

GLOBALLEX 2016

Lexicographic Resources for Human Language Technology

Workshop Programme

24 May 2016

09:00 – 09:10 Introduction by Ilan Kernerman and Simon Krek

09:10 – 10:30 Session 1

Patrick Hanks, *A common-sense paradigm for linguistic research*

Pamela Faber, Pilar León-Araúz and Arianne Reimerink, *EcoLexicon: New features and challenges*

Sara Carvalho, Rute Costa and Christophe Roche, *Ontotermology meets lexicography: The Multimodal Online Dictionary of Endometriosis (MODE)*

Gregory Grefenstette and Lawrence Muchemi, *Determining the characteristic vocabulary for a specialized dictionary using Word2vec and a directed crawler*

10:30 – 11:00 Coffee break

11:00 – 12:40 Session 2

Malin Ahlberg, Lars Borin, Markus Forsberg, Olof Olsson, Anne Schumacher and Jonatan Uppström, *Karp: Språkbanken's open lexical infrastructure*

Ivelina Stoyanova, Svetla Koeva, Maria Todorova and Svetlozara Leseva, *Semi-automatic compilation of a very large multiword expression dictionary for Bulgarian*

Raffaele Simone and Valentina Piunno, *CombiNet: Italian word combinations in an online lexicographic tool*

Irena Srdanovic and Iztok Kosem, *GDEX for Japanese: Automatic extraction of good dictionary example candidates*

Jana Klímová, Veronika Kolářová and Anna Vernerová, *Towards a corpus-based valency lexicon of Czech nouns*

12:40 – 14:20 Lunch break

14:20 – 16:00 Session 3

Jan Hajic, Eva Fucikova, Jana Sindlerova and Zdenka Uresova, *Verb argument pairing in a Czech-English treebank*

Sonja Bosch and Laurette Pretorius, *The role of computational Zulu verb morphology in multilingual lexicographic applications*

Martin Benjamin, *Toward a global online living dictionary: A model and obstacles for big linguistic data*

Luis Morgado da Costa, Francis Bond and František Kratochvíl, *Linking and disambiguating Swadesh texts*

Luis Espinoza Anke, Roberto Carlini, Horacio Saggion and Francesco Ronzano, *DEFEXT: A semi-supervised definition extraction tool*

16:00 – 16:30 Coffee break

16:30 – 17:00 Session 4

Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda and Guadalupe Aguado-de-Cea, *Modelling multilingual lexicographic resources for the web of data: The K Dictionaries case*

Ivett Benyeda, Péter Koczka and Tamás Varadi, *Creating seed lexicons for under-resourced languages*

17:00 – 18:00 GLOBALEX Discussion

Editors

Ilan Kernerman
Iztok Kosem
Simon Krek
Lars Trap-Jensen

K Dictionaries
Trojina, Institute for Applied Slovene Studies
“Jožef Stefan” Institute
Society for Danish Language and Literature

Workshop Organizers/Organizing Committee

Andrea Abel	EURALEX
Ilan Kernerman*	ASIALEX
Steven Kleinedler	DSNA
Iztok Kosem	eLex
Simon Krek*	eLex
Julia Miller	AUSTRALEX
Maropeng Victor Mojela	AFRILEX
Danie J. Prinsloo	AFRILEX
Rachel Edita O. Roxas	ASIALEX
Lars Trap-Jensen	EURALEX
Luanne von Schneidemesser	DSNA
Michael Walsh	AUSTRALEX

* Co-chairs of the Organising Committee

Workshop Programme Committee

Michael Adams	Indiana University
Philipp Cimiano	University of Bielefeld
Janet DeCesaris	Universitat Pompeu Fabra
Thierry Declerck	German Research Center for Artificial Intelligence
Anne Dykstra	Fryske Akademie
Edward Finegan	University of Southern California
Thierry Fontenelle	Translation Center for the Bodies of the EU
Polona Gantar	University of Ljubljana
Alexander Geyken	Berlin-Brandenburg Academy of Sciences and Humanities
Rufus Gouws	Stellenbosch University
Jorge Gracia	Madrid Polytechnic University
Orin Hargraves	University of Colorado
Ulrich Heid	Hildesheim University

Chu-Ren Huang
Miloš Jakubiček
Jelena Kallas
Ilan Kernerman
Annette Klosa
Iztok Kosem
Simon Krek
Robert Lew
Marie Claude l'Homme
Nikola Ljubešić
Stella Markantonatou

John McCrae
Roberto Navigli
Vincent Ooi
Michael Rundell
Mary Salisbury
Adam Smith
Pius ten Hacken
Carole Tiberius
Yukio Tono
Lars Trap-Jensen
Tamás Váradi
Elena Volodina
Eveline Wendl-Vogt
Shigeru Yamada

Hong Kong Polytechnic University
Lexical Computing – Sketch Engine
Institute of Estonian Language
K Dictionaries
German Language Institute
Trojina, Institute for Applied Slovene Studies
“Jožef Stefan” Institute
Adam Mickiewicz University
University of Montreal
University of Zagreb
Institute for Language and speech Processing
ATHENA
National University of Ireland Galway
Sapienza University of Rome
National University of Singapore
Lexicography Masterclass
Massey University
Macquarie University
Innsbruck University
Institute of Dutch Lexicology
Tokyo University of Foreign Studies
Society for Danish Language and Literature
Hungarian Academy of Sciences
Gothenburg University
Austrian Academy of Sciences
Waseda University

