordNet[1] (Dimitrova et al., 2014), although not establishing many (if any) relations with other nodes in the net (test, for instance, след in http://dcl.bas.bg/bulnet/).

Based on the analysis of Portuguese prepositions related to the expression of movement, this paper further explores the integration of prepositions in wordnets, showing how they can be modeled, which relations serve to encode their meaning and/or function, and how glosses and crosslinguistic equivalences can be inadequate to provide a clear grasp of the concept prepositions denote.

In the next sections, we review different approaches to prepositions, as well as further explore the motivations for integrating them in wordnets (section 2); we present our proposal for modeling prepositional concepts in wordnets, considering semantically full prepositions and argument-marking prepositions (section 3); we discuss the issues concerning semantic description, crosslinguistic equivalence, glosses and visual information (section 4); and lastly we present our final remarks (section 5).

[1] http://dcl.bas.bg/en/resursi/wordnet/

## 2. Prepositions

Prepositions are fairly common in natural languages, and their treatment is of high impact in NLP tasks (Hernández-Pastor & Periñán-Pascual, 2016).

The analysis of prepositions has many times been considered under the scope of the relation between prepositions and the nouns they co-occur with (*on Thursday*, *in the morning*), or the verbs that select them (*dream of*, *care about*) (Veerspoor, 1997), directly related to cases where the semantic contribution of prepositions to the meaning of the phrase or sentence seems or is void. In fact, this aspect of the combination of prepositions with other lexical items is what usually makes them difficult to be computationally processed and in many cases disambiguated (*Ele sonhou com a irmã*. = he dreamt of his sister; *Ele morou com a irmã*. = he lived with his sister).

However, many prepositions display a constant semantic content, which is crucial for the determination of the meaning of prepositional phrases and sentences (*since February* vs. *until February*; *at home* vs. *from home*) (Bannard & Baldwin, 2003).

In what concerns their semantic description, research on prepositions has taken three main directions:

i) large-scale symbolic accounts of preposition semantics (Dorr, 1997's 497 senses of English transitive and intransitive prepositions formalized in a lexical conceptual semantics framework; Canesson & Saint-Dizier, 2002's description of French prepositions in PrepNet; Jensen & Nilsson, 2003's description of prepositions through a finite set of universal binary role relations; Srikumar & Roth, 2013's set of relations established by prepositions);

ii) prepositional phrase disambiguation (O'Hara & Wiebe, 2003's account of prepositional phrases tokens according to case-roles, or McShane et al., 2005's

ontological semantic analyzer for disambiguating homonym prepositions); and

iii) distributional accounts of preposition semantics (such as Bannard & Baldwin, 2003's work on particles and transitive prepositions for a valence-conditioned classification of English prepositions).

In many of these cases, as well as in more conventional approaches such as traditional normative grammars, semantically full prepositions are commonly organized according to notions such as purpose, goal, location, temporality, cause, etc., across languages such as French (Saint-Dizier, 2008), English (Jensen & Nilsson, 2003) or Portuguese (Cunha & Cintra, 1984). According to these works, the semantic value of prepositions can be compared to those of other POS.

### 2.1 Related work

Several researchers have studied prepositions and the ontological organization of prepositions, adopting a similar approach to that of WordNet, given that prepositions are described according to their conceptual properties.

PrepNet (Saint-Dizier, 2005, 2008) is such an example. PrepNet is a database for prepositions structured in two levels: the abstract notion level (conceptual level, language independent) and the language realization level (which deals with the realizations for various languages). Abstract notions are organized in a first stage that characterizes the semantic family of the notions (localization, manner, quantity, company, etc.), a second stage that accounts for the different facets of each semantic family (source, destination, or via, for instance), and a third stage that captures the modalities of a given facet (such as basic manner, manner by comparison, manner with a reference point, etc.). The language representation level includes syntactic frames and semantic and domain restrictions.

PrepNet approach to the representation of the meaning of prepositions can be used as the base for integrating prepositions in wordnets, since the abstract notion can help in the establishment of prepositional higher nodes in wordnets as well as in the establishment of the sets of hyponyms. However, we observed that the facets and modalities expressed by prepositions are not necessarily the same in every language.

As mentioned before, accounting for prepositions in wordnets is listed as one of the existing challenges for the development and application of wordnets (Workshop Challenges for Wordnets, Bond & Piasecki, 2017). Nevertheless, in current days, not many of these lexical resources include prepositions. In fact, a survey of the information displayed on the presentation pages of each of the wordnets included in the Global WordNet Association list of wordnets in the world[2] show us that, from the 124 resources listed, 30 do not state the POS considered (from which we assume they encode the same POS treated in Princeton WordNet: nouns, verbs, adjectives and adverbs); 42 state they do not consider prepositions; 28 do not present webpages for the resources; 21 do not have functional webpages or are in maintenance and 2 do not provide information in English. Only the Bulgarian WordNet[3] (Dimitrova et al., 2014) states the inclusion of prepositions (as well as other functional words such as conjunctions), although these seem to be somewhat loose in the net, according to the observation of the nodes for some prepositions in http://dcl.bas.bg/bulnet/.

Following Amaro (2009), the motivation for integrating prepositions in wordnets comprises two sets of reasons:

i) theoretical (semantic) reasons: prepositions denote notions such as cause, location, temporality, etc., as demonstrated by several earlier and current studies;

ii) practical (functional) reasons: even semantically empty prepositions, which are idiomatic, add information useful for NLP purposes, contributing to the usability and relevance of wordnets.

The following sections illustrate further these aspects.

### 2.2 Dataset

The set of prepositions considered in this paper was compiled from prepositions commonly used in the expression of movement in Portuguese (Amaro, 2009), such as *de* (≈ from), *a* (≈ to), *até* (≈ until/to), *para* (≈ to, in the direction of, towards), *por* (≈ through), *em* (≈ in), *sobre* (≈ on top of, over), *entre* (≈ between), etc.

We also considered multiword expressions such as *acima de* (≈ above), *atrás de* (≈ behind), *ao lado de* (≈ next to, close to), *por baixo de* (≈ under), *em direção a* (≈ to, towards, in the direction of), and so on, since these fixed expressions behave like prepositions (see Cunha & Cintra, 1984; Baldwin et al., 2009). These correspond to multiword expressions that refer to prepositional meaning or have a prepositional function and are expressions that

i) do not undergo inflection, internal modification or word order variation, i.e. "words with spaces" (Sag et al., 2002):

(1) a. Ele colocou o livro mesmo ao lado da jarra.
he placed the book exactly at.the side of the vase (≈ next to)

b. *Ele colocou o livro mesmo aos lados da jarra.
he placed the book exactly at.the sides of the vase

c. *Ele colocou o livro ao lado mesmo da jarra.
he placed the book at.the exactly side of the vase

d. *Ele colocou o livro ao lado esquerdo da jarra.
he placed the book at.the left side of the vase

e. *Ele colocou o livro mesmo do lado à jarra.
he placed the book exactly of.the side at.the vase

ii) can often be replaced by simple prepositions, as illustrated in (2):

(2) a. The mouse ran in the direction of/to the table.

b. The man stood quiet in front of/before the judges.

## 3. Modeling prepositions in WordNet

### 3.1 Semantically full prepositions

Diverging from the approaches for modeling the semantics of prepositions in a deeper fashion and with specific sets of relations (Saint-Dizier, 2008; Srikumar & Roth, 2013, Schneider et al., 2015), we demonstrate that it is possible to model prepositions with full meaning (i.e., prepositions whose semantic content is crucial for the determination of the meaning of phrases, such as *before noon* vs. *after noon*) through relations already available in WordNet model, namely synonymy, antonymy, hyperonymy/hyponymy and cause/is caused by.

[2] http://globalwordnet.org/wordnets-in-the-world/
[3] http://dcl.bas.bg/en/resursi/wordnet/

These relations correspond to the ones defined in Fellbaum (1998) and Vossen (2002) and require only the adaptation of the tests and definitions to the specificity of this POS. Specifically, prepositions require a complement (usually a Noun Phrase) and cannot be linguistically tested without considering the entire Prepositional Phrase. Although based on the studied prepositions for Portuguese, the definitions and tests presented here are expected to serve for any language. For that reason, whenever possible, English examples will be used to illustrate the tests.
The adapted definitions and tests are presented below[4].

(3) Synonymy relation

Definition:
*P1 is synonym of P2 in C iff*
*if P1 then P2 and if P2 then P1*

Test:
if the mouse is *under* the table then the mouse is *underneath* the table, and if the mouse is *underneath* the table then the book is *under* the table: **True**

| | | |
|---|---|---|
| *under* | is synonym of | *underneath* |
| *underneath* | is synonym of | *under* |

--> {under, underneath}$_{Prep}$

Synonymy relations between prepositions, in Portuguese at least, are not very productive, even considering the synonymy notion bound to a given context. However, they still exist, in particular between atomic and multiword prepositions.
Prepositional synsets can also be related to each other by antonymy.

(4) Antonymy relation

Definition:
*P1 is antonym of P2 iff*
*i) P1 and P2 are co-hyponyms;*
*ii) P1+NPi/VPi is the opposite of P2+NPi/VPi and P2+NPi/VPi is the opposite of P1+NPi/VPi*

Test 1:
i) *under* and *on top of* are both hyponyms of *in, at*: **True**
ii) *under* the table is the opposite of *on top of* the table and *on top of* the table is the opposite of *under* the table: **True**

Test 2 (negation):
*if P1+NPi/VPi then not P2+NPi/VPi and*
*if P2+NPi/VPi then not P1+NPi/VPi*
if the cat is *under* the table, then the cat is not *on top of* the table: **True**
if the cat is *on top of* the table, then the cat is not *under* the table: **True**

| | | |
|---|---|---|
| *under* | is antonym of | *on top of* |
| *on top of* | is antonym of | *under* |

[4] The notations used in the definitions and tests correspond to: P = Preposition; NP = Noun Phrase; VP = Verb Phrase; AdjP = Adjectival Phrase; { } = synset/node in the net; / = or. The index *i* assures that the complements considered for P1 and P2 are the same.

Antonymy relations between prepositions, as it happens with adjectives (cf. Mendes, 2009), are quite relevant for further modeling prepositional concepts given that they allow to express opposite facets of several notions such as opposite locations with regard to a given ground object (ex.: *under* vs. *on top of*; *to inside of* vs. *to outside of* (see Figures 1 and 2)), opposite directions (ex.: *upwards* vs. *downwards*, *to* vs. *from*), opposite temporal relations (*after* vs. *before*), etc.

(5) Hyponymy/hyperonymy relation

Definition:
*P2 is hyponym of P1 and P1 is hyperonym of P2 iff*
i) *P2 is P1+NPi/VPi/AdjPi, but*
ii) *P1 is not P2+NPi/VPi/AdjPi*

Test 1:
*under* is *in*+the space below, but *in* is not *under*+the space below: **True**

| | | |
|---|---|---|
| {*under*}$_{Prep}$ | is hyponym of | {*in, at*}$_{Prep}$ |
| {*in, at*}$_{Prep}$ | is hyperonym of | {*under*}$_{Prep}$ |

Test 2 (conditions for replacement and anaphora):
*P2 is hyponym of P1; and*
*i) the complement of P1 denotes a reference that is equal or includes the reference denoted by the complement of P2;*
*ii) if P2 then P1, but if P2 then not P1*
*iii) P1 can be used as anaphoric element for P2.*

*under* is hyponym of *in*: **True**
the room includes the table: **True**
If the mouse is *under* the table, then the mouse is *in* the room: **True**
If the mouse is *in* the room, then the mouse is *under* the table: **False**
*The mouse is under the table. So, while it was in the room, nobody entered.*
*#The mouse was in the room. So, while it was under the table, nobody entered.*

The testing for hyponymy/hyperonymy relations requires considering the inclusion relations established between the prepositional complements, following the described in Vossen (2002: 21) for hyponymy relations between nouns.
The definitions and tests proposed here show the feasibility of modeling prepositional concepts in wordnets, with some level of meaning description. Figures 1, 2, 3 and 4 present examples of hyponymy nets for Portuguese prepositions related to the expression of movement and spatial relations.
The study of Portuguese prepositions related to the expression of movement also allowed us to observe that, although seeming quite similar to prepositions indicating location, and almost seeming compositionally built, prepositional expressions denoting goal and source locations (Figures 2 and 3) do not result from the combination of prepositions denoting location, in Figure 1.
First, if these expressions were regular and compositional, the occurrence of not allowed combinations would be minimal and accidental. However, on the contrary, it is not possible to express a source or goal location using the

prepositions *de* or *para* + *em* (the top nodes of the three subtrees presented):

(6) *Ele foi de em a escola para em a rua.
he went from in the school to in the street

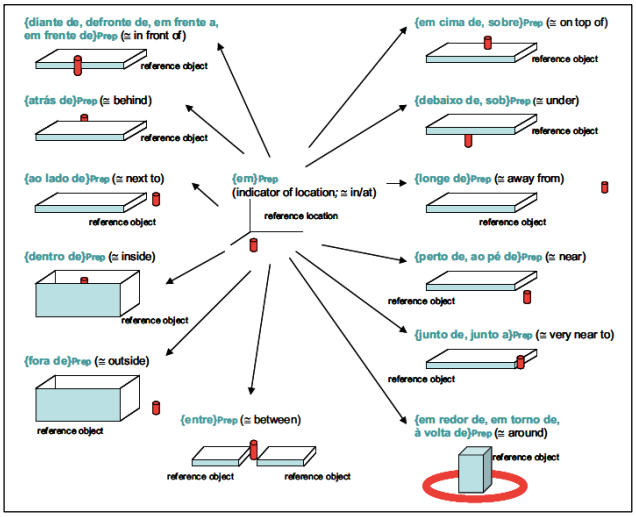



Figure 1: Hyponymy network of prepositional synsets denoting indicators of location

A closer view also reveals that several combinations of elements from the subnets presented are not possible:

(7) a. *para/de em cima de (≈ to/from on top of)
b. *para/de em baixo de (≈ to/from on under of)
c. *para/de em frente a (≈ to/from in front of)
d. para trás de/*de trás de/*em trás de (≈ to behind/from behind/in behind)
e. para debaixo de/de debaixo de/*em debaixo de (≈ to under/from under/under)
f. em torno de/*para torno de/*de torno de (≈ in around of/to around of/from around of)

However, intuitively, the concepts of location, source location and goal location seem to be strongly related. This is the case given that moving to a final location (goal) causes being in that location, and, on the contrary, moving from a given location (source) causes not being in that location. Being so, it is possible to link these concepts in wordnets through cause relations.
The definition and testing of cause/is caused by relation between prepositional nodes is presented below, as well as their application to the synsets {to}$_{Prep}$ (indicator of goal) and {from}$_{Prep}$ (indicator of source) and {in, at}$_{Prep}$ (indicator of location), for explanatory purposes.

(8) Cause/is caused by relation

Definition:
*P1 causes P2 iff*
*P1+Ni causes/has as consequence P2+Ni, but not the converse.*

Test:
a. (He moved) *to* the street causes/has as consequence (he is) *in* the street but (he is) *in* the street does not cause/have as consequence (he moved) *to* the street

{to}$_{Prep}$ causes {in, at}$_{Prep}$
{in, at}$_{Prep}$ is caused by {to}$_{Prep}$ (non-factive)

b. (He moved) *from* the street causes/has as consequence (he is) not *in* the street but (he is) not *in* the street does not cause/have as consequence (he moved) *from* the street

{from}$_{Prep}$ causes {in, at}$_{Prep}$ (negative)
{in, at}$_{Prep}$ is caused by {to}$_{Prep}$(negative) (non-factive)

In order to test the cause relation between a prepositional synset indicator of source location and another prepositional synset indicator of location, in (8)b, it is necessary to include negation, since the consequent state of moving from a given location amounts to not being in that location.
In WordNet, the negation label is used to explicitly express that a given relation does not hold. It is used to block unwanted implications, as non-inherited relations (Vossen 2002:16). The case presented here does not correspond exactly to the same situation, given that there is no prototypical relation to be inherited. The negation label is only used here for explanatory purposes.[5]

## 3.2 Argument-marking prepositions

One of the main reasons leading to the little attention dedicated to prepositions when it comes to their semantic content is directly related to semantically empty prepositions, that is, prepositions serving only functional or grammatical purposes. This set can be further divided in i) functional prepositions, i.e. prepositions that regularly indicate syntactic functions that do not depend on selection restrictions of specific lexical items (as, for instance, the preposition *a* in Portuguese, which regularly and invariably marks the indirect object of ditransitive verbs); and ii) argument-marking prepositions, i.e. prepositions whose only function is to mediate between a given predicate and its arguments (Sag & Wasow 1999: 157), as illustrated below for Portuguese and English:

(9) a. O rapaz gostou de cães.
the boy liked PREP dogs
b. O rapaz sonhou com cães.
the boy dreamt of dogs
c. O rapaz aproximou-se dos cães.
the boy came closer to.the dogs.

In what regards the integration of empty prepositions in wordnets, we propose that it is relevant to consider the second case since these prepositions, as illustrated in (9), concern:
i) cases in which the presence of the preposition is language dependent (9a);

[5] The relation established between {from}$_{Prep}$ and {in, at}$_{Prep}$ is that the first causes the negation of the last, and not that the relation between the nodes does not hold. For this reason, it is only possible to express this relation indirectly, linking *from* and *to* as antonyms, which motivates further the relevance of antonymy relation.

ii) cases in which the preposition choice does not correspond to the typical equivalent in other languages (9b, where the Portuguese preposition *com* corresponds to the English preposition *of*, instead of its frequent English translation *with*); and
iii) cases where the argument-marking preposition is homonym of the preposition denoting the opposite semantic content (in 9c, where the argument marking preposition *de* marks a goal location argument, whereas the semantically full preposition *de* denotes an indicator of source location).

On the contrary, truly functional prepositions can be effectively covered by syntactic rules, justifying their absence from the lexicon.

Being idiosyncratic, i.e. language dependent and not permutable by any other preposition, argument-marking prepositions are said to form a semantic component with the verb, since it is the verb+preposition that attributes case to the selected NP (see Neeleman, 1997).

Neeleman proposal results in complex lexical entries for verbs such as *gostar de* (≈ like), *sonhar com* (≈ dream of) and *aproximar-se de* (≈ go closer), for instance, and could motivate their encoding within the node for the verb form.

However, and as underlined by Godoy (2008), at syntactic level these prepositions form constituents with the selected NP and not with the verb, as illustrated in (10), (11) and (12).

(10) a. De cães, o rapaz gosta.
≈ PREP dogs, the boy likes
b. Com cães, o rapaz sonhou.
≈ of dogs, the boy dreamt
c. Dos cães, o rapaz aproximou-se.
≈ to the dogs, the boy moved closer

(11) a. O rapaz gosta de cães e ela também gosta.
≈ the boy likes PREP dogs and so likes she
b. O rapaz sonhou com cães e ela também sonhou.
≈ the boy dreamt of dogs and so dreamt she
c. O rapaz aproximou-se dos cães e ela também se aproximou.
≈ the boy moved closer to the dogs and so moved she

(12) a. O rapaz gosta de cães e de gatos.
≈ the boy likes PREP dogs and PREP cats
b. O rapaz sonhou com cães e com gatos.
≈ the boy dreamt of dogs and of cats
c. O rapaz aproximou-se dos cães e dos gatos.
≈ the boy moved closer to the dogs and to the cats

These examples show that, although required by a given verb, argument-marking prepositions do not form semantic or syntactic components with the verb that subcategorize for them: on the one hand, having no semantic content, these prepositions do not contribute to the semantic content denoted by the VP; on the other, they form syntactic constituents with the NP and not with the verb. Following Godoy's (2008) approach, we consider that these prepositions are not visible at semantic level, existing solely at syntactic level.

Argument-marking prepositions are true grammatical words and semantically empty lexical items, directly related to verbs that subcategorize them, raising the issue of how to represent these items in wordnets, since these prepositions do not denote concepts. Their inclusion in the lexicon, however, can be motivated by different reasons:

i) as idiosyncratic items, these prepositions are acquired by children in a similar process as all other lexical items, since their distribution and/or meaning do not result from the regular application of rules available in natural languages (cf. Godoy, 2008);
ii) argument-marking prepositions constitute a small and closed set of items, necessarily connected to the verbs that require their syntactic realization. So, the collection and treatment of argument-marking prepositions is always related to the collection and treatment of verbs.
iii) their representation as autonomous entries (instead of as part of verbal entries) allows for multiple linking, and avoids multiword expressions that not conform to the properties defined earlier (not undergoing inflection, internal modification or word order variation (as illustrated in (10), (11) and (12)).

These reasons, although strongly of lexicographic nature, motivate the inclusion of these items in wordnets as part of the set of prepositional items, but as extremely underspecified lexical entries. These can be related to other nodes in the net either using the conjunction label with role and involved relations (Vossen, 2012), either using specific selection relations, as proposed in Amaro (2010).

### 3.3 Informational gain

The cases presented clearly exemplify how the integration of prepositions in wordnets is possible, using mainly available relations with the necessary adaptations to definitions and testing conditions.

In terms of informational gains for these resources, the integration of prepositions allows, for instance, for a more complete description of the lexical items and of properties of POS, such as subcategorization properties of verbs, but also of the computational processes of the lexicon.

For instance, the integration of semantically full prepositions enables the model to represent in a more accurate way the expression of location. This is visible in two specific possibilities:

i) the automatic prediction of which specific lexical units can introduce location, source, goal, etc., considering the percolation of information in the net: if hyponyms inherit their hyperonyms properties, a given argument of a verb can be introduced by the indicated prepositional node or by any of its hyponyms:

(14) He put the books in / under/behind/inside the closet.

ii) the accurate expression of arguments considering the compositionality of PPs: preposition meaning+ complement meaning. For instance, the integration of prepositions makes it possible to encode, through the extension of involve relations, that *put* selects for an argument introduced by a preposition denoting an indicator of location (cf. 15), which is not expressible by the involved location relation as defined in Vossen (2002: 31), which requires a nominal synset (cf. 16).

(15) a. {put}$_{V}$ involve_location {in, at}$_{Prep}$
b. He put the books in the closet.

(16) a. ?{put}$_V$ involved_location {location}$_N$
b. #He put the books in the location.

The specific realization of the argument is naturally conditioned by the semantic properties of the elements in the predicate, corresponding in this case to the physical objects denoted by the direct object of the verb (*the book,* in (15b)) and by the complement of the preposition, in this case *the closet*.
This explains why sentences such as *He put the book inside the table* may be odd, or at least require the assumption that the table in question has an interior compartment, whereas sentences such as *John put the book inside the closet* may seem slightly redundant (as opposed to *John put the book in the closet*), since the container aspect of *closet* constitutes one of its defining semantic properties.
Finally, the integration of semantically full prepositions allows for encoding more accurately the semantic restrictions on argument selection, and thus semantic features, of verbs (Amaro 2009, Amaro et al. 2013). As stated above, *put* can be described as selecting for an argument of the type location (see 16). However, as illustrated below, this semantic type is most of the times built from the semantics of the preposition used: *table, closet, fridge* are hardly thought of as locations, or represented as hyponyms of *location*, but result in well-formed sentences when arguments of the verb *put* introduced by a preposition indicator of location (cf. (17)).

(17) John put the bottle in the table/closet/fridge/window.

For these reasons, the integration of prepositions in wordnets constitutes a relevant informational gain for the model and for the lexicons described.

## 4. Lexicographic issues: glosses, crosslinguistic equivalences and visual description

Focusing on synonymy and hyponymy relations, we modeled some subsets of Portuguese prepositions directly related to the expression of movement. These subnets concern the expression of location (Figure 1), goal location (Figure 2), source location (Figure 3) and path (Figure 4).
These subnets reveal some underlying issues concerning the description of prepositional meaning, illustrating several strategies to account for them.

### 4.1 Glosses: using natural language to describe prepositional meaning

The first issue requiring further reflection concerns the use of natural language to describe the meaning of prepositions, starting with the description of the initial node for each subnet.
Considering the network depicted in Figure 1, the Portuguese preposition *em* is the top node for this subnet, roughly corresponding to the English prepositions *in/at*. This preposition denotes the more general and underspecified concept of indicator of location, with regard to a reference location, which is then specified by its hyponyms. But glossing the concept denoted by this preposition as "indicator" is not a coincidence.

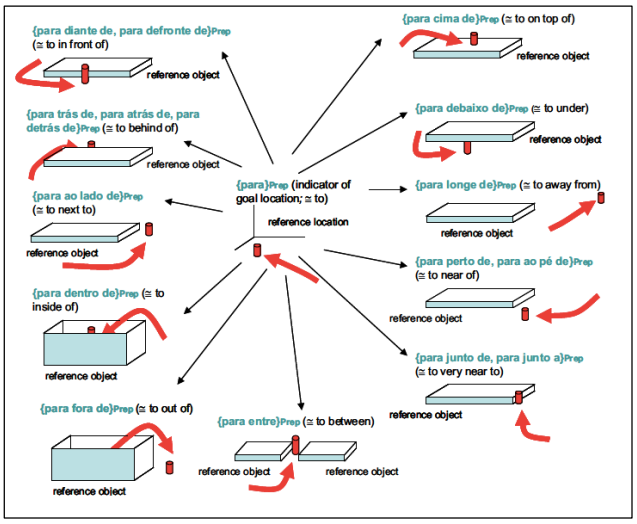



Figure 2: Hyponymy network of prepositional synsets denoting indicators of goal location

From traditional grammars (Cunha & Cintra, 1984) to current linguistic models (Saint-Dizier, 2008; Srikumar & Roth, 2013, Schneider et al., 2015), prepositions are described as items that connect other elements in a sentence. Jensen & Nilsson (2003), for instance, propose a finite set of universal binary role relations to describe the semantic content of prepositions. In their perspective, prepositions denote a relation between the concept denoted by a given lexical item and semantic roles considered in a given ontology. In other words, prepositions can be described as *indicators* of concepts relating to space, temporality, causality, and so on. These ontological analyses can provide us with the top concepts susceptible to be lexicalized by prepositions, but also with an initial proto-hyperonym from which to draw our initial glosses, i.e. the notion or concept of "indicator". Accordingly, we can gloss prepositions as indicators of location, of time, of cause, etc.

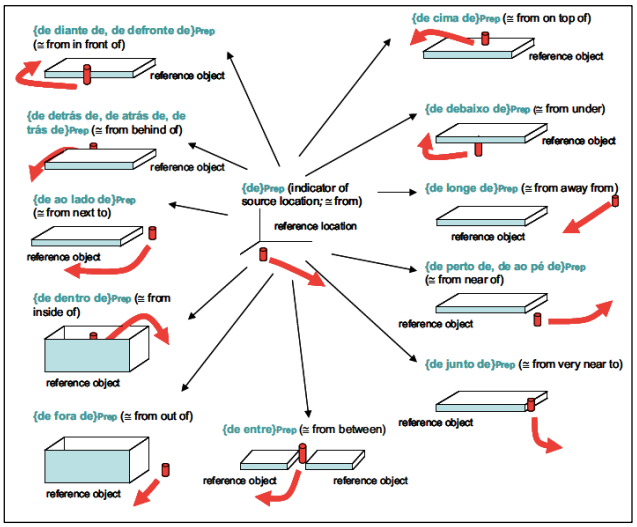



Figure 3: Hyponymy network of prepositional synsets denoting indicators of source location

Building glosses for hyponym prepositional nodes is yet another issue. It is not easy to gloss prepositional concepts

without resorting to the lexical items we intend to describe. For instance, *entre* (≈ between) can be glossed, more or less artificially, as "*em* (in/at) the space that separates objects". However, prepositional expressions such as *debaixo de* (≈ under), *em cima de* (≈ on top of), *ao lado de* (≈ next to), *atrás de* (≈ behind), etc., are not as easily glossed.

Although not as straightforwardly as for other POS, we can gloss the meaning of prepositions using two main strategies. Consider, for instance, the synset {*para fora de*}$_{Prep}$ (≈ to outside of), hyponym of {*para*}$_{Prep}$ (≈ to; indicator of goal location). We can build its gloss using:

i) the hyperonym lexical item + NP/VP/Adj concerning the hyponym specific properties (Aristotelian formula). Example: {*para fora de*}$_{Prep}$ (≈ to outside of) gloss: *para* + uma localização exterior a (*to* + a location exterior to);

ii) the proto-concept of "indicator", providing the specific notion or relation at stake. Example: {*para fora de*}$_{Prep}$ (≈ to outside of) gloss: indicador de localização final exterior ao objeto ou localização de referência (indicator of final location exterior to the reference object or location).

Both strategies have pros and cons:

i) the first strategy results in regular and direct glosses, although somewhat artificial, that allow the direct replacement of the glossed lexical units. Example: *Ele foi para fora da sala. --> Ele foi para uma localização exterior à sala*) (≈ He went to outside of the room --> He went to a location outside of the room);

ii) the second strategy results in more informational descriptions that help to understand more complex concepts, for instance in more abstract cases such as in *He cried in anger*; *The offer was received with fear*.

The decision for one or the other of the strategies must respect the goals and purpose of the resource and its target audience.

Nonetheless, the construction of glosses is directly related to the second issue to be accounted in wordnet model when considering prepositions, namely how to establish crosslinguistic equivalences for prepositional nodes and if these are accurate and feasible using glosses alone.

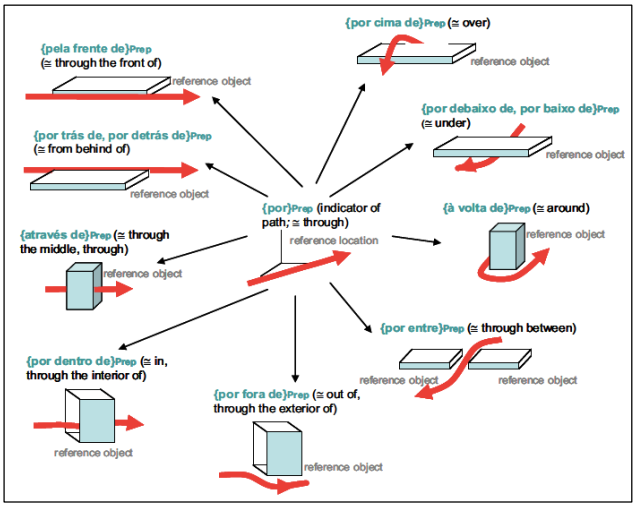



Figure 4: Hyponymy network of prepositional synsets denoting indicators of path

## 4.2 Crosslinguistic equivalence for prepositions and visual information

As mentioned in previous sections, several authors have studied prepositions and prepositional meaning, departing from different languages, and the concepts denoted can be fairly commonly grouped under notions of temporality, space, cause, etc., organized in several ways. For explanation purposes, Figures 5 and 6 present different proposals concerning different approaches and languages.

AGENT _____ animate being acting intentionally (ex: *treatment **by** physician*)
CAUSE _____ inanimate force/actor
CAUSED_BY ____ inversed CAUSE
PATIENT _____ affected entity/effected entity (ex: *treatment **of** children.*)
PART_OF _____ part of whole/member of (ex: *side **of** the head, cells **in** the eye, agent **from** the CIA*)
COMPRISE ____ inverse PART_OF; whole constitued of parts
BY MEANS OF __ means to end/instrument (ex: *treatment **with** medicine*)
SOURCE _____ source, origin, point of departure (ex: *haemorrhege **from** the intestine*)
PURPOSE _____ purpose
LOCATION ____ place, position (ex: *inflammation **of** the eyes*)
TEMPORALITY __ temporal anchoring, duration, inception, etc. (ex: ***for** two days, **from** last year*)
MATERIAL _____ material (ex: *cushion **of** leather.*)
CHARACTERIZE _ property ascription (ex: *children **with** diabetes*)

Figure 5: Top ontology of prepositional role relations presented in Jensen & Nilsson (2003: 8) for English

- Localization:
  - source
  - destination
  - via/passage
  - fixed position
- Quantity
  - numerical or referential quantity
  - frequency and iterativity
  - proportion or ratio
- Manner
  - manners and attitudes
  - means (instrument or abstract)
  - imitation, agreement or analogy
- Accompaniment
  - adjunction
  - simultaneity of events
  - inclusion
  - exclusion
- Choice and exchange
  - exchange
  - choice or alternative
  - substitution
- Causality
  - cause
  - goal or consequence
  - intention
  - purpose
- Opposition
- Ordering
  - priority
  - subordination
  - hierarchy
  - ranking
  - degree of importance
- Instrument
- Other groups
  - theme
  - in spite of
  - comparison

Figure 6: Abstract notions and facets denoted by prepositions (Saint-Dizier 2008: 764-765) for French

These Figures illustrate sets of notions commonly related to the meaning of prepositions, evidencing that these can be more or less regular across languages. However, even in typologically close languages such as English and French, or Portuguese (as opposed to English and Guarani, for instance), establishing equivalences in prepositional meaning can be tricky. In our perspective, this happens for two main reasons:

i) speakers tend to actualize prepositional meaning considering the distributional properties of the prepositions (i.e., the meaning of the predicates with which semantically full prepositions can occur contributes to the actual definition of their meaning), and distributional properties are inherently language dependent;

ii) prepositions constitute a close class with highly polysemous items, even within the same semantic domain (observe, for instance, the English prepositions *over*, *under*, *through*, *in*, *to*, within the semantic field of movement (cf. Figures 1 to 4)).

The description of sense 1b (of the 17 listed) of the preposition *under* in the *American Heritage Dictionary of the English Language* clearly illustrates this:

(18) 1b. To or into a lower position or place than: *rolled the ball under the couch*.

The equivalence of this sense of *under* in Portuguese has to deal with differences in i) polysemy: *to* = *para* (goal location?) and *into* = *em* (goal position?: not possible in Portuguese); ii) distribution: *rolled the ball*, manner of motion verb (*roll*) non existent in Portuguese.
Also, if we add to this the issues concerning conceptual differences and the description of prepositional meaning through glosses, the potential for inaccuracy and confusion grows further. For instance, *perto de* corresponds to *near* or to *close*? Are these synonyms? And *junto de*? Does it denote a closer location (see Figure 1)? So, how to accurately describe prepositional meaning?
Considering the subset of prepositions studied (Portuguese prepositions concerning the expression of movement), the visual description, as illustrated in the Figures 1 to 4 above, seems to be an efficient strategy. In fact, and given the issues described above, several authors have used spatial models to describe the meaning of prepositions (Galton, 1993, 1997; Herzog, 1995; Asher & Sablayrolles, 1996; Lockwood et al., 2005, among others), thus further motivating our approach.
The visual description proposed uses static elements in the case of location (e.g. in Figure 1) and dynamic ones (arrows) in the cases where there is a component of movement associated to the meaning of the preposition (e.g. in Figure 4), as well as color to highlight the core elements of the descriptions:

- reference objects and locations are depicted in gray and soft colors, with deeper tones whenever 3-dimensional perspective is relevant;
- core objects and representations are depicted in red and bright color, and lines with initial or final arrows are used to represent movement and direction, whenever relevant.

Visual descriptions should be as flat and repetitive as possible, to avoid introducing additional elements and contributing to possible different interpretations.
The use of visual descriptions allows, thus, for straightforwardly representing the meaning of these prepositions[6], while it also illustrates the polysemy of prepositional items (in whatever languages are encoded or 'translated' in the net) and the existence of conceptual voids or gaps, given the fact that visual information is language independent. Naturally, this implies a more complex database able to cope with and display visual information, as well as user-friendly graphic editors for lexicographers.
Also, as less-intensively connected items in a model in which the relations established with the other nodes primarily represent the meaning of a unit, prepositions (as well as of other POS in similar conditions) can profit from the use of visual information for a more rich semantic description. Glosses can, thus, be used for adding useful information of a different nature, such as distributional information, for instance.

[6] The conception of visual descriptions for prepositions related to other notions (causality, manner) may pose specific challenges in itself, which although very interesting are out of the scope of this paper.

## 5. Final remarks

The integration of prepositions in wordnets, in itself, is currently a non-controversial issue that responds to an identified and open challenge for this model, in particular when it comes to semantically full prepositions. However, the encoding of prepositions reveals further lexicographic challenges concerning the description of their meaning.
In this paper, we aimed at showing that the integration is possible and quite easy, requiring mainly the linguistic adaptation of the tests and conditions that mediate the establishment of the relations of synonymy, antonymy, hyperonymy and cause between prepositional nodes. We demonstrate that the integration of prepositions results in a more complete description of other lexical items, such verbs and verbal selection properties, but it also allows for accounting for computational processes of meaning compositionality.
In what concerns lexicographic strategies, the integration and description of prepositions show the difficulties of establishing equivalences between the concepts denoted by prepositions in different languages, as well as using glosses in natural language to describe their meaning. The use of visual information obviates these issues, while posing issues on implementation.
Finally, the integration of prepositions makes wordnets more useful and usable resources, by augmenting the words described and the quantity of information encoded, and contributes to test other lexicographic strategies, as for instance freeing glosses to serve other lexicographic purposes, instead of being used to describe the meaning of lexical units when the semantic relations available are not sufficient.

## 6. Acknowledgments

I thank the anonymous reviewers for the careful reading and for all the suggestions and critiques that definitely contributed to the improvement of this paper.

## 7. Bibliographical References

Amaro, R. (2009). Computation of Verbal Predicates in Portuguese: Relational Network, Lexical-Conceptual Structure and Context - the case of verbs of movement. PhD thesis. University of Lisbon.

Amaro, R., Mendes, S. & P. Marrafa (2010). Encoding Event and Argument Structures in Wordnets. In Sojka P., Horák, A., Kopeˇcek, I. & K. Pala (eds.), Text, Speech and Dialogue, LNAI 6231, Berlim Heidelberg: Springer-Verlag, pp. 21–28.

Amaro, R., Mendes, S., and Marrafa, P. (2013). Increasing Density through New Relations and PoS Encoding in WordNet.PT. *International Journal of Computational Linguistics and Applications*, 4(1): 11–27.

American Heritage Dictionary of the English Language (2016). Fifth Edition, Houghton Mifflin Harcourt Publishing Company.

Asher, N. & P. Sablayrolles (1996). A Typology and Discourse Semantics for Motion Verbs and Spatial PPs in French. In Pustejovsky, J. & B. Boguraev (eds.), Lexical Semantics: the Problem of Polysemy, Oxford: Clarendon Press, pp. 163-209.

Baldwin, T., Kordoni, V. & Villavicencio, A. (2009). Prepositions in applications: a survey and introduction to the special issue. Computational Linguistics, 35(2):119–149.

Bannard, C. & T. Baldwin (2003). Distributional Models of Preposition Semantics. In Proc. of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, Toulouse, pp. 169–180.

Bond, F. & M. Piasecki (2017). Introduction: Contemporary Challenges for Development and Application of Wordnets. In Proc. of the Workshop on Challenges for Wordnets. Galway. http://ceur-ws.org/Vol-1899/wordnet_preface.pdf.

Cannesson, E. & P. Saint-Dizier (2002). Defining and representing preposition senses: A preliminary analysis. In Proc. of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, USA, pp. 25–31.

Crystal, D. (ed.) (1995). The Cambridge Encyclopedia of the English Language. Cambridge University Press, Cambridge.

Cunha C. & L. Cintra (1984). Nova Gramática do Português Contemporâneo. Edições João Sá da Costa, Lda, Lisboa.

Dimitrova, T., E. Tarpomanova & B. Rizov (2014). Coping with Derivation in the Bulgarian Wordnet. In Proc. of the 7th Global Wordnet Conference, pp. 109-117. http://www.aclweb.org/anthology/W14-0115

Dorr, B. J. (1997). Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. In Machine Translation, 12 (4): 271–322.

Fellbaum, C. (ed.) (1998). WordNet. An Electronic Lexical Database, MA: The MIT Press.

Galton, A. (1997). Space, time, and movement. In Stock, O. (ed.), Spatial and temporal reasoning, Kluwer Academic Publishers, pp. 321-352.

Godoy, L. (2008). Preposições e os verbos transitivos indiretos: interface sintaxe-semântica lexical. In Revista da ABRALIN, vol. VII, pp. 49-68

Gouws, R. H. (2004). State of the art: Lexicology and Lexicography: Milestones in metalexicography. In P. G. J. van Sterkenburg, Piet (ed.) Linguistics Today: Facing a Greater Challenge. John Benjamins Publishing Company. Volume 1, pp. 187-206.

Hernández-Pastor, D., Periñán-Pascual, C. (2016). Developing a knowledge base for preposition sense disambiguation: A view from Role and Reference Grammar and FunGramKB. Onomázein, no. 33, Editorial Pontificia Universidad Católica de Chile.

Herzog, G. (1995). Coping with Static and Dynamic Spatial Relations. In Amsili, P., M. Borillo & L. Vieu (eds.), Proc. of TSM'95, Time, Space, and Movement: Meaning and Knowledge in the Sensible World, Château de Bonas, pp. 47–59.

Jensen, P. & J. F. Nilsson (2003). Ontology-Based Semantics for Prepositions. In Proc. of ACL-SIGSEM workshop: The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, Toulouse.

Klein, W. (2001). Lexicology and lexicography. In N. Smelser, & P. Baltes (eds.), International encyclopedia of the social & behavioral sciences: vol. 13, pp. 8764-8768.

Lockwood, K., K. Forbus & J. Usher, J. (2005). SpaceCase: A model of spatial preposition use. In Proc. of the 27th Annual Conference of the Cognitive Science Society, Stressa.

Mcshane, M, S. Beale & S. Nirenburg (2005). Disambiguating Homographous Prepositions and Verbal Particles in an Implemented Ontological Semantic Analyzer. Working Paper 01-05, Institute for Language and Information Technologies University of Maryland.

Neeleman, A. (1997). PP-Complements. In Natural language and linguistic theory, n. 15, pp. 89-137.

O'Hara, T. & J. Wiebe (2003). Preposition semantic classification via Treebank and FrameNet. In Proc. of the 7th Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, pp. 79–86.

Pribbenow, S. (1993). Computing the meaning of localization expressions involving prepositions: The role of concepts and spatial context. In Zelinsky-Wibbelt, Cornelia (ed.), The Semantics of Prepositions. From Mental Processing to Natural Language Processing. De Gruyter, pp. 441–470.

Sag, I. & Wasow, T. (1999). Syntactic Theory: A Formal Introduction, CSLI Publications, Stanford, CA.

Saint-Dizier, P. (2005). PrepNet: a framework for describing prepositions: preliminary investigation results. In Bunt, H., J. Geerzen & E. Thijse (eds.), Proc. of the 6th International Workshop on Computational Semantics (IWCS'05). ITK, Tilburg, pp. 25-34.

Saint-Dizier, P. (2008). Syntactic and Semantic Frames in PrepNet. In International Joint Conference on Natural Language Processing (IJCNLP 2008), ACL, Hyderabad (Inde), p. 763-768.

Schneider, N., Srikumar, V., Hwang. J. D, & Palmer, M. (2015). A Hierarchy with, of, and for Preposition Supersenses. In Proc. of LAW IX - The 9th Linguistic Annotation Workshop, Denver, Colorado, ACL, pp. 112–123.

Srikumar, V. & Roth, D. (2013). Modeling Semantic Relations Expressed by Prepositions. In Transactions of the Association for Computational Linguistics, 1 pp. 231–242.

Verspoor, C. M. (1997). Contextually-Dependent Lexical Semantics. PhD dissertation, University of Edinburgh.

Vossen, P. (ed.) (2002). EuroWordNet General Document", version 3. University of Amsterdam, http://www.vossen.info/docs/2002/EWNGeneral.pdf

# Presenting the *Nénufar Project*: a Diachronic Digital Edition of the *Petit Larousse Illustré*

**Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo**
PRAXILING UMR 5267 CNRS & Univ Paul Valéry Montpellier 3 - Montpellier, France
name.surname@univ-montp3.fr

**Mohamed Khemakhem[1,2,3], Laurent Romary[1,2,4]**
[1] Inria – ALMAnaCH, Paris
[2] Centre Marc Bloch, Berlin
[3] Université Paris Diderot, Paris
[4] Berlin-Brandenburgische Akademie der Wissenschaften, Berlin
name.surname@inria.fr

**Abstract**

This paper presents the Nénufar project, which aims to make several successive (free of copyright up to 1948) editions of the **French Petit Larousse Illustré** dictionary available in a digitised format. The corpus of digital editions will be made publicly available via a web-based querying interface, as well as distributed in a machine readable format, TEI-LEX0.

**Keywords:** TEI, Petit Larousse, dictionaries

## 1. Introduction

The digitisation of historical dictionaries has recently taken on strong momentum, moving past the mere publication of scanned texts to the conversion of paper dictionaries into easily exploitable lexical databases encoded using well established digital standards. At the same time, a number of the main historical French dictionaries (16th to 19th century) are also currently being digitised and made available online. Two main initiatives in this regard are *Grand Corpus des dictionnaires Garnier*[1] and the ARTFL project[2], which provide access to the content by means of search interfaces (though access is partly restricted and sources aren't downloadable)[3]. On the other hand there is a lack of similar initiatives for 20th century French dictionaries.

The **Nénufar**[4] **project** aims to make several successive editions of the *Petit Larousse Illustré* (PLI) available in a digitised format. The PLI makes an especially good candidate for such a project since it is the only French dictionary that has been updated every year since it was first published, in this case in 1905[5]. Under the French copyright law, collective works such as the PLI fall under the public domain after 70 years from the publication, which means that we can at present take into account all editions up to 1948.

Each new edition of the PLI differs from the previous one in terms of lexical entries (with a number of words entering or exiting); but changes are also found in updated definitions and at times in the orthographic and grammatical norms which are referred to, all of which provides lexicographers, linguists and historians with an invaluable source of information on the evolution of French language and culture during the first half 20th century. At the same time, the evolution of language notwithstanding, the PLI is also an important source of linguistic information on contemporary French, and its digitisation will feed into the existing ecosystem of French language technologies (see (Mariani et al., 2012) for an overview).

## 2. The Project

Nénufar is a project headed by laboratoire Praxiling at the Paul Valéry University of Montpellier in collaboration with INRIA, and is supported by funding from the Délégation Générale à la Langue Française et aux Langues de France (DGLFLF) and the Huma-Num consortia CORLI[6] and CAHIER[7]. It continues a previous project, initiated in the early 2000s, which saw the publication of a first version of the 1905 edition in 2005[8].

The original edition was publicly accessible for searching from a web interface, but this is no longer the case; moreover, the XML encoding used was not fully TEI compliant. The first goal of the Nénufar project is thus to re-encode the 1905 edition, transforming the existing version into a TEI compliant XML, as well as correcting remaining OCR errors and improving the detection and annotation of the main lexicographic elements of each entry.

The availability of an already existing digitised version of the first edition makes the digitisation of later editions much easier: by comparing two OCRed versions of two subsequent editions it is possible to identify changes in the more recent edition, but also undetected OCR errors from the previous one.

[1] http://www.classiques-garnier.com/

[2] http://artfl-project.uchicago.edu

[3] Gallica also provides access to OCRed scans of old dictionaries, http://gallica.bnf.fr/.

[4] Nouvelle édition numérique de fac-similés de référence.

[5] The PLI is still published today and is the best selling dictionary for the French language.

[6] https://corli.huma-num.fr/

[7] http://cahier.hypotheses.org/

[8] This first initiative was headed by laboratoire Lexique, Dictionnaires et Informatique, under the lead of Jean Pruvost, who is now an advisor in Nénufar.

While the PLI was published every year since 1905 the project will prioritise the digitisation of only a selected set of issues, which correspond to major re-editions of the dictionary - namely the 1924, 1936, 1948 ones.
Currently the 1924 edition is being digitised, and we calculated that 1/3 of its entries were modified with respect to the 1905 one.
A first release of the Nénufar corpus, including the 1905 and the 1924 editions, will take place by the end of 2018. New editions will be subsequently made available. Alongside with the lexicographic part, it will also contain additional onomastic information (from the encyclopaedic section of the PLI, listing proper names of people, places, ....) and a digitised version of all figures with their captions.

## 3. The Formats

The question of publication formats is crucial for a project such as this one, which caters to different research communities. On the one hand, in order to fit the requirements of the general public as well as of traditional historical lexicographers, we need to provide a browsable web interface, which enables users to search for entries and see their evolution over time in a user-friendly way. On the other hand, the needs of digital lexicographers and language technologists can only really be met by making the sources of each edition available in a standardised format, something that would not only allow for more specialised querying, but would also be best suited for long term preservation.
Currently two formats are under discussion for the publication of retrodigitised dictionaries such as PLI, namely the TEI dictionaries module[9], the Ontolex-Lemon model (RDF) (McCrae et al., 2017). Those two formats serve different purposes: TEI represents the dictionary as a digital edition, and is better suited to the needs of lexicographers and linguists, while Ontolex-Lemon is the reference format for the publication of dictionaries as Linked Open Data, and thus is more relevant for the domain of Language and Semantic Web technologists.
As to the encoding of PLI in TEI, the first step was to transform the 2005 mark-up in a TEI compliant format, which is the one presented in Appendix B. This first encoding remains very adherent to the structure of the typographic entry, as can be seen in Appendix A, and thus uses the *entryFree* TEI tag, which allows for maximum freedom in the representation and encoding of the different parts of a lexical entry. For this reason it is the one that will be used internally in the Nénufar database to derive the HTML displayed on the browsable web interface.
However an excessive freedom in terms of entry modelling can become a hindrance to interoperability with other projects. For this reason a recent a joint ENeL[10] / DARIAH[11] / PARTHENOS[12] initiative has proposed a more strict TEI representation for dictionaries, called TEI-Lex0 (Bański et al., 2017). TEI-Lex0 derives from the lexicographic module of TEI and is fully TEI compliant, but aims to provide more clear guidelines for the encoding of retrodigitised dictionaries.
With respect to the more general TEI guidelines for dictionaries, TEI-Lex0 is aimed at providing a schema which will allow most modern dictionaries to be represented in a way that enables interoperability, comparability and further ease of exploitation. To that end, the internal structure and information of lexical entries have been revised and optimised to be more clearly explicit and uniform.
We believe that the PLI can constitute an excellent test case for this new format, which we intend as the distribution format for the downloadable resource. In Appendix C you can find the same entry transformed into the TEI-Lex0 format.
As you can see, going from the current format to the new one requires some changes; some of them (such as the insertion of the *type* attribute in the *form* tag) are straightforward, but others are more complex to implement.
Firs of all the *entryFree* tag is replaced by *entry*, which allows for less freedom as to the tags it may contain. As a consequence, the original structure cannot be left as it is. In particular the *sense* tag needs to be inserted, to group a definition with its related examples and citations. This implies adding information which, in the original entry is not explicitly marked by visible typographic features (such as numbering, symbols or formatting, as is the case in other dictionaries). By close analysis of the PLI entries, we consider that every new definition instantiates a new sense, and that no sense hierarchy is inferable.
Another issue is the fact that free text is not allowed within the *sense* tag. Thus *pc* tags need to be used to wrap up punctuation elements such as columns, as they cannot be considered neither as part of the definition, nor of the citation.
Despite the work required to transform the current format into TEI-Lex0, the advantages are obvious; TEI-Lex0 will allow for different dictionaries to be queried using the same strategy and also facilitate the development of common tools.
One of the current applications of this format is in the GROBID-Dictionaries infrastructure, which aims to automatically machine-learn the TEI-Lex0 structure of a dictionary entry from OCRed dictionary pages (Khemakhem et al., 2017). Within the Nénufar project experiments are ongoing to digitise new editions with GROBID-Dictionaries.
As to the Ontolex-Lemon version, at the time of writing this paper (March 2018) a working group is active drafting the specifications for a dictionary module, which will enable to represent retro-digitised dictionaries using the Ontolex-Lemon core with additional properties. The specifications are not yet finalised, and the final modelling of PLI in this new format will be the object of further research; it is important however to underline how PLI entries from the 1905 edition are currently being used as examples to discuss the new module issues[13].
As to the availability of the two versions, the TEI edition will be downloadable from the Ortolang[14] platform, and the Ontolex-Lemon will be queryable via a SPARQL endpoint.

[9](Budin et al., 2012), see also `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html`
[10]`http://www.elexicography.eu/`
[11]`https://www.dariah.eu/`
[12]`http://www.parthenos-project.eu/`
[13]`https://www.w3.org/community/ontolex/wiki/Lexicography`
[14]`http://www.ortolang.fr`

Finally, two modelling issues are of a more generic nature and will affect both formats. On the one hand homographs are generally but not systematically treated as separate entries in the PLI; this may represent a problem as to the encoding of grammatical properties at the entry level and may require adjustments. On the other a normalisation of data categories for grammatical features is required and currently on-going; the grammatical labels (gender, number, language, ...), represented with in the original by (often unsystematic) French abbreviations, will be normalised using existing controlled vocabularies; in this sense, the CLARIN Concept Registry may [15] constitute a valid solution.

## 4. The Content

Dictionaries are the "tools of a language and a culture" (Pruvost, 2006) and the PLI, whose millions of copies over more than 110 years have found place in the majority of French households, has played and still plays a great role in the democratisation of linguistic knowledge(Cormier et al., 2006); for this reason the diachronic investigation of its successive editions sheds a new light on the evolution of French language and society.

First and foremost the Nénufar corpus will constitute a privileged source of information on the evolution of orthography. The name of the project itself is inspired by a surprising controversy sparked in 2016 by the proposed change in the spelling of the French word for waterlily, from *nénuphar* to *nénufar*. Despite the fact that the new spelling was strongly ostracised by the people and by the media, an inspection of early editions of PLI shows that the *nénufar* spelling was already present in the 1905 edition and remained the preferred orthography for the word for the whole of the first half of the 20th century. Other orthographies attested in the earlier versions PLI would be considered almost shocking today, such as *à priori* (with an accent), *fiord* instead of *fjord*, *ognon* as an alternate spelling for *oignon* (the French for *onion*).

Apart from the evolution of orthography, the older editions of the PLI are rich in information about phonetics ([distrik], [lo-kouass] for *district* et *loquace* en 1906), neologisms (*antimilitarisme* in 1911, *boche*, the equivalent of the English pejorative word for German, in 1917, etc.) and changes in the definitions. As to the these, some are rather amusing, such as the one for *aviation*, which in 1905 reads "on a fait de nombreuses tentatives à ce sujet mais le problème n'est pas encore résolu" (several tests have been carried out but the problem hasn't been solved yet) and in 1911 becomes "les aéroplanes ont victorieusement résolu le problème du plus lourd que l'air" (planes have victoriously solved the heavier-than-air controversy). In other cases (as in the older entries for *juiverie* or *nègre, négresse*) definitions bear testimony of the evolution of society, of which the PLI is the mirror.

## 5. Conclusion

In this paper we presented Nénufar, an ongoing project aimed to the digitisation of chosen editions of the *Petit Larousse Illustré* from the first half of the 20th century.

A first TEI and web release of the Nénufar corpus will be available in 2018 with an open license, thus enabling research in the domains of linguistics, history and language technologies to research and use this

To ensure interoperability, the project is carried out in close contact with on-going international initiatives aimed at promoting standard and best practices in the retro-digitisation of legacy dictionaries[16]. Moreover, it is currently used as a test bed for GROBID-Dictionaries, a technology which will considerably speed up the encoding of OCRed resources.

The current project is specifically targeting the PLI, but the best practices developed within Nénufar will be applicable to other legacy dictionaries.

## 6. Bibliographical References

Bański, P., Bowers, J., and Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In *eLex2017*.

Budin, G., Majewski, S., and Mörth, K. (2012). Creating Lexical Resources in TEI P5. *Journal of the Text Encoding Initiative*, (Issue 3), November.

Cormier, M.-C., Pruvost, J., Mitterand, H., Garnier, Y., and Collectif. (2006). *Les dictionnaires Larousse : Genèse et évolution*. PU Montréal, Montréal, March.

Khemakhem, M., Foppiano, L., and Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In *electronic lexicography, eLex 2017*, Leiden, Netherlands, September.

Joseph Mariani, et al., editors. (2012). *La langue française à l' Ère du numérique – The French Language in the Digital Age*. White Paper Series. Springer-Verlag, Berlin Heidelberg.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *eLex2017*.

Pruvost, J. (2006). *Les dictionnaires français : Outils d'une langue et d'une culture*. Ophrys, Paris.

[15] https://concepts.clarin.eu/ccr/browser/

[16] In addition to what was mentioned in this paper, Nénufar is planning on collaborating with the ELEXIS project, which recently kicked off and aims at building a European Infrastructure for E-lexicography (http://www.elex.is/)

# Appendices

## A The dictionary entry *verre* (glass) in the PLI .

**VERRE** (*vè-re*) n. m. (lat. *vitrum*). Corps solide, transparent et fragile, produit de la fusion d'un sable siliceux mêlé de potasse ou de soude : *le verre est très cassant*. Objet fait de verre : *verre de montre*. Vase à boire, fait de verre ; ce qu'il contient : *un verre de vin*. *Verre double*, verre très épais. *Maison de verre*, maison où il n'y a rien de secret. *Petit verre*, liqueur alcoolique qu'on prend dans un verre de petite dimension : *boire un petit verre*. — Le *verre*, dont l'invention est attribuée aux Phéniciens, est obtenu par la fusion dans des *creusets* (ou *pots*) d'un mélange de silice (sable) avec des sels de soude, de potasse (*verre ordinaire*) ou de plomb (*cristal*). Les creusets sont placés dans des *fours* où la température est poussée jusqu'à 1.000°. Cueilli avec une *canne* que l'on plonge dans les creusets par une ouverture (*ouvreau*) pratiquée dans la paroi du four, le verre pâteux est travaillé, soufflé, moulé, étiré, pour donner des bouteilles, des vitres, des objets de gobeleterie, des tubes, etc. Les glaces sont obtenues par *coulage ;* on sort du four le creuset tout entier et l'on en verse le contenu sur une immense table de fonte. Tous les objets de verre, avant d'être livrés au commerce et indépendamment des façons qu'on leur fait subir ou des décors dont on les agrémente, doivent être *recuits* c'est-à-dire refroidis lentement, pour être moins cassants. Outre les mille objets à l'usage domestique, le verre sert encore à fabriquer les verres optiques et les instruments si nombreux utilisés dans les laboratoires. Ramolli au four et comprimé fortement, il donne la *pierre de verre*, qu'on emploie au revêtement des murs et même au pavage des rues.

Véronique.

## B The first TEI-XML encoding

```
<entryFree xml:id="verre">
  <form>
    <orth>VERRE</orth>
  </form>
  <pron>(vè-re)</pron>
  <gramGrp>
    <pos>n.</pos>
    <gen>m.</gen>
  </gramGrp>
  <etym>
    (<lang>lat.</lang> <mentioned>vitrum</mentioned>)
  </etym>.
  <def>Corps solide, transparent et fragile, produit de la fusion d'un sable
    siliceux mêlé de potasse ou de soude</def> :
  <cit type="example"><quote>le verre est très cassant.</quote></cit>
  <def>Objet fait de verre</def> :
  <cit type="example"><quote>verre de montre.</quote></cit>
  <def>Vase à boire, fait de verre ; ce qu'il contient</def> :
  <cit type="example"><quote>un verre de vin.</quote></cit>
  <re type="exp"><form>Verre double</form>, <def>verre très épais.</def></re>
  <re type="exp"><form>Maison de verre</form>,
    <def>maison où il n'y a rien de secret.</def>
  </re>
  <re type="exp"><form>Petit verre</form>,
    <def>liqueur alcoolique qu'on prend dans un verre de petite dimension</def> :
    <cit type="example"><quote>boire un petit verre.</quote></cit>
  </re> -
  <def value="encycl">
    Le <emph rend="italic">verre</emph>, dont l'invention est attribuée
    aux Phéniciens, est obtenu par la fusion dans des <emph rend="italic">
      creusets</emph> (ou <emph rend="italic">pots</emph>) d'un mélange de
    silice (sable) avec des sels de soude, de potasse (<emph rend="italic">
      verre ordinaire</emph>) ou de plomb (<emph rend="italic">cristal</emph>.)
    Les creusets sont placés dans des <emph rend="italic">fours</emph> où la
    température est poussée jusqu'à 1.000°. Cueilli avec une <emph rend="italic">
      canne</emph> que l'on plonge dans les creusets  par une ouverture
    (<emph rend="italic">ouvreau</emph>) pratiquée dans la paroi du four,
    le verre pâteux est travaillé, soufflé, moulé, étiré, pour donner des
    bouteilles, des vitres, des objets de gobeleterie, des tubes, etc.
    Les glaces sont obtenues par <emph rend="italic">coulage</emph> ;
    on sort du four le creuset tout entier et l'on en verse le contenu
    sur une immense table de fonte. Tous les objets de verre, avant d'être
    livrés au commerce et indépendamment des façons qu'on leur fait subir
    ou des décors dont on les agrémente, doivent être
    <emph rend="italic">recuits</emph> c'est-à-dire refroidis lentement,
    pour être moins cassants. Outre les mille objets à l'usage domestique,
    le verre sert encore à fabriquer les verres optiques et les instruments
    si nombreux utilisés dans les laboratoires.
    Ramolli au four et comprimé fortement,
    il donne la <emph rend="italic">pierre de verre</emph>, qu'on emploie
    au revêtement des murs et même au pavage des rues.
  </def>
</entryFree>
```

# C The TEI-LEX0 encoding

```
<entry>
  <form type="lemma">
    <orth>VERRE</orth>
    <pron>(vè-re)</pron>
  </form>
  <gramGrp>
    <pos>n.</pos>
    <gen>m.</gen>
  </gramGrp>
  <etym> (<lang>lat.</lang>
    <mentioned>vitrum</mentioned>)</etym><pc> .</pc>
  <sense>
    <def>Corps solide, transparent et fragile, produit
      de la fusion d'un sable siliceux
      mêlé de potasse ou de soude</def><pc> :</pc>
    <cit type="example">
      <quote>le verre est très cassant</quote>
    </cit>
  </sense><pc>.</pc>
  <sense>
    <def>Objet fait de verre</def><pc> :</pc>
    <cit type="example">
      <quote>verre de montre</quote>
    </cit>
  </sense><pc>.</pc>
  <sense>
    <def>Vase à boire, fait de verre ;
      ce qu'il contient</def><pc> :</pc>
    <cit type="example">
      <quote>un verre de vin</quote>
    </cit>
  </sense><pc>.</pc>
  <re type="exp">
    <form type="lemma">
      <orth>Verre double</orth>
    </form>, <def>verre très épais</def></re><pc>.</pc>
  <re type="exp">
    <form type="lemma">
      <orth>Maison de verre</orth>
    </form>, <def>maison où il n'y a rien de
      secret</def>
  </re><pc>.</pc>
  <re type="exp">
    <form type="lemma">
      <orth>Petit verre</orth>
    </form><pc>,</pc>
    <sense>
      <def>liqueur alcoolique qu'on prend dans un
        verre de petite dimension</def><pc> :</pc>
      <cit type="example">
        <quote>boire un petit verre</quote>
      </cit>
    </sense>
  </re><pc>.</pc>
  <sense>
    <def value="encycl"> - Le <emph rend="italic">verre</emph>, dont l'invention
      est attribuée aux Phéniciens, [...see encoding above....]</def>
  </sense>
</entry>
```

# Crafting a lexicon of referential expressions for NLG applications

**Ariel Gutman, Alexandros Andre Chaaraoui, Pascal Fleury**
Google Research Europe, Zurich
{relgu, alexandrosc, fleury}@google.com

**Abstract**

To engage users, a natural language generation system must produce grammatically correct and eloquent sentences. A simple NLG architecture may consist of a template repository coupled with a lexicon containing grammatically-annotated lexical expressions referring to the entities that are present in the domain of the system. The morphosyntactic features associated with these expressions are crucial to render grammatical and natural-sounding sentences. Existing electronic resources, like dictionaries or thesauri, lack wide-scale coverage of such referential expressions. In this work, we focus on the creation of a large-scale lexicon of referential expressions, relying on n-gram models, morpho-syntactic parsing, and non-linguistic knowledge. We describe the collected linguistic information and the techniques used to perform automatic extraction from large text corpora in a way that scales across languages and over millions of entities.

**Keywords:** Natural Language Generation, Lexicon Extraction, Referential Expressions

## 1. Introduction

Dialogue systems, such as voice-driven personal assistants or conversational chat-bots, as well as other natural language generation (NLG) applications are bound to produce appropriate, grammatical and well-formulated utterances, in order to engage the human user. One often-overlooked prerequisite for such behaviour is the use of correct lexical information regarding the entities in the domain of the system (e.g., place names, names of people, etc.). In this paper, we shall describe several techniques that make it possible to acquire such information automatically at a large scale.

A typical architecture of an NLG system has distinct modules for content planning, sentence planning and sentence realization, as outlined by Reiter and Dale (2000) or Walker and Rambow (2002). A simple sentence realization module may contain the following two components:

1. A template repository, which stores the various messages which the system can generate. These templates, each created for a specific communicative intent of the system, may correspond broadly speaking to the notion of *constructions* of the *construction grammar* framework (Goldberg, 1995): they are a mixture of lexical, syntactic and surface form specifications for each utterance.

2. The lexicon, containing the lexical forms (lexemes) and the relevant grammatical information of the entities in the domain of the system.

The usage of a template-based sentence realization system is, of course, quite old (see Weber and Mendoza (1973) for a description of a very early system which produces haikus). In their simplest form, template-based systems have been contrasted with true NLG (Reiter, 1995). Yet the addition of the second component, namely a linguistically annotated lexicon, makes them truly NLG-worthy. NLG lexica have typically been hand-crafted, but this is not possible if the scale of the required domain is very big (e.g. weather reports for all localities on Earth).

As stated above, in this paper we are concerned with the automatic crafting of such large-scale lexica in a multilingual setting. Morphosyntax and surface form variations are very language-specific, as will be illustrated below with some languages for which we created lexica: Czech, English, French, Swedish and Russian. We are especially interested in acquiring information about *referential expressions*, i.e. expressions which have specific referents in the world (either real or fictional), e.g. *Paris*, *The Beatles*, or *James Bond*. Such expressions are often termed *proper nouns* or *proper names*; in either case we note that they can superficially seem as compositional noun phrases, such as *The Great Lakes*.

Being noun phrases, these referential expressions exhibit grammatical properties that can affect the selection and form of surrounding words, due to phenomena such as grammatical agreement, preposition selection and the like. Therefore, they cannot simply be plugged into an empty slot in the template, as part of the template may need to be re-edited. Instead, the template needs to be specified in such a way that this lexical information is taken into account. Moreover, in some cases, the combination of information from multiple referential expressions is needed to generate the grammatically correct form of a sentence. This happens, for example, with the gender of a list of conjoined nouns in French: a single masculine noun in it will trigger masculine agreement with any element dependent on the list.

An important property of referential expressions, in contrast to more conventional lexemes of a language, is their large scale. Thus, the Second Edition of the 20-volume Oxford English Dictionary contains about 300,000 entries (Simpson and Weiner, 1989), yet the number of referential expressions is theoretically unlimited and in practice could reach tens of millions, depending on the domain of the NLG system. This immense richness of referential expressions is often overlooked since many NLU systems, such as parsers, do not require grammatical information about these names: it suffices for an NLU system to mark these names as such. If moreover, the referential expression is compositional, its proper name nature can be overlooked.

Thus, most electronic lexical resources concentrate on the common lexemes of language, such as common nouns,

verbs or adjectives. For instance, Sagot (2010) presents a lexical database of French containing about 110,000 lemmas, out of which only about half are proper nouns. Moreover, the grammatical information needed for proper nouns is often not encoded in standard lexical resources or dictionaries. For example, in some languages various toponyms require different locative prepositions (for instance, islands require in general the preposition "on" in English, though some larger islands, or island groups, are exempt). Such information is usually not present in dictionaries, or it can only be deduced from examples given there.

In this paper, we present three different systems to acquire large-scale lexical data consisting mainly of referential expressions (as well as common nouns), in a multilingual setting. Two of the systems use data-mining methods to extract information from corpora, in which referential expressions are marked and linked to an entity's identifier in a non-linguistic knowledge base of entities, such as a geographical repository or a database of people. The corpora we used include Wikipedia pages, as well as selected news sites. The difference between the two approaches is related to the amount of grammatical annotation the corpus has. For some languages, which we call "high-resource languages", a parser may be at our disposition, while for others, called here "low-resource languages" we have no such tools. The third system is a last-resort rule-based system which "guesses" the grammatical properties of a given referential expression using available knowledge at the time of generation.

We present below a simple example of the type of information we want to acquire, and subsequently the three systems.

## 2. A simple example of a lexicon

Consider an NLG system which produces weather reports for various localities. It may contain a template as the following:

```
It is sunny in (Location).
```

In this template, the placeholder `(Location)` is to be replaced with a name of a location (a *toponym*):

```
It is sunny in Paris.
```

Yet it is easy to see that such a simplistic template would generate ungrammatical sentences if the location requires a different preposition, as is typically the case with islands or lakes:

```
It is sunny on Tenerife.
It is sunny at Lake Como.
```

This last example also illustrates that the possible choices are constrained by the referential expression, but also by the wanted semantics, as *on Lake Como* would be another perfectly acceptable phrase in this context, but with a slightly different meaning.

To accommodate such cases, the template has to be rewritten so that the correct preposition is chosen:

```
It is sunny (Locative
preposition + Location).
```

Once the template has been amended, the system now relies on the correct preposition being specified in the lexicon for each entity (see Table 1).

| Name | Preposition |
|---|---|
| Paris | in |
| Tenerife | on |
| Lake Como | at |

Table 1: Samples of different locative prepositions in English.

A further complication is presented by toponyms such as *the Isle of Man*, for which we expect the following message:

```
It is sunny in the Isle of Man.
```

Yet the determiner *the* is not an integral part of the toponym, as is evident from the fact that it can be removed in certain expressions (*Britain's Isle of Man*) and would not appear in a listing of countries or on a map. Thus, the lexicon needs to be augmented with information about determiners as shown in Table 2.

| Name | Preposition | Determiner |
|---|---|---|
| Paris | in | - |
| Tenerife | on | - |
| Lake Como | at | - |
| Isle of Man | on | the |

Table 2: Locative prepositions and the required determiner for different English toponym samples.

An English lexicon may additionally contain traditional grammatical information about gender and number, to be used for instance in pronominalization or verbal agreement, or phonological information, such as whether a lexeme starts with a vowel. To exemplify the latter, contrast *Australia* with *Uruguay*, where only the former has a vocalic onset, yielding expressions like *an Australian city* versus *a Uruguayan city*. In languages with richer morphology like Russian, the lexicon may additionally enumerate the various case inflections of a given name, which are often idiosyncratic for proper nouns, or provide other necessary pieces of grammatical information, such as animacy in Russian. Apart from the grammatical information, the lexicon may be enriched with multiple names for a given entity, be it short or long versions of the same name (*Frankfurt* vs. *Frankfurt am Main*) or various nicknames of entities (*the Big Apple* vs. *New York*).

## 3. N-gram-based lexicon extraction

For low-resource languages, i.e. languages for which some amount of written material can be found in the web, we have at our disposition a corpus of texts lacking grammatical annotation. A prerequisite of the lexicon extraction

process, however, is that the potential referential expressions are identified in the corpus, and are linked to the relevant entities in the knowledge base of the system, a process known as named-entity extraction (Momchev, 2010). Since in this case we do not possess any grammatical annotation of the text, we rely on the insight that functional words in the vicinity of the referential expressions may give us information regarding the grammatical features of the expression, a method that has been shown to explain similar aspects of child language acquisition (Gutman et al., 2015). For instance, if we want to deduce the gender of the French toponym *Paris*, we may observe the presence of the masculine determiner *le* in the expression *le grand Paris* and deduce that Paris is a masculine toponym. At the same time, we may observe the text *Paris est belle*, from which we would deduce that it is actually a feminine toponym, probably due to the feminine gender of the latent concept *ville* ("city"). This hints at the fact that such proper nouns usually do not have a fixed grammatical gender, a property which could potentially also be modeled by the extracted annotations.

In practice, however, in order to use this procedure, we provide for each language only a short table of functional words (typically determiners) associated with their grammatical properties. For example, for French we used the data presented in Table 3. In this table, grammatical features are shown in the columns, the functional words in rows, and the modeled *attributes* in the cells. Note that some function words do not provide any information regarding a given feature, so the corresponding table cell is empty, e.g. the plural determiners that are gender-neutral (or underspecified) in French. Conversely, one form may be associated with competing features: in German, the determiner *die* can be either feminine singular or gender-neutral plural, and the determiner *der* could be masculine singular nominative or feminine singular genitive.

| | Gender | Number | Elision |
|---|---|---|---|
| *le* | masc. | sg. | - |
| *la* | fem. | sg. | - |
| *l'* | | sg. | + |
| *les* | | pl. | |
| *un* | masc. | sg. | |
| *une* | fem. | sg. | |
| *des* | | pl. | |

Table 3: Gender, number and whether elision is applied or not for French definite and indefinite articles.

Additional data given to the system is whether these words should appear before or after the corresponding referential expression (French and German determiners appear before), and the size of the n-gram window around the named entity to examine. In practice, looking at bigrams proved to be sufficient. For features like elision-triggering, which is a sandhi phenomenon (i.e., word-edge variation which is due to morpho-phonological conditions), the system only considers the unigram adjacent to the referential expression.

Given this data, the assignment of grammatical features to referential expressions is straightforward: for every mention $m$ of referential expression $E$ in the set of mentions $M_E$, for each grammatical feature $F$, and for each possible attribute value $a_F$ of the feature, the system identifies the functional words $t$ in the window $\Omega_m$ of n-grams adjacent to the mention of the referential expression. This contributes a certain weight $w_{a_F,t}$ to the total score of the given attribute of the expression $a_{F,E}$. The score is normalized by the number of mentions $|M_E|$.

$$score(a_{F,E}) = \frac{\sum_{m \in M_E, t \in \Omega_m} w_{a_F,t}}{|M_E|} \quad (1)$$

Selection of the right attribute for a given feature $F$ of a referential expression $E$ is then done by taking the highest scoring attribute (in the set of possible attributes $A_F$), above a certain threshold $min_{a_F}$:

$$a_{F,E} = \arg\max_{a \in A_F} \{score(a_{F,E}) | score > min_{a_F}\} \quad (2)$$

The confidence threshold $min_{a_F}$ may be used in order to filter out cases where there is not enough supporting evidence for an attribute in the whole corpus. Yet in practice, as we shall see below, setting this threshold to zero allows us getting maximal coverage without compromising the quality of the results significantly.

As for the calculation of the weight $w_{a_F,t}$ this could in principle be learned from an annotated corpus. Yet since we do not have such annotations, we take a simple approach of distributing a weight of 1 over all possible attributes $A_{F,t}$ of a feature $F$ specified for a certain functional word $t$:

$$w_{a_F,t} = \begin{cases} \frac{1}{|A_{F,t}|} & \text{if } a_F \in A_{F,t} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

For example, the weight of the attribute *masculine* of the French determiner *le* is 1, while the weight of the same attribute for *les* is 0 (since no gender is specified for *les*). In the experiments we did with French and Swedish there were no cases of fractional weights, since every functional word has at most one attribute specified for each feature.

Using this approach we extracted about 800,000 lexicon entries. We selected a sample of 100 entities to evaluate the precision of the grammatical features of gender, number and elision. The results are given in Table 4, using two different confidence thresholds: 0% (i.e. no threshold) and 10%. These results are compared to a baseline result, which consists of uniformly selecting the majority group (i.e. masculine, singular and no elision). As expected, using a higher threshold increases the precision,[1] though this comes with a decreased coverage of about 40%, compared to the zero-threshold results.[2] The rest of the figures in this paper are given for the case when a zero confidence threshold is used.

[1] Surprisingly, the precision goes slightly down for the number feature. This can probably be ascribed to the usage of a small sample and the very high initial precision rate.

[2] To be more exact, out of the sample of 100 entities, only 58 entities get the gender or elision features assigned with the 10% confidence threshold, and similarly only 72 entities get the number feature assigned.

The referential expressions in the sample are a mixture of proper nouns (e.g. *Dheepan* or *Nathalie Rihouet*), proper names (*Miss France 2007*), acronyms (*FICP = Fichier national des Incidents de remboursement des Crédits aux Particuliers*) as well as common nouns (*neuvaine*) or noun phrases (*perche à selfie*). All refer to entities in the domain of the system and as mentioned before French toponyms or company names do not always have a fixed gender. For this evaluation we relied on the gender as it appears in the French Wiktionary.[3] If no gender was given, we did not include the entity in our evaluation and therefore we did not calculate a recall value.

| French | Gender | Number | Elision |
|---|---|---|---|
| Baseline | 60% | 82% | 76% |
| 0% threshold | 87% | 97% | 98% |
| 10% threshold | 98% | 96% | 100% |

Table 4: Precision results obtained for French grammatical features applying n-gram based lexicon extraction, with two different confidence thresholds. For comparison, a baseline of selecting the majority group is given as well.

The low score obtained for the gender feature, when no threshold filtering is used, can be explained by the fact that plural articles (as well as the elided article *l'*) neutralize the gender property. For example, the determiners in *l'Autriche* or *les Maldives* do not provide any information about the gender. Yet if our corpus contains a mistyped expression such as *le Maldives* (and such typos are frequent in web corpora), the system will erroneously deduce that *Maldives* is masculine in the lack of counter-evidence. This is rectified to some degree by filtering the results using a minimal scoring threshold, which we did not, however, use in the evaluation procedure. For instance, setting the threshold to 0.1 (i.e. the evidence for gender is present in at least 10% of the occurrences of every given expression) increases the gender precision to 90% while purging 30% of expressions.

The same technique was applied to Swedish, using various Swedish determiners. We used the various forms of the definite article *den*, the indefinite article *en*, the demonstrative *denna*, the possessive pronouns as *min* ("my"), as well as other determiners: *vilken* ("which"), *någon* ("some"), *ingen* ("no"), and *annan* ("another"). All these determiners exhibit number variation as well as gender variation in the singular (common or neuter gender). For Swedish we used a smaller corpus and extracted about 35,000 entities.

The precision results are shown in Table 5, evaluated on a sample of 115 common nouns and 150 proper names. The baseline results are given for an equal mix of proper and common nouns.

Here too, the lower result for gender can be explained by neutralisation of the gender feature in plural determiners. In an expression like *de nya Flugbussarna* ("the new Airport-busses") there is no information regarding the gender of the referential expression *Flugbussarna.*

[3] `http://fr.wiktionary.org`.

| Swedish | Gender | Number |
|---|---|---|
| Baseline (mixed) | 52% | 85% |
| Common nouns | 90% | 97% |
| Proper names | 66% | 92% |

Table 5: Precision results obtained for Swedish grammatical features applying n-gram-based lexicon extraction, with no confidence threshold. For comparison, a baseline of selecting the majority group is given as well.

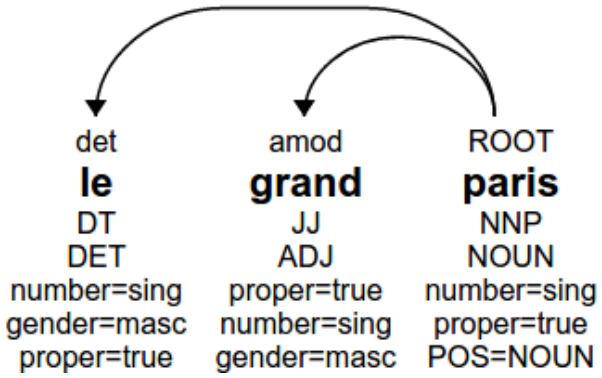



Figure 1: Extracting grammatical properties (number=singular and gender=masculine) from a determiner (DET) and an attributive adjective (ADJ). The labels on the arcs permit the extraction system to find the words which may carry the relevant information (det=determiner arc, amod=attributive modifier arc).

## 4. Dependency-tree-based lexicon extraction

For languages for which we have access to a morpho-syntactic parser, we use a more involved system. Specifically, the morpho-syntactic parser presented in Andor et al. (2016), annotates our corpora with dependency relations and with some morphological annotations. Occasionally, the referential expression itself is annotated with the desired grammatical features (such as the grammatical gender and number) yet this is not always the case for proper nouns. Essentially, we use the same technique as before, but instead of guessing that a nearby determiner is related to the target expression, we can identify the correct determiner by virtue of the available syntactic parse (following a dependency arc). Moreover, we are not limited to specific functional items, but we can also rely on agreement morphology apparent on verbs or adjectives.

For example, we can extract the gender of *Paris* both from a determiner and an attributive adjective in the phrase *le grand Paris* and from the predicative adjective in the sentence *Paris est belle*, corresponding to the dependency trees shown in Figures 1 and 2.

Note that in both cases the parser does not give us the grammatical gender of the name *Paris*, possibly due to the difficulty of assigning such a gender.

Similarly, we can directly count which prepositions govern each referential expression in order to infer the most common locative preposition. Of course, to infer phonological sandhi features (such as the *elision* feature), the extraction

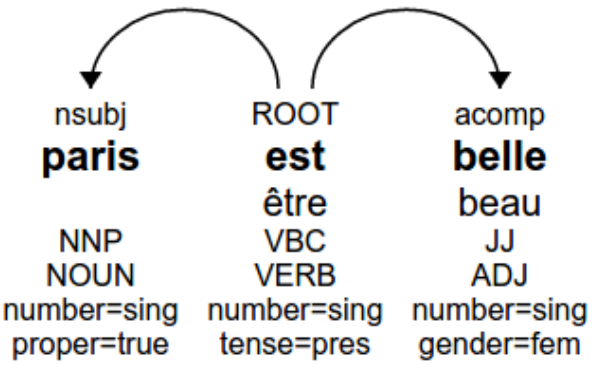



Figure 2: Extracting grammatical properties (number=singular and gender=feminine) from a predicative adjective (ADJ). Here the extraction system follows two arcs: *Paris* is the nominal subject (nsubj) of the verb *est* ("is", being the root of the tree), while *belle* is an attributive complement (acomp) of the verb.

system must still take into consideration linear adjacency rather than dependency relations.

For this process, we used a much larger corpus, and managed to extract about 7 million French lexical entries, being mostly proper names. Thus, for evaluation we used a larger evaluation set, consisting of about 46,000 entries. The precision results are given in Table 6.

| French | Gender | Number | Elision |
|---|---|---|---|
| Precision | 70% | 98% | 95% |

Table 6: Precision results obtained for French grammatical features applying dependency-tree-based lexicon extraction.

We note that the results are worse than the n-gram based model, especially for the gender feature. This is expected, since we are able to infer such properties also when no article is present (for instance by looking at a predicative adjective, as in Figure 2), but this necessarily increases the noise in the system.

Using this system we have also extracted the locative preposition of toponyms. Here we got a precision level of 88%.

## 5. Lexicon inference based on minimal information

In some cases our methods of lexicon extraction are not practicable at all, or they failed for a specific entity. Yet we may still have at our disposition *non-linguistic knowledge* about the entity coupled with some default (typically *official*) name (for instance, we may have a database of geographical names or of movie actors). In such cases we can still apply some last-resort rules to guess the relevant grammatical properties, either by detecting some morpho-syntactic pattern in the name itself, and/or by relying on the non-linguistic information.

A trivial case is if a French name starts with an article: in that case we can infer the grammatical properties directly from that article, as in the toponyms *Le Havre* or *La Rochelle.*

A less-trivial example is using the ending of a French name to infer its gender. Our investigation shows that relying on a simple heuristic of assigning feminine gender to French names ending with *-e* is correct in about two thirds of the cases.

As for non-linguistic information, if we know, for instance, that an English geographical name represents an *island*, we can guess with high probability that it should take the locative preposition *on*. Additionally, we can detect the word "island" in the name itself and apply the same heuristic. Similarly, for names of people, we may assume that the gender of the named person corresponds to the grammatical gender of the name.

We have applied this method specifically to a set of approximately 11,000 Czech toponyms, with the goal of obtaining their locative prepositions to form prepositional phrases such as *v Praze* ("in Prague"), *ve Vancouveru* ("in Vancouver"), or *na Ukrajině* ("in Ukraine"). Based on the knowledge base of the system, entities have been classified in different categories that share linguistic properties with regards to the locative preposition: expressions referring to islands, mountains, peninsulas, airports, train stations, highways, universities, castles or lakes, were assigned the locative preposition *na*, while other expressions were assigned the locative preposition *v* or its allomorph *ve*, based on the presence of certain consonantal onsets in the referential expression. Results were evaluated with a golden set of 1,200 manually annotated toponyms, where subsets were chosen based on the entity's frequency in the corpus (see Table 7).

| Sample set | Set size | Precision |
|---|---|---|
| Head - 1st tertile | 400 | 96% |
| Torso - 2nd tertile | 400 | 98% |
| Tail - 3rd tertile | 400 | 99% |

Table 7: Precision of locative preposition assignment for Czech toponyms using lexicon inference based on the type and the orthographic name of the entity.

Note that Czech nouns inflect for the locative case after these prepositions. In order to acquire the paradigm of the Czech names we still had to use an n-gram-based lexicon extraction process, in which we could identify case inflections by virtue of their co-occurrence with certain prepositions.

## 6. Conclusions

In this paper we presented various techniques to assemble information about referential expressions known more generally as *proper names*. We showed that given a corpus with annotation of referential expressions alone, we may use minimal grammatical knowledge of functional words in the language in order to infer grammatical properties. If we do have grammatical annotation we may use these to improve upon the impoverished technique.

Finally, we suggested that even when no linguistic knowledge apart from the name of an entity is available, we may

still rely on that name together with non-linguistic information about the entity to infer some grammatical properties with some confidence. In this respect, as illustrated in Figure 3, the three presented methods can be combined; especially the lexicon inference can serve as a last-resort method to assign linguistic properties to expressions which are only rarely found in the available corpora.[4] Conversely, if certain grammatical properties are generally predictable from the orthography of a name or the entity's type, we may choose to mainly rely on this method and only store in our lexicon the exceptions to the rule (which can be gathered using lexicon extraction).

In future work, we aim to address methods for selecting and grouping various referential expressions referring to the same entity. While in the simplest case we may just select the most frequently occurring referential expression as the relevant one (as we did in the above experiments), the situation is more complicated if we want to reconcile several expressions into a paradigm, as in a case-inflecting language. This can be achieved if we have some minimal knowledge of the relevant paradigms present in the language, similarly to the techniques used by Clément et al. (2004) for French verbs. A further problem is to find several different referential expressions, or paradigms of such, differing in some semantic dimension. For example, one expression could be an *official* name, and another the everyday *colloquial* name. This is in fact quite a difficult task, which warrants a separate discussion.

## 7. Acknowledgements

We would like to thank Ivan Korotkov, Jana Strnadova and Daniel Calvelo Aros for creating and working on different parts of the above described systems, as well as many other linguists and engineers who contributed to our work.

## 8. Bibliographical References

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. *CoRR*.

Clément, L., Lang, B., and Sagot, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *LREC 04*, pages 1841–1844, Lisbonne, Portugal.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language & Culture. University of Chicago Press.

Gutman, A., Dautriche, I., Crabbé, B., and Christophe, A. (2015). Bootstrapping the syntactic bootstrapper: Probabilistic labeling of prosodic phrases. *Language Acquisition*, 22(3):285–309.

Momchev, N. (2010). Annotating web documents with Wikipedia entities. Master's thesis, Sofia University.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

Reiter, E. (1995). NLG vs. templates. In *Proc of the Fifth European Workshop on Natural-Language Generation (ENLGW-1995)*, Leiden.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May.

John Simpson et al., editors. (1989). *The Oxford English Dictionary*. Oxford University Press, Oxford, second edition.

Marilyn Walker et al., editors. (2002). *Computer Speech and Language: Special Issue on Spoken Language Generation*, volume 16(3–4). Academic Press.

R.L. Weber et al., editors. (1973). *A Random Walk in Science*. Institute of Physics Publishing, Bristol and Philadelphia.

[4]Note that for a given language, we typically only use one of the corpus-based lexicon-extraction methods, depending on the availability of a dependency parser for that language. The combination of these two methods is required for the construction of a multi-lingual lexicon.

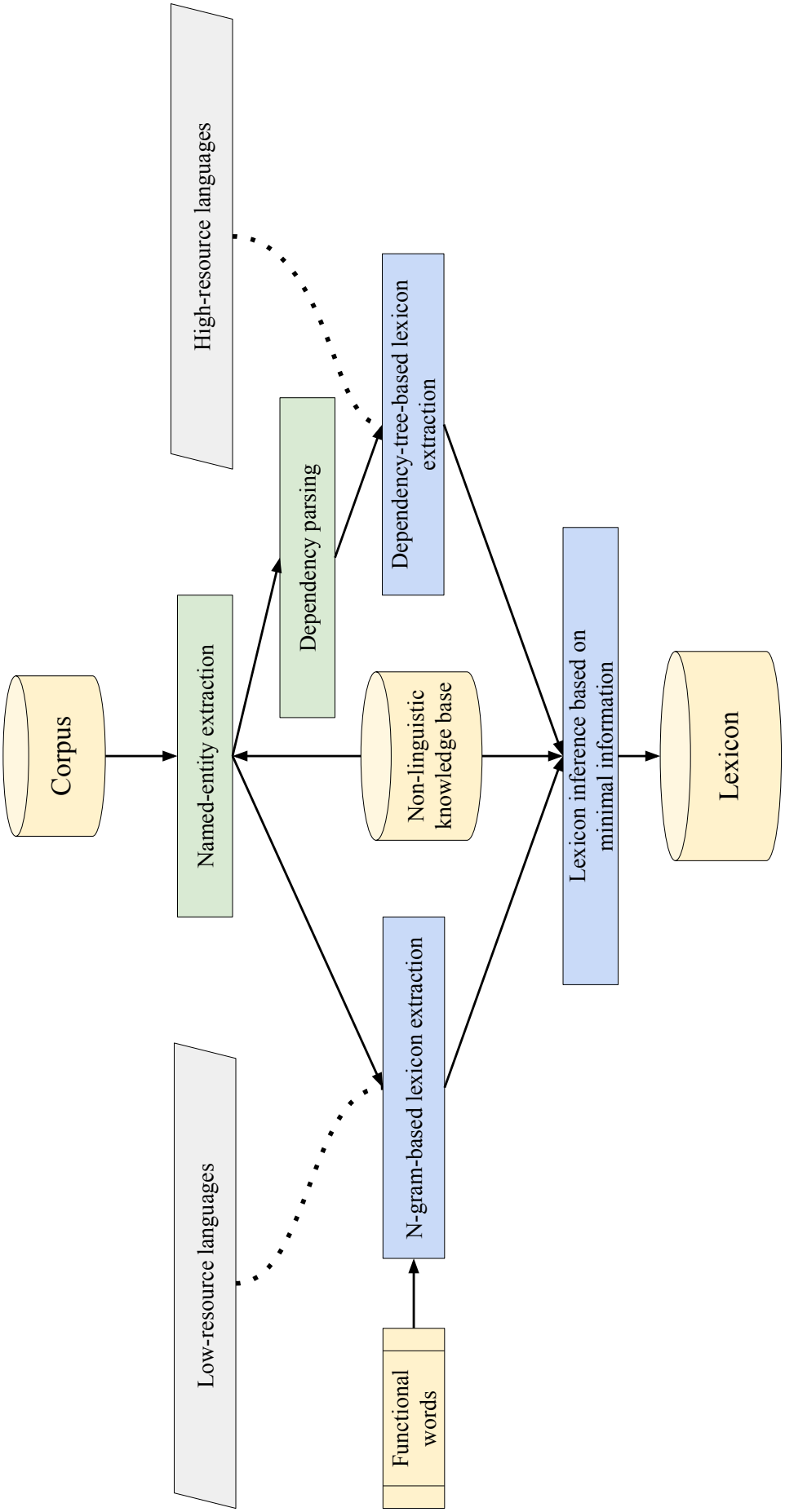



Figure 3: Flow diagram of the proposed architecture for crafting a linguistically annotated lexicon. Note that the lexicon inference based on minimal information is optional.

# Enhancing Usability for Automatically Structuring Digitised Dictionaries

**Mohamed Khemakhem**[†*§], **Axel Herold**[†‡¶], **Laurent Romary**[†*‡]
† Inria - ALMAnaCH, Paris
* Centre Marc Bloch, Berlin
§ Université Paris Diderot
¶ École Pratique des Hautes Études, Paris
‡ Berlin-Brandenburgische Akademie der Wissenschaften, Berlin
{mohamed.khemakhem, axel.herold, laurent.romary}@inria.fr

**Abstract**

The last decade has seen a sharp rise in the number of NLP tools that have been made available to the community. The usability of several e-lexicography tools represents a serious obstacle for researchers with little or no background in computer science. In this paper we present our efforts to overcome this issue in the case of a machine learning system for the automatic segmentation and semantic annotation of digitised dictionaries. Our approach is based on limiting the burdens of managing the tool's setup in different execution environments and reducing the complexity of the training process. We illustrate the possibility to reach this goal by adapting existing functionalities and using out-of-the box software deployment tools. We also report on the community's feedback after exposing the new setup to real users from different professional backgrounds.

**Keywords:** electronic lexicography, usability, digitised dictionaries, TEI, Docker

## 1. Introduction

Web applications have been the main deployment solution for many NLP tool designers to shortcut the need to deal with installation and configuration issues that many desktop applications continue to represent for end users. A web architecture does not rely on the user being familiar with local software tools such as command line shells or software development environments that allow expert and more personalised use of some advanced libraries. A strong current development is the integration of sets of tools into unified web-based working environments for general Humanities research such as the European CLARIN[1] and DARIAH[2] initiatives. In the more specialised field of lexicography, tools such as the Lexonomy[3] dictionary writing system (Měchura, 2017) represent a typical class of web-based applications.
While much of this high level way of accessing NLP tools also accounts for desktop applications, locally installed tools and possibly other software they rely on still have to be updated regularly. Different tools may even form a complex "eco-system" with subtle dependencies between individual modules. The main concern for users with regard to web-based tools is the security and possibly the confidentiality of their data. Therefore desktop applications still exist after the general movement towards web-based solutions.
GROBID-Dictionaries[4] is a machine learning system which has been developed to serve as a web application for structuring digitised dictionaries (Khemakhem et al., 2017). It also exhibits the desktop functionality required for the preprocessing of data during the training process. Although it has a decent documentation, the process of setting up the desktop version of the tool remains very challenging for users with limited programming knowledge. Annotating the preprocessed XML data also represented a serious challenge in earlier versions of the tool because initially it did not provide mechanisms for sanity checks or for visualising annotations for humans.
In this paper we focus on the desktop functionality built into GROBID-Dictionaries. We present new features which have been implemented to enhance the usability of the tool. In Section 2. we provide an overview of the architecture and setup of the system. We detail the different stages of the training process in Section 3. We then address the technical challenges related to the installation of the system as well as the annotation process and present our solution to overcome them in Section 4. In Section 5. we report on first experiences with the new setup and features based on feedback collected from users who were previously not familiar with GROBID-Dictionaries.

## 2. GROBID-Dictionaries

The work carried out by Khemakhem et al. (2017) resulted in a successful adaptation and extension of GROBID – an existing machine learning platform (Lopez and Romary, 2015) – to be used for the automatic identification of lexical information in digitised lexical resources. The resulting system is called GROBID-Dictionaries to reflect the dependency with the parent project. GROBID-Dictionaries has been tested using several lexical resources with promising results.

### 2.1. Architecture

The system's architecture is cascaded. Textual and typographical information are processed by means of multi-level classifications performed by machine learning models.
Figure 1 sums up the architecture described in Khemakhem et al. (2017). Each blue object represents a *conditional random field* (CRF) model. These models are used to classify

[1] https://www.clarin.eu/
[2] https://www.dariah.eu/
[3] http://www.lexonomy.eu/
[4] https://github.com/MedKhem/grobid-dictionaries

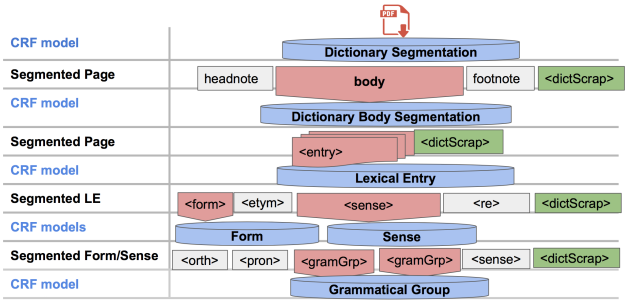



Figure 1: General architecture of GROBID-Dictionaries

the input text together with its typographical features. The other objects represent resulting text clusters to be either directly wrapped into proper TEI elements (elements with angle brackets) or they are temporarily tagged with pivot elements that are transformed into valid TEI constructs only in the final output (e. g., `headnote`, `footnote`, `body`). For the sake of simplicity, Figure 1 does not include all possible tags for the *Form* and *Grammatical Group* models. A complete description of all possible TEI structures resulting from these two models can be found in the TEI P5 dictionary chapter[5][6] in Budin et al. (2012).

## 2.2. Configuration

GROBID-Dictionaries depends on core utilities and libraries provided by GROBID[7]. The installation of the system must be preceded by the installation and setup of the parent project. Therefore GROBID-Dictionaries needs to be cloned as an extension module within GROBID's project structure and must be built after its parent project.
Due to differences in technical preferences of the project leaders, two different automation build technologies need to be used for building each project: Gradle[8] for GROBID and Maven[9] for GROBID-Dictionaries. Successful builds of the system are packaged as Java libraries in two formats:

- a JAR (Java ARchive): this file is required for all processing stages which precede the training of each model, and
- a WAR (Web Application Resource or Web application ARchive): in the case of GROBID-Dictionaries this is not only a standalone web application but also a self-contained one that can be run after the training of the CRF models. It provides a graphical user interface to the existing web services, each corresponding to one or more of the cascading classification models.

GROBID-Dictionaries has been developed, tested and documented for the Linux and Mac operating systems. The behaviour of the resulting libraries is expected to be the same when run on other operating systems. However, there is no explicit guarantee for such uniform behaviour.

## 3. MATTER Annotation Workflow

The annotation workflow in GROBID-Dictionaries follows the MATTER methodology (Model–Annotate–Train–Test–Evaluate–Revise, see Figure 2) introduced by Pustejovsky and Stubbs (2012). Projected onto GROBID-Dictionaries and the processing of lexical resources, the individual steps are as follows:

**Model:** define a CRF model for predicting different text structures at one stage and determine the corresponding feature set. This phase requires the involvement of a programmer to create the defined models and integrate them into the cascading architecture.

**Annotate:** assign a TEI tag to each text block representing a lexical entity defined within a model's scope. This task must be performed on an XML representation of the data and must be strictly synchronised with the corresponding feature set file. The annotation guidelines[10] need to be respected.

**Train:** use each annotated batch of data to train a corresponding model. The cascading architecture of the models should be respected here.

**Test:** this step gives just a rough idea about how the trained model behaves on unseen data. There are many ways to accomplish this goal. The easiest one is to run the corresponding web service from the web application on a held-out sample.

**Evaluate:** a precise evaluation with different measures is enabled at the end of the training process as long as annotated data are provided under the dedicated location in the dataset.

**Revise:** the last stage is about reviewing the modelling and annotation steps that have been described in the guidelines. Four possible measures are the outcome of this step:

- annotate more data when an improvement in the results was achieved,
- refine the annotation guidelines for new variations noticed in the last training batch
- proof-read the performed annotations when minor anomalies are noticed
- think about redefining the modelling when the results represent unexplainable anomalies. This could be translated either into a simple feature engineering process or into a change of the logic behind and the scope of the models or their architecture.

[5] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-form.html`
[6] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-gramGrp.html`
[7] `https://github.com/kermitt2/grobid`
[8] `https://gradle.org`
[9] `https://maven.apache.org`
[10] `https://github.com/MedKhem/grobid-dictionaries/wiki/How-to-Annotate%3F`

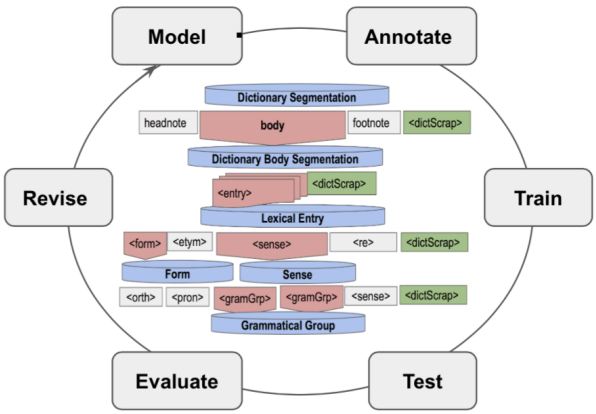



Figure 2: Implemented MATTER Workflow

## 4. Enhanced Usability

Section 2. presented a detailed picture of the technical setup required to install and execute the different parts of the system. Thus it is clear that a certain expertise and understanding of the system architecture is mandatory to successfully install the tool. Section 3. highlighted the challenges of the iterative training cycle which involves costly manual work in terms of carrying out data annotation. Such requirements impose a twofold obstacle: on the one hand, the tool's target community mostly consists of users, such as lexicographers or linguists, who have limited programming skills. If these users are not able to get technical support, the tool will not be usable for a large proportion of its target community. In the other hand, the GROBID-Dictionaries project aims to constantly improve its architecture and to provide more fine-grained lexical information. In the long term, the goal of the project is to provide generic machine learning models which will be able to exploit different types of digitised dictionaries. Collecting and working with different types of lexical data (or at least samples thereof) drawn from a preferably diverse user community is a crucial step in the further development of GROBID-Dictionaries.
The usability of the tool is a vital aspect as this enables a broad user community to productively make use of GROBID-Dictionaries. Therefore, issues of usability are of similar importance to the tool's earlier defined purpose and the research challenges it encounters.

### 4.1. Unified Execution Environment

As a first measure, we have investigated different ways for streamline the setup process and to guarantee a unique behaviour of the system across different execution environments.
One possible solution would have been to use a system image runnable on a virtual machine. Such an image should have a Linux based operating system, a Java development kit (JDK) and the different automated build systems installed. GROBID and GROBID-Dictionaries should also already be cloned and built correctly. This type of solution suffers from two main issues. Firstly, the size of the image would be huge as it would include several unnecessary tools and system files that are still part of the operating system. Secondly, the static nature of such an image would make it complicated to update after a new version of GROBID-Dictionaries is released. Updates to GROBID-Dictionaries are published frequently since the tool is under continuous development. However, a system image containing the above mentioned components can be built in a more efficient way using a different technique. Docker[11] is a state of the art software technology which is also based on the virtualisation of the execution environment. In contrast to the static image approach sketched out initially, Docker allows for the flexible composition of an image. An image is shaped by instructions written in a Docker file[12]. These instructions ensure that only the required components are included in the image. Moreover, several alternatives are available to efficiently update a build within an image starting from pushing a newly created image to the online Docker Hub repository[13], to linking the corresponding GitHub and Docker Hub repositories coupled with activating the automatic build to synchronise the image after each update of the code.

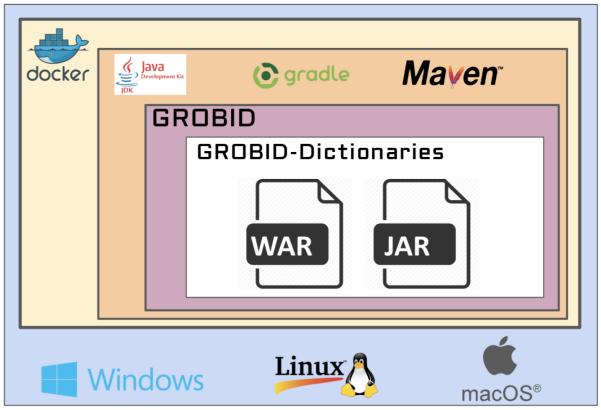



Figure 3: A GROBID-Dictionaries image in a Docker container

To run a Docker image of GROBID-Dictionaries (see Figure 3), a user needs to install the version of the Docker software corresponding to the user's operating system and pull the latest image of the tool from Docker Hub. The pulled image (orange box) will not be run directly on top of the operating system of the host machine but rather inside a Docker controlled container (yellow box). Thus testing the tool on Docker is enough to guarantee a unified behaviour, regardless of the particular system configuration of a user's computer environment.
It is also possible to synchronise files on the host machine with a running image in the Docker container. This feature allows the tool hosted inside a Docker container to directly interact with files stored on the host machine. We took advantage of this alternative to make the dataset directory shared between the two environments. With this mechanism, the user can exploit the full functionality of the tool living

[11] https://www.docker.com
[12] https://github.com/MedKhem/grobid-dictionaries/blob/master/Dockerfile
[13] https://hub.docker.com/r/medkhem/grobid-dictionaries/

in the Docker image to train the machine learning models on the data residing locally on his machine.
In addition, thanks to the self-contained nature of the tool's web application coupled with its fluid setup and manipulation through the Docker image, using the GROBID-Dictionaries image enables both of the desktop and web based functionality to be run on the user's local machine. Such a feature represents an asset for researchers who are concerned about the security of their data and experiments.

## 4.2. Lightening MATTER Process

The second major category of improvements specifically targets the annotation workflow. Annotating data for the training process involves challenging manual work and requires precautionary measures to ensure data integrity and validity.

### 4.2.1. Creating Training Data

To train a model in GROBID-Dictionaries based on a PDF file containing the raw text and the typographical features of a lexical resource, two additional files are necessary: a TEI document containing the corresponding reference encoding and a feature file describing textual and typographical information of each printed line or token.
To generate the training files, embedded functionalities of the tool should be used following one of the two following options:

- *pre-annotated training data:* this used to be the default mode for automatically creating training data, inherited directly from GROBID's core functionality. This mode is useful when a model was trained on a substantial amount of data. The task of the annotator is then to correct the automatically placed TEI tags by moving, adding or removing them.

- *raw training data:* this constitutes new functionality we have implemented to shortcut the checkout and cleaning of the tags automatically generated by using the default mode. The idea is simply to create training data without pre-annotations. Despite being obvious, starting to annotate a document from scratch was not possible before integrating this new feature. Such a mode breaks with the old practice of correcting the predictions made by a model trained on different samples, to make it possible to start annotating totally fresh data. Besides giving more choices to the annotator, such a mode saves time and efforts especially if an old model was trained with multiple TEI elements.

A legitimate question remains as yet unanswered: how can a user generate training data based on a selection of specific pages from possibly hundreds of pages a dictionary may comprise? After annotating different lexical samples in PDF format, we could qualify splitting an existing document into separate pages, or sequences of pages, as a very critical step. With some supposedly dedicated PDF manipulation tools producing damaged pages, we found only one tool reliably useful for the purpose of separating PDF pages[14] which seems to produce a quality split as good as the original document. Using workaround solutions for this purpose, such as the print-to-file functionality in web browsers, is also not recommended.

### 4.2.2. Training Data Annotation

As previously stated, GROBID-Dictionaries generates a pre-processed XML representation from PDF files containing the raw text of a lexical resource. To create training data for the tool, the user is then required to introduce semantic mark-up for the different models. Typically, an XML aware editor should be used to perform this task. Some advanced editors such as oXygen[15] allow for the visual annotating of XML files (see Figure 4 for an example).
We aimed take advantage of the visual feature to avoid performing inline annotation directly on the text of the XML elements. This is catered for by a new feature in GROBID-Dictionaries that for each model now provides both a schema description (in Relax NG)[16] and a presentational stylesheet (in CSS). The schema description enables the editing software to check or even enforce schema compliance of the training data. The stylesheet can be exploited by the editing software to allow users to mark up the training data semantically by highlighting portions of the text and then enclosing the highlighted portion with a suitable XML tag. The colours attributed to each element can be customised by a simple modification in the stylesheet.

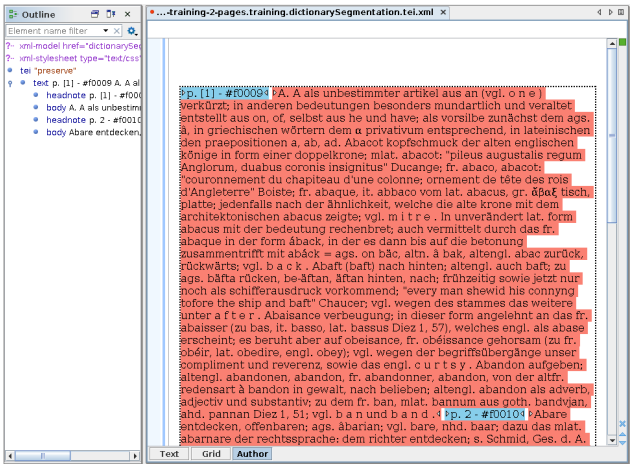



Figure 4: Training data annotation in oXygen author mode for the first model: page headers vs. page body

### 4.2.3. Train, Test and Evaluate

For this segment of the MATTER workflow, the user is provided with straightforward shell commands to execute, a graphical mode to test and varied measures to evaluate and decide whether a model has reached an acceptable level of accuracy. A simple but effective trick could however be employed at this stage to verify the accuracy of the annotations performed in the previous step. Where in a normal case the annotated data should be split between training and evaluation datasets, the training dataset could be also used as an evaluation dataset to verify any inconsistencies that might have accrued during the annotation process. In

[14] http://community.coherentpdf.com

[15] https://www.oxygenxml.com/

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

such a setup, a correct annotation should give 100 % accuracy, which means that model could reproduce what it has learnt correctly. Any other result should lead to the last step described in Section 2.

## 5. User Experience

We had the opportunity to expose the system with its new setup and features to a mixed group of users in the course of a winter school on lexicography that was held at the Berlin-Brandenburg Academy of Sciences and Humanities at the end of 2017[17]. During this event we collected information about the usability of the tool. Additionally, we asked participants to respond to a questionnaire after the winter school to gain further insight into their experience of working with the tool. Given the relatively small number of participants, the responses to the questionaire do not allow for a rigid quantitative evaluation. Nevertheless, based on the responses and our own experiences during the tutorial we are able to present a qualitative evaluation.

### 5.1. Setup

A group of nine users participated in the experiment which was carried out during three hands-on sessions of four hours each. The users were free to join one or more sessions of the tutorial. The goal of the tutorial was to familiarise the participants with the MATTER workflow as implemented in GROBID-Dictionaries, while excluding the first modelling step which requires programming skills. Note that none of the participants was familiar with the tool prior to the tutorial.

After a short introduction to the architecture of the system, the users were guided through the process of installing and running the docker image[18]. Once the docker image was running, the participants were then able to reproduce the results reported in Khemakhem et al. (2017) which are based on a modern English monolingual dictionary. As the next step, several users used the possibility to experiment with their own lexical samples by repeating the workflow they had learnt and crafting new models for their individual datasets. Two of the participants succeeded in training and using all of the implemented models for their own datasets, thus adapting all of the functionality currently implemented in GROBID-Dictionaries.[19]

### 5.2. Gathered Insights

We asked the participants of our tutorial to respond to a questionnaire after the winter school. The questionnaire was created as a Google Form[20]. The results of the inquiry can be summed up by the following points:

**Tool setup / user profile** The first three questions focus on establishing the professional background of the participants. The tutorial group consisted of lexicographers, linguists, computational linguists, a computer scientist, a web developer and a philologist. Participants were free to name more than one field of expertise. Of the nine respondents, seven reported previous knowledge of machine learning techniques but only four of them had actually worked with machine learning tools before.

When asked whether they encountered any problems with actually running the tool from the docker image, the majority of the participants (seven) responded that this was not the case. The setup failed once on a Windows based computer with insufficiently sized memory that was running an advanced version of the operating system. Consequently there was not enough memory left to run the Docker software which requires more than the 1 GB of free memory. The participant could still continue the tutorial by sharing a machine with her colleague. Without taking into account the answer of another respondent who involuntarily reported encountering an installation issue, almost 90% of the users were able to launch the tool without any problem.

**Sample data / Initial training** The lexical resources brought to the tutorial were considerably varied. They included different types of dictionaries (some digitised, some born digital with no explicit semantic markup) such as general monolingual, bilingual and etymological dictionaries as well as a dictionary from a language documentation field project (see Table 1).

We asked the participants whether they successfully trained at least the first two models and thus were able to perform the general dictionary segmentation (page segmentation) and the dictionary segmentation (entry recognition). Despite the variety of their datasets, 100% of the answers were positive. This supports the assumption of the implemented cascading approach to be sample independent.

| Type | Language(s) | Size |
|---|---|---|
| general, bilingual | Greek, English | ≈ 17 000 entries |
| general, monolingual | Basque | ≈ 16 000 pages |
| etymological, bilingual | Hittite (a languages of the ancient Near East), English | ≈ 470 pages |
| lang. documentation | French, Yemba (an African language family) | ≈ 2 1000 entries |
| lang. documentation | German (Bavarian dialects in Austria) | ≈ 75 000 entries |
| general, monolingual | English | ≈ 370 pages |

Table 1: Dictionaries experimented with during the tutorial. Note that two participants worked on the same resource and another two used the resource that we provided.

[17] https://lexmc.sciencesconf.org/

[18] see instructions at https://github.com/MedKhem/grobid-dictionaries/wiki/Docker_Instructions

[19] A more detailed description of the conditions of the experiment can be found in a blogpost at https://digilex.hypotheses.org/250 as shared by of one of the participants.

[20] https://goo.gl/Zt2gDy

**Creating training data** Two questions focus on the usability of the graphical annotation of the training data using oXygen's author mode. None of the participants found graphically marking the training data a hard task and six described it as a straightforward process. Compared to creating the training data by manipulating the XML structure directly with a text editor, most of the participants (seven) confirmed that the graphical approach was easier.

**Training workflow** Although just two participants could finish the training for all models of the tool, all those who were not able to train the remaining models during the tutorial expect to be able to complete the training on their own. Moreover, all the participants reported being confident that they were able to re-apply what they had learnt on other lexical resources. It's important though to clarify why some users could not successfully train all of the models until the end of the tutorial. This was mainly due to the fact that the participants were free to attend only parts of the tutorial sessions and due to the considerably long time spent downloading the huge Docker image with the available internet connection.

**Future use of the tool** Based on the apparently successful mastering of the training workflow, all but one participant were willing to continue using GROBID-Dictionaries after the tutorial. It is worth noting that the participant who does not intend to continue using GROBID-Dictionaries is working with non-lexical data and still plans to adapt the parent project GROBID to his type of data.

Having motivated inter-disciplinary experts participating in the tutorial as well as testing the tool on new lexical samples provided us with the opportunity to spot some issues and several possible improvements. We were able fix some of the minor triggered implementation issues in the course of the tutorial. Other issues have been filed as new tickets on GitHub, e. g. issues concerning the treatment of lexical entries that stretch over more than two pages in print. Some technical issues related to the GROBID core still need to be resolved such as support for some classes of special characters which are wrongly encoded in the preprocessing of the raw input text. The annotation guidelines should also be further refined to provide clearer definitions of constructs to be annotated, such as related entries.

## 6. Conclusion

Whereas Khemakhem et al. (2017) presented the basis of the approach to implement GROBID-Dictionaries and initial experimental results, this paper provides a more in-depth description of the machine learning system, with the focus on its architecture, technical setup and the training workflow. Enhancing the usability of the tool has been addressed as a fundamental feature given the fact that the tool is in its early development stage and the involvement of end users is a key factor in the evolution of the tool. Therefore several measures have been implemented to guarantee a straightforward installation and user-friendly annotation process. The exposure of the tool to real users has confirmed many of our choices to alleviate the challenges of a complex ML workflow. This experiment also provided us with the possibility to promote the tool as well as to collect in-depth feedback, which will help us to efficiently set our priorities. The recent version and setup of the tool, presented in this paper, does not only enhance its usability but also supports the reproducibility of findings resulting from its use.

## 7. Acknowledgements

This work has been supported by the "Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies" (PARTHENOS) project. We would like to thank Karim Ben Ammar B.Sc. for his valuable technical advice regarding the efficient use of the Docker technology.

## 8. References

Budin, G., Majewski, S., and Mörth, K. (2012). Creating lexical resources in tei p5. a schema for multi-purpose digital dictionaries. *Journal of the Text Encoding Initiative*, (3).

Khemakhem, M., Foppiano, L., and Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In *electronic lexicography, eLex 2017*, Leiden, Netherlands, September.

Lopez, P. and Romary, L. (2015). Grobid - information extraction from scientific publications. *ERCIM News*.

Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In *electronic lexicography, eLex] 2017*, Leiden.

Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

# The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches

**Pilar León-Araúz, Antonio San Martín**
University of Granada, University of Quebec in Trois-Rivières
Buensuceso, 11 18001 Granada (Spain), 3351, boul. des Forges, Trois-Rivières (Quebec, Canada)
pleon@ugr.es, antonio.san.martin.pizarro@uqtr.ca

**Abstract**

Many projects have applied knowledge patterns (KPs) to the retrieval of specialized information. Yet terminologists still rely on manual analysis of concordance lines to extract semantic information, since there are no user-friendly publicly available applications enabling them to find knowledge rich contexts (KRCs). To fill this void, we have created the KP-based EcoLexicon Semantic Sketch Grammar (ESSG) in the well-known corpus query system Sketch Engine. For the first time, the ESSG is now publicly available in Sketch Engine to query the EcoLexicon English Corpus. Additionally, reusing the ESSG in any English corpus uploaded by the user enables Sketch Engine to extract KRCs codifying generic-specific, part-whole, location, cause and function relations, because most of the KPs are domain-independent. The information is displayed in the form of summary lists (word sketches) containing the pairs of terms linked by a given semantic relation. This paper describes the process of building a KP-based sketch grammar with special focus on the last stage, namely, the evaluation with refinement purposes. We conducted an initial shallow precision and recall evaluation of the 64 English sketch grammar rules created so far for hyponymy, meronymy and causality. Precision was measured based on a random sample of concordances extracted from each word sketch type. Recall was assessed based on a random sample of concordances where known term pairs are found. The results are necessary for the improvement and refinement of the ESSG. The noise of false positives helped to further specify the rules, whereas the silence of false negatives allows us to find useful new patterns.

**Keywords:** EcoLexicon Semantic Sketch Grammar, knowledge patterns, sketch grammars, semantic relations, Sketch Engine

## 1. Introduction

Terminologists rely on corpus analysis for the extraction of conceptual information because most of the knowledge shared by experts is expressed in texts (Bourigault & Slodzian, 1999). For a long time, the only accessible way of analyzing corpus information for terminological work consisted in manually reading concordance lines. This is time-consuming and inefficient because for a given term a terminologist can be confronted with thousands of concordance lines, many of which may not carry any useful information for the terminologist.

Useful concordance lines for conceptual analysis are called knowledge-rich contexts (KRCs) (Meyer, 2001) and one of the most common approaches to find them is to search for knowledge patterns (KPs) in corpora. KPs are the linguistic and para-linguistic patterns that convey a specific semantic relation in real texts (Meyer, 2001). For instance, some of the simplest examples of generic-specific KPs are *x is a kind of y*, *As include Bs, Cs and Ds* (Meyer, 1994) and *comprise(s), consist(s), define(s), denote(s), designate(s), is/are, is/are called, is/are defined as, is/are known as* (Pearson, 1998).

KPs are considered one of the most reliable methods for the extraction of semantic relations (Auger & Barrière, 2008; Barrière, 2004; Bowker, 2003; Cimiano & Staab, 2005; Condamines, 2002; L'Homme & Marshman, 2006; Lafourcade & Ramadier, 2016; Lefever, Kauter, Hoste, Van de Kauter, & Hoste, 2014; Marshman, 2002, 2014; Marshman, Morgan, & Meyer, 2002). They have been applied in many terminology-related projects leading to the development of knowledge extraction tools, such as Caméléon (Aussenac-Gilles & Jacques, 2008) and TerminoWeb (Barrière & Agbago, 2006).

However, no user-friendly application allowing terminologists to find KRCs in their own corpora is publicly available. For this reason, in León-Araúz, San Martín & Faber (2016), we created a KP-based sketch grammar for Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) with the intention of allowing other users to extract KRCs through word sketches from their own corpora previously compiled with our grammar, which is mostly domain-independent.

Word sketches are defined as automatic corpus-derived summaries of a word's grammatical and collocational behavior (Kilgarriff et al., 2004). Rather than looking at an arbitrary window of text around the headword—as occurs in previous corpus tools—Sketch Engine is able to look for each grammatical relation that the word participates in (Kilgarriff et al., 2004). The default word sketches provided by Sketch Engine represent different relations, such as verb-object, modifiers or prepositional phrases. However, except for the recently implemented generic-specific word-sketches, they only represent linguistic relations. Figure 1 shows an example of three default word sketches in Sketch Engine.

| nouns modified by "bird" | | 24.17 |
|---|---|---|
| specie + | 17,353 | 9.09 |
| flu + | 11,801 | 9.82 |
| feeder + | 11,199 | 9.80 |
| watching + | 10,862 | 9.75 |
| sanctuary + | 6,860 | 8.86 |
| life + | 6,168 | 5.24 |

| verbs with "bird" as object | | 21.89 |
|---|---|---|
| be + | 42,519 | 2.88 |
| see + | 14,373 | 5.05 |
| kill + | 9,920 | 7.53 |
| have + | 7,084 | 1.77 |
| watch + | 6,672 | 6.26 |
| find + | 6,307 | 4.08 |

| verbs with "bird" as subject | | 21.15 |
|---|---|---|
| be + | 88,426 | 3.38 |
| have + | 22,998 | 3.40 |
| fly + | 10,469 | 8.89 |
| watch + | 8,554 | 8.19 |
| sing + | 7,164 | 8.74 |
| do + | 7,039 | 3.59 |

Figure 1. Example of word sketches for *bird* in the English Web 2013 (enTenTen13) corpus

In León-Araúz, San Martín & Faber (2016), we developed 64 new sketch grammar rules focusing on the extraction of semantic relations, expanding the functionality of word sketches to the summarized representation of semantic

behavior. This new sketch grammar for the English language includes some of the most common semantic relations used in the field of terminology: generic-specific, part-whole, location, cause and function. For the first time, this sketch grammar is now publicly available under the name of the EcoLexicon Semantic Sketch Grammar (ESSG). It is built in Sketch Engine to query the EcoLexicon English Corpus (see section 3.1), but users can also reuse it with any corpus following the instructions on <http://ecolexicon.ugr.es/essg>.

This paper describes the process of building a KP-based sketch grammar with special focus on the last stage, namely, the evaluation with refinement purposes. We conducted a shallow precision and recall evaluation of the 64 English sketch grammar rules created so far for hyponymy, meronymy and causality, which are an expansion and refinement of the ones presented in León-Araúz, San Martín & Faber (2016).

## 2. Building a KP-based sketch grammar

Although some authors (Marshman, 2004; Meyer, 2001) have inventoried patterns, they normally are a simplification of what is actually found in a corpus. For instance, when formalizing the pattern *is a type of* we should also take into account all of its possible variants. The verb *to be* may be in its plural form or substituted by a comma; if it is in the plural, various hyponyms will be enumerated to the left of the pattern; the verb *to be* may be preceded by a modal verb; the word *type* may be preceded by an adjective and an adverb; and it may be substituted by other synonyms such as *kind*, *sort*, *example*, *group*, etc. All of these possible variations must be accounted for when developing the grammar rules.

Corpus querying in Sketch Engine is based on an extension of the Corpus Query Language (CQL) formalism (Jakubíček, Kilgaiff, McCarthy, & Rychlý, 2010), allowing for the formalization of grammar patterns in the form of regular expressions combined with POS-tags. CQL expressions in Sketch Engine can be used as one-time queries (giving access to matching concordance lines) or stored in a sketch grammar, which will produce word sketches. For instance, if users query "[tag="JJ.*"] [lemma="energy"]" in SketchEngine, they will obtain all the concordances in which *energy* is preceded by an adjective in the corpus of their choice. For their part, sketch grammars are collections of CQL expressions that allow users to produce word sketches without any knowledge of the CQL formalism. A single word sketch may be the result of a combination of multiple long CQL expressions.

In the development of the ESSG we have considered different issues that are specific to each relation. For instance, there are certain patterns that always take the same form and order (e.g. *such as*), whereas others show such a diverse syntactic structure that the directionality of the pattern must also be accounted for. We also had to take into account the fact that a single sentence could produce more than one term pair because of the enumerations that are often found on each side of the pattern (e.g. *x, y, z and other types of w*). This entails performing greedy queries in order to allow any of the enumerated elements fill the target term. However, this may also cause endless noisy loops. Sometimes it is necessary to limit the number of possible words on each side of the pattern. In this sense, we observed that enumerations are more often found on the side of hyponyms, parts, and effects than on the side of hypernyms, wholes, and causes. Consequently, the loops were constrained accordingly in the latter case. Table 1 shows a summarized and simplified version of the patterns included for each semantic relation evaluated in this study (only a sample of 5 patterns per semantic relation for space reasons).

| |
|---|
| **Generic-specific:** HYPONYM ,|(|:|is|belongs (to) (a|the|…) type|category|… of HYPERNYM // types|kinds|… of HYPERNYM include|are HYPONYM // types|kinds|… of HYPERNYM range from (…) (to) HYPONYM // HYPERNYM (type|category|…) (,|() ranging (…) (to) HYPONYM // HYPERNYM types|categories|… include HYPONYM |
| **Part-whole:** WHOLE is comprised|composed|constituted (in part) of|by PART // WHOLE comprises PART // PART composes WHOLE // PART is|constitutes (a|the|…) part|component|… of WHOLE // WHOLE has|includes|possesses (…) part|component|… (,|() (:|such as|usually|namely|…) PART // WHOLE has|includes|possesses (a|the|…) fraction|amount|percent… of PART |
| **Cause:** CAUSE (is) responsible for EFFECT // CAUSE causes|produces|… EFFECT // CAUSE leads|contributes|gives (rise) to EFFECT // CAUSE-driven|-induced|-caused EFFECT // EFFECT (is) caused|produced|… by|because|due (of|to) CAUSE |

Table 1: Simplified version of the patterns included in each grammar

By way of example, Table 2 shows the actual CQL representation of a generic-specific KP-based rule, followed by an explanation and three natural language examples of concordances matched with the grammar.

| | |
|---|---|
| 1:"N.*" [word=",|\("]? [tag="IN/that|WDT"]? "MD"* [lemma="be|,|\("] "RB.*"* [word="classified|categori.ed"] ([word="by"] [tag!="V.*"]+)? [word="in|into"] [tag!="V.*"]* [lemma="type|kind| example|group|class| sort|category|family|species|subtype| subfamily|subgroup| subclass|subcategory|subspecies"]? [tag!="V.*"]* 2:[tag="N.*" & lemma!="type|kind|example| group|class| sort|category|family|species|subtype|subfamily|subgroup| subclass|subcategory|subspecies"] | |
| 1:"N.*" | The hypernym is a noun. |
| [word=",|\("]? | An optional comma or bracket. |
| [tag="IN/that|WDT"]? | Optionally "that" or "which". |
| "MD"* | Any modal verb from zero to infinite times. |
| [lemma="be|,|\("] | Lemma "be" or a comma or a bracket. |
| "RB.*"* | Any adverb from zero to infinite times. |
| [word="classified|categori.ed"] | Classified, categorized, or categorized. |
| ([word="by"] [tag!="V.*"]+)? | Optionally, "by" followed by anything from one to infinite times that does not contain a verb. |
| [word="in|into"] | In or into. |
| [tag!="V.*"]* | Anything from zero to infinite times that does not contain a verb. |
| [lemma="type|kind| example|group|class|sort|category|family|species|subtype| subfamily|subgroup|subclass|subcategory|subspecies"]? | Optionally any of the lemmas "type", "kind", "example", "group", "class", "sort", "family", etc. |
| [tag!="V.*"]* | Anything from zero to infinite times that does not contain a verb. |
| 2:[tag="N.*" & lemma!="type|kind|example| group|class|sort|category|family |species|subtype|subfamily| subgroup|subclass|subcategory| | The hyponym is any noun other than "type", "kind", "example", "group", "class", "sort", "family", etc. |

| | |
|---|---|
| subspecies"] | |
| Stony-iron meteorites are classified into pallasites and mesosiderites.<br>Modern reefs are classified into several geomorphic types: atoll, barrier, fringing, and patch.<br>Littoral materials are classified by grain size in clay, silt, sand, gravel, cobble, and boulder. | |

Table 2. CQL representation of a generic-specific KP-based rule followed by its explanation

For the development of sketch grammar rules we followed the following methodology:

1. *Collection of KPs*: this first stage only includes the collection of patterns in plain English (no formalism or encoding language used).

-Patterns referenced by other authors.
-Patterns already known.
-Recursive method: term pairs linked by already known semantic relations are searched for to find new patterns. Then these patterns are used to find new term pairs, and so on.

2. *CQL encoding*: it consists of translating the KPs collected during the first stage into CQL sketch grammar rules.

-Splitting or lumping: some KPs collected in the first stage can be lumped into a single CQL sketch grammar rule, while others collected as a single KP need to be split.
-Addition of adverbs, punctuation, modal verbs, relative phrases, adjectives, determiners, etc.

3. *Enrichment and refining*: CQL rules are enriched and refined trying to keep the balance between noise and silence.

-Enrichment: Testing the CQL rules with additional optional elements to spot new variations of the pattern (for instance, the possibility of an adverb in a place where it was not previously accounted for).
-Refining: Detection of erroneous concordance lines obtained with the CQL rules. Analysis of the source of the error, and determination of whether it is appropriate to change the CQL rule.

4. *Evaluation*: this includes a precision and recall analysis, which is described in section 3.2. After the evaluation, the enrichment and refining step is repeated to include the new patterns and modifications that the analysis of noise and silence has proved necessary.

# 3. Evaluation of the ESSG

## 3.1 EcoLexicon English Corpus

For evaluating the ESSG, we applied them to the EcoLexicon English Corpus (EEC). The EEC is a 23.1-million-word corpus of contemporary environmental texts compiled by the LexiCon Research Group for the development of the environmental terminological knowledge base EcoLexicon (Faber & Buendía, 2014; Faber, León-Araúz, & Reimerink, 2016; San Martín et al., 2017)[1]. It can be queried within the knowledge base, but the corpus has also recently been made freely available in Sketch Engine Open Corpora[2]. Each text in the EEC is tagged according to a set of XML-based metadata. This allows constraining corpus queries based on pragmatic factors such domain, user, geographic variant, genre, editor, year and country of publication.

The EEC is tagged with the Penn Treebank tagset (TreeTagger version) ver. 3.3, which allows for more fine-grained queries in CQL. It employs the default sketch grammar for English in combination with the ESSG In this way, word sketches in the EEC incorporate automatic corpus-derived summaries of a concept's semantic relations (Figure 2). Thus, the aim of our sketch grammar is twofold: (1) offering semantic word sketches in our freely available EEC, (2) and providing other users (i.e. terminologists) with the possibility of reusing it in their own corpora.

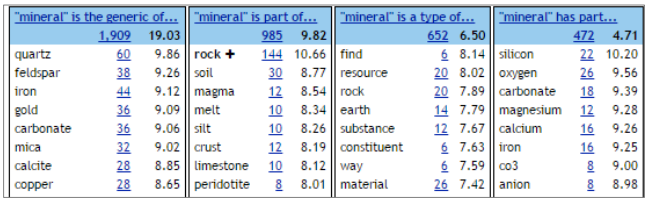



Figure 2. Word sketches of *mineral* in the EEC extracted with the ESSG

## 3.2 Precision and recall metrics

Precision is measured on a random sample of concordances of one of the terms that has most frequently been annotated as part of each word sketch. This leads to the identification of false positives and the analysis of their causes, which results in the refinement of sketch grammar rules. Given that at this stage the goal of the evaluation was to use the results to improve our sketch grammar before objectively assessing their global efficiency as knowledge extraction devices, we chose to analyze only the results of one particular term. This allowed us to reduce the workload of the evaluation process. Moreover, since sketch grammars are conceived for the compilation of word sketches that users might find interesting to look at, the keyword is chosen based on a term susceptible to being queried, avoiding, for instance, top-level concepts.

Recall, in turn, is measured on a random sample of concordances where the most frequent term pair is found, enriching the grammar rules through the identification of new useful KPs based on the false negatives encountered. The recall analysis is based on a particular term pair because that makes having a sample of manually curated positive concordances viable. The steps for each measure are as follows. Steps from 1 to 3 are common to both, with the only difference that for the precision analysis we select one particular term and for the recall analysis we select a particular term pair.

1. All concordances where each relation has been annotated are retrieved. For example, for the hyponymic relation the query [ws(".*-n","\"%w\" is a type of...",".*-n")] provides all the results where hypernyms and

[1] ecolexicon.ugr.es/

[2] the.sketchengine.co.uk/open

hyponyms (variables 1 and 2) have been annotated while compiling the corpus.

2. The results are sorted by frequency with Sketch Engine's functionality Node form, showing the terms/term pairs that have most frequently annotated as one/both of the variables.

3. One of the most frequently annotated terms/term pairs is selected avoiding top-level concepts (i.e. *factor*, *parameter*) and terms that usually act as a modifier. Given the fact that users will query word sketches to find meaningful term pairs, we considered that broad top-level concepts are markedly less susceptible of being searched and thus we did not select them. Terms usually acting as modifiers were avoided as well since sketch grammars can only find single-word terms as variables for the moment.

*Precision:*

4. A sample of 1000 randomized concordances of the selected term is analyzed in order to quantify true and false positives.

5. The causes of false positives are analyzed and further constraints are defined in order to refine the grammar rules.

*Recall:*

4. A new query is performed in order to find all contexts where the pair occurs. For instance, the query (meet [lemma="wind"] [lemma="wave"] -15 15) within <s/> provides all contexts within the same sentence where *wind* and *wave* are found in a word span of ±15.

5. From a randomized sample of 1000 concordances, we manually select all explicit occurrences of the relation in question, whether it is through KPs covered by the grammar or not.

6. A subcorpus is created based on the selected concordances, where we again perform the query in step 1 and then apply a negative filter. In this way, all concordances filtered are the ones that have not been identified by the grammar (false negatives).

7. The causes of false negatives are analyzed and further patterns are found in order to enrich the grammar.

## 4. Evaluation Results and Enhancement of the ESSG

The keywords selected for the precision analysis are: *species*, as a hypernym; *rock*, as a part; and *erosion*, as an effect. The term pairs selected for the recall analysis are: *breakwater-structure*, for hyponymy; *mineral-rock*, for meronymy; and *wind-wave* for causality. The concordances were extracted from the EEC.

As shown in Figure 3, hyponymic rules for *species* as a hypernym are 69.5% precise, whereas meronymic and causality rules scored 71.4% and 55.2% respectively. Recall was 45.2% for the hyponymic pair, 65% for the meronymic pair and 60% for the causality pair. Meronymic rules are thus the ones that perform better in terms of both precision and recall. Causal rules score better results for recall than for precision.

Considering that Sketch Engine only displays statistically relevant word sketches, the precision rate reached by the ESSG seems good enough to get reasonable results when users query the corpus to get semantic word sketches, such as those shown in Figure 1. The study of false positives (Section 4.1) and false negatives (Section 4.2) will contribute to the improvement and refinement of the grammar.

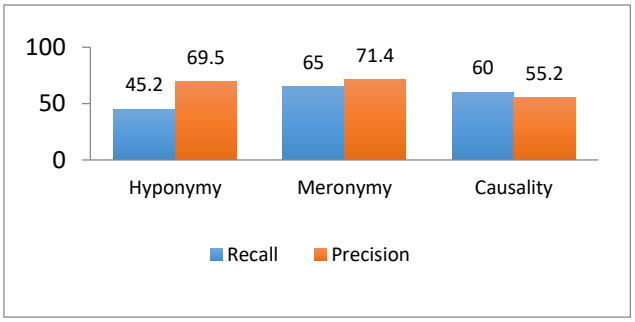



Figure 3. Precision and recall of hyponymic, meronymic and causality sketch grammar rules

### 4.1 Precision: analyzing false positives

Some FPs are caused by inherent limitations of using KP-based extraction of semantic relations with word sketches. Thus, we currently have no way of avoiding them.

1. POS-tagger mistake (mainly, tagging verbs as nouns).
   *…other species, especially those growing in natural ecosystems…*
2. Polysemous keywords: word sketches are unable to perform word sense disambiguation. Consequently, if the keyword is polysemous, the word sketch will show the results of all the senses combined (e.g. *species* as the hypernym of chemicals).
   *…scavenge the reactive oxygen species, including superoxide and hydrogen peroxide…*
3. The cause is a clause, not a noun.
   *They also trampled and over-grazed land, causing erosion and...*
4. Error induced by anaphora.
   *… a Dimilin–propanil mixture on these and other nontarget aquatic species.*
5. A correct relation is detected by mistake.
   *For Caulerpa taxifolia, the other Mediterranean invasive Caulerpa species, a decrease in specialist grazers such as Mullus surmuletus…*
6. The relation is only correct if transitivity is applied.
   *The basement to the arc is made up of at least 3000 m of Triassic (about 240 Ma) sedimentary rock…*

There are other types of FP that can be completely or partially avoided by refining our sketch grammar:

7. The detected hyponym/part/cause is a general word used as part of the pattern itself (i.e. *type*, *part*, *cause*).
   *More than a dozen Queensland frog species, especially the stream-dwelling types…*

All hyponymic grammar rules could be refined by negating for both variables (i.e. hyponym and hypernym)

the words that are used as anchoring words in the patterns. For instance, the rule that caused this FP could be transformed as follows (changes are highlighted in red): 1: [tag="N.*" & lemma!="type|kind|example|group|class|sort|category|family|species|subtype|subfamily|subgroup|subclass|subcategory|subspecies"] [word=",|\("] [word="especially|primarily|namely|usually|typically|characteristically|generally|mainly|particularly|chiefly|mostly|principally"] [tag!="V.*|IN"]* 2: [tag="N.*" & lemma!="type|kind|example|group|class|sort|category|family|species|subtype|subfamily|subgroup|subclass|subcategory|subspecies"]

8. Wrong detection of noun phrase.
   *…populations of the same or closely related species by a physical barrier such as a large river or...*
9. Error induced by the fact that the right elements of the pair are separated by too many words.
   *Streaming winds and following seas toppled expensive summer cottages into the surf, scrubbed the wooden-shingled roofs from quaint boutiques and restaurants, and caused extensive dune erosion.*

The solution in these cases (8 and 9) mostly lies in constraining very long loops. For instance, as mentioned above, in order to find enumerations of different terms at the left and right of the patterns we included broad loops such as [tag!="V.*"] (any word not being a verb). Instead, we should specify how enumerations are usually codified. With [tag="DT|RB.*|JJ.*|N.*" |word="and|or|,|;"]{0,10} we could gain in precision. However, an analysis will be needed to determine whether we would lose recall.

10. Error induced by a relative clause.
    *Ice sheets that form during glaciations cause erosion…*
    In this case, introducing relative clause markers (i.e. *that*, *which*) as a compulsory element between variables 1 and 2 would enhance causal grammar rules.
11. Error induced by negative sentences.
    *…water to enter into the test section from the head tank without causing immediate erosion and…*
    Constraints should be added to easily filter out these matches, adding a list of negative words (*never*, *without*, *no*, *not*, etc.) to all grammar rules.

### 4.2 Recall: analyzing false negatives

As a result of the recall analysis, the following patterns will be updated (changes are highlighted in gray):

- HYPERNYM ,|( such as|like (a|the|…) HYPONYM
- (a|the|one|two|some|…) part|component|building block… of WHOLE (is) called|referred… (to) (as) PART
- (a|the|one|two|some|…) part|component|building block… of WHOLE is PART
- PART (,|() (a|the|…) part|component|building block… of WHOLE
- PART (is) contained|present in WHOLE
- PART composes|constitutes|makes (up) WHOLE
- PART is|constitutes (a|the|…) part|component|building block… of WHOLE
- CAUSE causes|produces|creates… EFFECT
- EFFECT (is) caused|produced|created… by|because|due (of|to) CAUSE

The following are new patterns encountered during recall analysis, some of which might be integrated into existing patterns:

- major HYPERNYM is|include HYPONYM
- HYPERNYM (is) used as HYPONYM
- HYPERNYM serve|act as HYPONYM
- HYPERNYM ,|( e.g. |viz (a) HYPONYM)
- HYPONYM or any ADJ and ADJ HYPERNYM
- HYPERNYM (HYPONYM…
- HYPERNYM: HYPONYM
- HYPERNYM, these being HYPONYM
- WHOLE (is) rich in PART
- PART-rich WHOLE
- WHOLE is an aggregate of PART
- WHOLE and|or its part|component|… PART
- PART in|within WHOLE
- WHOLE with a proportion of PART
- percentage of WHOLE in PART
- EFFECT is the product of CAUSE
- CAUSE acts as generator of EFFECT
- CAUSE acts to cause|produce|create… EFFECT
- CAUSE contributes to the generation of EFFECT
- EFFECT generation by|due to CAUSE
- generation of EFFECT by|due to CAUSE

## 5. Conclusions and future work

The evaluation performed on the ESSG has shown that even a shallow precision and recall analysis is an efficient way of detecting ways of refining and enriching the sketch grammar. Additionally, although the ultimate purpose of the evaluation was not to assess the global performance of the ESSG, the results suggest that the combination of word sketches with KPs has the potential of providing a reliable user-friendly method for the extraction of semantic relations in specialized corpora. Nonetheless, the evaluation indicates as well that there is still room for improvement as far as the level of precision and recall is concerned.

In future work, a larger evaluation study of all of our refined sketch grammar rules will be performed. This will include the study of each relation with no keyword limitations, the assessment of each pattern separately and the evaluation of word sketch precision for multiple term types. In addition to incorporating the improvements revealed by the precision and recall evaluations, the ESSG in the EEC will be enhanced by the inclusion of multiword terms based on those contained in the knowledge base EcoLexicon (by means of corpus annotation) and new collocation rules.

## 6. Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness.

## 7. Bibliographical References

Auger, A., & Barrière, C. (2008). Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*, *14*(1), 1–19. http://doi.org/10.1075/term.14.1.02aug

Aussenac-Gilles, N., & Jacques, M.-P. (2008). Designing and evaluating patterns for relation acquisition from texts with Caméléon. *Terminology*, *14*(1), 45–73. http://doi.org/10.1075/term.14.1.04aus

Barrière, C. (2004). Knowledge-Rich Contexts Discovery. In *Seventeenth Canadian Conference on Artificial Intelligence (AI'2004)* (Vol. 3060, pp. 187–201). London, Ontario: CSCSI. http://doi.org/10.1007/978-3-540-24840-8_14

Barrière, C., & Agbago, A. (2006). TerminoWeb: a software environment for term study in rich contexts. In *Conference on Terminology, Standardisation and Technology Transfer (TSTT 2006)*. Beijing.

Bourigault, D., & Slodzian, M. (1999). Pour une terminologie textuelle. *Terminologies Nouvelles*, *19*, 29–32.

Bowker, L. (2003). Lexical Knowledge Patterns, Semantic Relations, and Language Varieties: Exploring the Possibilities for Refining Information Retrieval in an International Context. *Cataloging & Classification Quarterly*, *37*(1–2), 153–171. http://doi.org/10.1300/J104v37n01_11

Cimiano, P., & Staab, S. (2005). Learning concept hierarchies from text with a guided agglomerative clustering algorithm. In C. Biemann & G. Paas (Eds.), *Proceedings of ICML 2005. Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*. Bonn.

Condamines, A. (2002). Corpus analysis and conceptual relation patterns. *Terminology*, *8*(1), 141–162. http://doi.org/10.1075/term.8.1.07con

Faber, P., & Buendía, M. (2014). EcoLexicon. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress* (pp. 601–607). Bolzano: EURALEX.

Faber, P., León-Araúz, P., & Reimerink, A. (2016). EcoLexicon : New Features and Challenges. In *GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with LREC 2016* (pp. 73–80).

Jakubíček, M., Kilgaiff, A., McCarthy, D., & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. In *Proceedings of the PACLIC 24* (pp. 741–747).

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX International Congress* (pp. 105–116). Lorient: EURALEX.

L'Homme, M.-C., & Marshman, E. (2006). Terminological Relationships and Corpus-based Methods for Discovering them: An Assessment for Terminographers. In L. Bowker (Ed.), *Lexicography, Terminology and Translation. Text-based studies in honour of Ingrid Meyer* (pp. 67–80). Ottawa: University of Ottawa Press.

Lafourcade, M., & Ramadier, L. (2016). Semantic Relation Extraction with Semantic Patterns Experiment on Radiology Reports. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4578–4582). Portorož, Slovenia: European Language Resources Association (ELRA).

Lefever, E., Kauter, M. Van De, Hoste, V., Van de Kauter, M., & Hoste, V. (2014). HypoTerm: Detection of hypernym relations between domain-specific terms in Dutch and English. *Terminology*, *20*(2), 250–278. http://doi.org/10.1075/term.20.2.06lef

León Araúz, P., San Martín, A., & Faber, P. (2016). Pattern-based Word Sketches for the Extraction of Semantic Relations. In *Proceedings of the 5th International Workshop on Computational Terminology* (pp. 73–82). Osaka.

Marshman, E. (2002). The cause-effect relation in a biopharmaceutical corpus: English knowledge patterns. In *Proceedings of the 6th international Conference on Terminology and Knowledge Engineering* (pp. 89–94). Nancy.

Marshman, E. (2004). The cause-effect relation in a French-language biopharmaceuticals corpus: Some lexical knowledge patterns. In *Proceedings of the Workshop on Computational and Computer-assisted Terminology, in association with the Language Resources and Evaluation Conference 2004* (pp. 24–27).

Marshman, E. (2014). Enriching terminology resources with knowledge-rich contexts: A case study. *Terminology*, *20*(2), 225–249. http://doi.org/10.1075/term.20.2.05mar

Marshman, E., Morgan, T., & Meyer, I. (2002). French patterns for expressing concept relations. *Terminology*, *8*(1), 1–29. http://doi.org/10.1075/term.8.1.02mar

Meyer, I. (1994). Linguistic Strategies and Computer Aids for Knowledge Engineering in Terminology. *L'actualité terminologique/Terminology Update*, *27*(4), 6–10.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent advances in computational terminology* (pp. 279–302). Amsterdam/Philadelphia: John Benjamins.

Pearson, J. (1998). *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.

San Martín, A., Cabezas-García, M., Buendía Castro, M., Sánchez Cárdenas, B., León-Araúz, P., & Faber, P. (2017). Recent Advances in EcoLexicon. *Dictionaries: Journal of the Dictionary Society of North America*, *38*(1), 96–115.

## 8. Language Resource References

LexiCon Research Group (2018). EcoLexicon, http://ecolexicon.ugr.es/

LexiCon Research Group (2017). EcoLexicon English Corpus, https://www.sketchengine.co.uk/ecolexicon-corpus/

LexiCon Research Group (2018). EcoLexicon Semantic Sketch Grammar, http://ecolexicon.ugr.es/essg

# Combining Dictionaries, Wordnets and other Lexical Resources– Advantages and Challenges

**Bolette Sandford Pedersen[1], Sanni Nimb[2], Sussi Olsen[1], Nicolai Hartvig Sørensen[2]**
University of Copenhagen[1], The Danish Society of Language and Literature[2]
Njalsgade 139, DK-2300 S[1], Christians Brygge 1, DK-1219
bspedersen@hum.ku.dk, sn@dsl.dk, saolsen@hum.ku.dk, nhs@dsl.dk

**Abstract**

In this paper we account for the advantages, challenges and pitfalls that we have encountered when compiling language technology (LT) resources based on dictionary information and vice versa. We describe the main lines in our collaborative work during the last decade and based on this experience, we provide some suggestions and recommendations in order for dictionaries to become more standardised and multifunctional and thereby also more directly useful for LT.

**Keywords:** lexical resources, wordnets, framenets, annotated corpora, language technology, international standards, language transfer

## 1. Compiling LT resources from dictionaries and vice versa

In this paper we account for the advantages, challenges and pitfalls that we have encountered when compiling language technology (LT) resources based on dictionary information and vice versa. Our focus is on a medium-resourced language, namely Danish, where LT resource scarcity has prompted us to look seriously into the perspective of re-using existing lexical resources.

To this end, it is important to stress that dictionaries are not just systematic collections of words with information about morphology and syntax; they are cultural testimonies in the sense that they describe the society and culture in which they are being compiled. Ideally, the LT systems that we develop for use in both our private and professional lives should reflect the same dimensions. However, if we solely adapt our future LT systems on the basis of English language models, there is a danger that this dimension is completely overlooked.

In order to address this challenge, the Danish language and language technology community has in recent years focused on methods for building language technology resources that:

- employ existing high-quality lexical data of Danish,
- comply with international standards, *and*
- incorporate elements of language transfer from better resourced languages where relevant[1]

In addition to this combination of approaches, focus has been into keeping a reference point across all the developed resources in terms of common sense identifiers or a common "core" so to speak. This approach has enabled the teams to not only produce LT resources from traditional dictionary work, but also go the other way: To exploit LT resources when developing a new Danish thesaurus.

Where a close collaboration between a dictionary publisher and a university institute (as seen in our case between The Society for Danish Language and Literature and the Centre for Language technology at the University of Copenhagen), is not seen so often, the idea of developing lexical cores as a basis for new resources, is not a new or unique approach. Examples are such as The DANTE database (Atkins 2010) which is a lexical database which provides a fine-grained, corpus-based description of the core vocabulary of English. SALDO (Borin et al. 2013) is a Swedish semantic and morphological lexical resource primarily intended for use in LT applications, which however, is closely entangled with two paper dictionaries as well as with the Swedish wordnet. Similar to SALDO, Cornetto stands for Combinatorial and Relational Network as Toolkit for Dutch Language Technology and is a lexical semantic database that combines a wordnet with framenet-like information for Dutch (cf. Vossen et al. 2013). The combination of the two lexical resources (the Dutch wordnet and the Referentie Bestand Nederlands) is claimed to provide a richer relational database to be used in LT.

Our own starting point for the collaborative work between resources, which has been realised for more than a decade, is the monolingual dictionary *Den Danske Ordbog* (DDO) and the Danish wordnet, DanNet; the latter compiled a decade ago with DDO as its primary source (Pedersen et al. 2009), but still complying with wordnet standards (Fellbaum 1998, Vossen 1999). To compile the wordnet we used a bottom-up strategy based on the hypernym given for each sense definition in the dictionary expressed in a specific genus proximum field. As consequence of this compilation approach, the two resources are linked at

[1] See for instance Pedersen et. al. (2018) for transfer of frame-semantic information from English.

sense level, allowing for the combination of all types of information across the two resources.

For instance, the links have been used to enrich the online version of the DDO, enabling users to browse related words in terms of hyponymy (Sørensen & Trap-Jensen 2010). The exact order of the hyponyms in the online presentation 'Beslægtede Ord' (Related Words, available 2009-17) was based on a calculation of semantic relatedness depending on information in the wordnet: a set of semantic relations and the ontological types. Another direct use of the combined data is the graphical representation of DanNet's hierarchies and relations at *andreord.dk* where the (restricted) definitions of DDO as well as domain information and citations from the dictionary are included. In Section 2 we describe the common sense inventory in more detail.

Most recently, the linked data has furthermore resulted in new resources in terms of an *annotated corpus*, *a Danish thesaurus* and a *Berkeley-style frame-lexicon* all of which we briefly account for in Section 3.

In Section 4 we sketch out some recommendations for a future larger degree of multi-functionality in the next generation of dictionary projects. In particular, we discuss the perspectives of future, truly digitally born lexical resources which are not limited or influenced by (former) physical issues, and which can therefore be compiled and interlinked with a higher degree of consistency.

## 2. One sense inventory as a common reference point

The DDO is corpus-based and continuously being extended with new words and senses. Entries are organized in main and sub-senses in a structure which to a high degree reflects the logical relations between a core sense and its either narrower or broader sense derivations as well as metaphorically derived senses. However, this general principle is sometimes downgraded for communicative purposes. For instance, very deep sub-sense structures are avoided, and very frequent senses have instead been upgraded to main senses, no matter whether there exists a logical relation to a core sense or not. What is also important to notice is that the first edition of DDO was published in print in six volumes. This influenced to a very high degree the sense structure of less frequent words. For such words the core and sub-senses were often merged into one definition in order to save space for a more detailed description of the very frequent words. Furthermore, many cases of regular polysemy are implicit in the dictionary, covered by only one sense.

When we compiled the Danish wordnet, DanNet (Pedersen et. 2009) from the DDO in a semi-automatic fashion, these informal deviations from the general structure caused some extra adjustment work in terms of reorganization of senses and collapses of some senses into the same synsets. Likewise, the adjustment and reorganization of the implicit DDO hyponymy structure was somewhat time consuming. For instance, we realized that many of the hyponymies found in the DDO had incorporated a great mixture of *natural* and *functional* kinds in Cruse's terminology (Cruse 2000), mixing natural taxonomies with layman's view of the concept's *function*. For instance, edible plants could have either 'plant' or 'vegetable' as their hypernym in the DDO depending somewhat on the lemma's frequency in the corpus and on its subsequent allotted physical space and unfolding in the original dictionary.

## 3. Developing new resources based on DDO/DanNet

### 3.1 Combinations of information from wordnet and dictionary: A thesaurus and a Frame lexicon

The semantic links between DanNet and the DDO further facilitated the compilation of a comprehensive thesaurus for Danish (Nimb et al. 2014 a; Nimb et al. 2014 b). Large hierarchies of words (i.e. all furniture or clothes), including links to the corresponding DDO senses, were directly transferred to the relevant thesaurus chapters. Data extracted from DDO in the form of definitions and synonyms was used to arrange the hyponyms into subgroups, and the categorization of senses profited from our experiences with the wordnet compilation.

Several of the semantic relations from DanNet were adapted in order to structure the thesaurus XML manuscript. By use of these formal semantic criteria, the vocabulary was annotated with core semantic types such as acts, events, properties, persons, artifacts etc., enabling us to keep track of the semantic grouping of words throughout the thesaurus project as well as to identify and extract precisely restricted semantic groups from the finished manuscript. In this way, approx. 1/5 of the words and expressions in the thesaurus were identified as acts or events and subsequently used for starting up the Danish frame lexicon. See Nimb et al. (2017) and Nimb (2018) for more details.

The chapter division in the thesaurus made it possible to identify precise semantic domains such as acts of 'communication' and 'cognition' and thereby to assign the appropriate frame in Berkeley FrameNet covering these exact domains to a large quantity of lexical units at a time. The resulting frames have been tested on restricted corpus data (Nimb et al. 2017), and the project has afterwards been extended in order to compile frames for the entire Danish act/event vocabulary. In a future project, we plan to study whether the sense links between the frame data and DanNet can be used to extend the wordnet with framenet information, i.e. especially to improve the verb hierarchies of DanNet.

### 3.2 A semantically annotated corpus

The common backbone sense inventory was also further exploited for annotating a corpus – annotations which were subsequently used for training a Danish sense tagger (Martinez et al. 2015 and Pedersen et al. 2018). Hence, the so-called SemDaX corpus (Pedersen et al. 2016) contains about 100,000 words with semantic annotations of varying granularity, annotated by humans. The most coarse-grained sense annotations are annotations of all content words with so-called *supersenses*, derived from Princeton WordNet's lexicographical files.

In addition to the supersense annotations, SemDaX comprises lexical sample annotations for a small set of highly ambiguous nouns. The fine-grained annotations are based on the set of senses in DDO. Each noun has been annotated with the full DDO sense inventory as well as with two different automatically clustered sense inventories of different granularity (Pedersen et al. 2018) based on their ontological type in DanNet.

All manual annotations were carried out in the annotation tool WebAnno (Yimam et al., 2013). The aim of the corpus is to serve as training and test data for word sense disambiguation, as well as to estimate the usefulness of the different sense annotation schemes by analyzing the data and the inter-annotator agreement.

## 4. Future dictionaries: How can they become more suitable for multiple purposes?

The Danish lexical core approach was initiated with the combination of a dictionary and a wordnet based on the common sense inventory. This initiative gave interesting insights and results and led on to other lexical products as described in the above. To sum up, the linked data combining hierarchical information, semantic relations, dictionary definitions, and dictionary synonyms has enabled us to compile a thesaurus and consequently also a frame lexicon in a very efficient way. The logical information from the dictionary sense structure combined with the ontological information in the wordnet has furthermore allowed us to carry out several comparative annotation studies with both full sense inventories and sense *clusters*. Using this corpus for word sense disambiguation has given us insights wrt. how to identify the most adequate levels of sense granularity – both for human annotators and for automatic systems.

The work has further provided insights into where dictionaries for human users lack explicit information which is needed for human language technology. One example is the logical relation between senses which should preferably be more specific and for instance described by more specific links. Another is the discrepancies in hypernym structure where space issues in the printed dictionary to some extent influenced the structure so that for instance regular polysemous lemmas did not systematically refer to their correct hypernyms.

Also the assignment of very coarse-grained semantic information, such as whether the sense is a first, a second or a third order type of entity (cf. Lyons 1977) would be very useful to have implicitly expressed in dictionaries, preferably by the use of simple attributes. Often dictionary definitions use polysemous words across the three semantic classes (i.e. figurative, abstract words that also have a concrete sense). This has as consequence that it is not at all easy to extract whether a standalone definition defines something concrete or abstract – or maybe even covers both cases – without having to look deeper into citations, other senses of the word etc. The same goes for many cases of regular polysemy. Precise attributes on regular polysemy patterns should preferably be included in dictionaries, allowing the editor to check out and mark which of the regular senses are accounted for in the description, based on corpus inspection.

Our work with dictionaries in an LT context has also inspired us the other way around regarding which supplementary information types seem useful for LT resources and have not previously been fully acknowledged as such. Surprisingly enough, for instance, the *function* relation (labelled the 'telic role' in Pustejovsky 1995, and 'functional/nominal' kinds by Cruse 2000) receives very little attention in the wordnet literature, and only very few wordnets contain – to our knowledge – this information type even if it proves quite crucial in many inference tasks in particular when it comes to tasks involving artifacts. The relation is highly represented in many DDO definitions where a concept's function is very often described – and when it is not, the integration with other resources is much more complicated. In fact, in Nimb & Pedersen 2000 we concluded that a concept's function often constitutes the very core of the figurative sense of the same word[2]. To this end, we would recommend that also this relation becomes formally explicit via the logical relations between senses as well as the function role formally explicit in dictionaries.

With regards to sense structure, one can only hope that future digitally born dictionary versions (where physical limitations is no longer an issue), will by and by result in a more consistent sense description where lesser frequent words are treated with same consistency as frequent words. Combined with a higher level of standardization – in our case partly introduced via the international wordnet and framenet standards – some of the obstacles that we have encountered in our work can hopefully gradually be reduced. However, there is no doubt that it requires explicit focus.

In fact, the newly embarked ELEXIS infrastructure has

[2] For instance, the telic role of *window*, namely to give access to a broader view of the surroundings from the inside of something, determines the figurative sense in a phrase like *a window to the world*.

exactly the goal of explicitly addressing cooperation and information exchange among lexicographical and LT research communities. The aim is to achieve a higher degree of standardisation and inter-functionality of existing and future dictionaries. The infrastructure is a newly granted project under the Horizon 2020 INFRAIA call, and the plan is to work with strategies, tools and standards for extracting, structuring and linking of lexicographic resources.

## 5. Bibliographical References

Atkins, B. T. S. (2010). The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data. In : G.-M. de Schryver (ed.) *A Way with Words: Recent Advances in Lexical Theory and Analysis*. A Festschrift for Patrick Hanks. Kampala: Menha Publishers.

Borin, Lars, Markus Forsberg, Lennart Lönngren (2013). SALDO: a touch of yin to WordNet's yang. In : *Language Resources and Evaluation, Volume 47, Issue 4*, pp 1191–1211.

Cruse, D.A (2000). *Meaning in Language*. Oxford: Oxford University Press.

DDO = *Den Danske Ordbog*. (E. Hjorth et al). 2003-2005. Det Danske Sprog- og Litteraturselskab & Gyldendal, Copenhagen.

Fellbaum, Christiane (ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT press.

Lyons, John (1977). *Semantics*. Cambridge University Press

Martínez Alonso, Héctor; Anders Johannsen; Sussi Olsen; Sanni Nimb; Nicolai Hartvig Sørensen; Anna Braasch; Anders Søgaard; Bolette Sandford Pedersen. (2015). Supersense tagging for Danish. In : *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015*, Linköping Electronic Conference Proceedings #109, ACL Anthology, Linköping University Electronic Press, Sweden.

Martínez Alonso, Héctor; Anders Johannsen; Sanni Nimb; Sussi Olsen; Bolette Sandford Pedersen. (2016). An empirically grounded expansion of the supersense inventory. In : *Proceedings of Global Wordnet Conference 2016*.

Nimb, Sanni (2018). The Danish FrameNet Lexicon: method and lexical coverage. In : *Proceedings from the International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons.* Miyazaki, Japan

Nimb, Sanni; Braasch, Anna; Olsen, Sussi; Pedersen, Bolette Sandford; Søgaard, Anders. (2017). From Thesaurus to FrameNet. In: *Proceedings of eLex 2017*, Leiden.

Nimb, S. & B.S. Pedersen (2000). Treating Metaphorical Senses in a Danish Computational Lexicon - Different Cases of Regular Polysemy. In : *Proceedings from The Ninth Euralex International Congress pp. 679-691*, Universität Stuttgart Germany.

Nimb, Sanni, Henrik Lorentzen, Liisa Theilgaard, Thomas Troelsgård, Lars Trap-Jensen (2014 a). *Den Danske Begrebsordbog*. Det Danske Sprog- og Litteraturselskab og Syddansk Universitetsforlag

Nimb, Sanni, Lars Trap-Jensen, Henrik Lorentzen (2014 b) The Danish Thesaurus: Problems and Perspectives. In: Andrea Abel, Chiara Vettori & Natascia Ralli (eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen: EURAC Research, pp. 191-199

Pedersen, Bolette S., Manex Agirrezabal, Sanni Nimb, Sussi Olsen, Ida Rørmann (2018). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In: *Proceedings of GWC2018*, Singapore.

Pedersen, Bolette S., Sanni Nimb, Jørg Asmussen, Nicolai Sørensen, Lars Trap-Jensen, Henrik Lorentzen. (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. In: *Language Resources and Evaluation*, Computational Linguistics Series, pp.269-299.

Pedersen, Bolette S., Sanni Nimb, Anders Søgaard, Mareike Hartmann, Sussi Olsen (2018). A Danish FrameNet Lexicon and an Annotated Corpus Used for Training and Evaluating a Semantic Frame Classifier. In: *Proceedings of LREC 2018*, Miyazaki, Japan.

Pedersen, Bolette S.; Braasch, Anna; Johannsen, Anders Trærup; Martínez Alonso, Héctor; Nimb, Sanni; Olsen, Sussi; Søgaard, Anders; Sørensen, Nicolai. (2016). The SemDaX Corpus - sense annotations with scalable sense inventories. In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*. Portorož, Slovenia.

Pustejovsky, James (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA.

Sørensen, Nicolai Hartvig. & Trap-Jensen, Lars (2010). Den Danske Ordbog som begrebsordbog. In : Harry Lönnroth, Kristina Nikula (eds.) *Nordiska Studier i Leksikografi 10*, NFL-skrift nr 11, Tammerfors 2010, pp. 164-179.

Vossen, P. (ed). (1999). *EuroWordNet : A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, The Netherlands.

Vossen, P., I. Maks, R. Segers, H. van der Vliet, M. Moens, K. Hofmann, E. Tjong Kim Sang, and M. de Rijke (2013). Cornetto: a lexical semantic database for Dutch. In : *Essential speech and language technology for Dutch, results by the Stevin-programme*, P. Spyns and J. Odijk, Eds., Springer series Theory and Applications of Natural Language Processing, 2013, pp. 165-184.

Yimam, S.M., Gurevych, I., Eckart de Castilho, R., & Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In: *Proceedings of ACL-2013*, demo session, Sofia, Bulgaria.