Table of contents

Towards a Corpus-based Valency Lexicon of Czech Nouns <i>Jana Klímová, Veronika Kolářová, Anna Vernerová</i>	1
Ontoterminology meets Lexicography: the Multimodal Online Dictionary of Endometriosis (MODE) <i>Sara Carvalho, Rute Costa, Christophe Roche</i>	8
Verb Argument Pairing in Czech-English Parallel Treebank <i>Jan Hajič, Eva Fučíková, Jana Šindlerová, Zdeňka Urešová</i>	16
DEFEXT: A Semi Supervised Definition Extraction Tool <i>Luis Espinosa-Anke, Roberto Carlini, Horacio Saggion, Francesco Ronzano</i>	24
Linking and Disambiguating Swadesh Lists: Expanding the Open Multilingual Wordnet Using Open Language Resources <i>Luis Morgado da Costa, Francis Bond, František Kratochvíl</i>	29
The Role of Computational Zulu Verb Morphology in Multilingual Lexicographic Applications <i>Sonja Bosch, Laurette Pretorius</i>	37
CombiNet. A Corpus-based Online Database of Italian Word Combinations <i>Valentina Piunno</i>	45
Creating seed lexicons for under-resourced languages <i>Ivett Benyeda, Péter Koczka, Tamás Váradi</i>	52
GDEX for Japanese: Automatic Extraction of Good Dictionary Example Candidates <i>Irena Srdanović, Iztok Kosem</i>	57
Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case <i>Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, Guadalupe Aguado-de-Cea</i>	65
EcoLexicon: New Features and Challenges <i>Pamela Faber, Pilar León-Araúz, Arianne Reimerink</i>	73
Determining the Characteristic Vocabulary for a Specialized Dictionary using Word2vec and a Directed Crawler <i>Gregory Grefenstette, Lawrence Muchemi</i>	81
Semi-automatic Compilation of the Dictionary of Bulgarian Multiword Expressions <i>Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetlozara Leseva</i>	86

Author Index

Aguado-de-Cea, Guadalupe	65	León-Araúz, Pilar	73
Benyeda, Ivett	52	Leseva, Svetlozara	86
Bond, Francis	29	Montiel-Ponsoda, Elena	65
Bosch, Sonja	37	Morgado da Costa, Luís	29
Bosque-Gil, Julia	65	Muchemi, Lawrence	81
Carlini, Roberto	24	Piunno, Valentina	45
Carvalho, Sara	8	Pretorius, Laurette	37
Costa, Rute	8	Reimerink, Arianne	73
Espinosa-Anke, Luis	24	Roche, Christophe	8
Faber, Pamela	73	Ronzano, Francesco	24
Fučíková, Eva	16	Saggion, Horacio	24
Gracia, Jorge	65	Srdanović, Irena	57
Grefenstette, Gregory	81	Stoyanova, Ivelina	86
Hajič, Jan	16	Šindlerová, Jana	16
Klímová, Jana	1	Todorova, Maria	86
Koczka, Péter	52	Urešová, Zdeňka	16
Koeva, Svetla	86	Váradi, Tamás	52
Kolářová, Veronika	1	Vernerová, Anna	1
Kosem, Iztok	57		
Kratochvíl, František	29		

Introduction

The field of lexicography has been shifting to digital media, with effect on all stages of research, development, design, evaluation, publication, marketing and usage. Modern lexicographic content is created with help of dictionary writing tools, corpus query systems and QA applications, and becomes more easily accessible and useful for integration with numerous LT solutions, as part of bigger knowledge systems and collaborative intelligence.

At the same time, extensive interlinked language resources, primarily intended for use in Human Language Technology (HLT), are being created through projects, movements and initiatives, such as Linguistic Linked (Open) Data (LLOD), meeting requirements for optimal use in HLT, e.g. unique identification and use of web standards (RDF or JSON-LD), leading to better federation, interoperability and flexible representation. In this context, lexicography constitutes a natural and vital part of the LLOD scheme, currently represented by wordnets, FrameNets, and HLT-oriented lexicons, ontologies and lexical databases. However, a new research paradigm and common standards are still lacking, and so are common standards for the interoperability of lexicography with HLT applications and systems.

The aim of this workshop is to explore the development of global standards for the evaluation of lexicographic resources and their incorporation with new language technology services and other devices. The workshop is the first-ever joint initiative by all the major continental lexicography associations, seeking to promote cooperation with related fields of HLT for all languages worldwide, and it is intended to bridge various existing gaps within and among such different research fields and interest groups. The target audience includes lexicographers, computational and corpus linguists and researchers working in the fields of HLT, Linked Data, the Semantic Web, Artificial Intelligence, etc.

GLOBALEX 2016 is sponsored by the five existing continental lexicography associations and the international conferences on electronic lexicography:

- AFRILEX – The African Association for Lexicography
- ASIALEX – The Asian Association for Lexicography
- AUSTRALEX – The Australasian Association for Lexicography
- DSNA – The Dictionary Society of North America
- EURALEX – The European Association for Lexicography
- eLex – Electronic Lexicography in the 21st Century conferences

This workshop constitutes the initial step in forming GLOBALEX – a global constellation for all continental, regional, local, topical or special interest communities concerned with lexicography. GLOBALEX will promote knowledge sharing and cooperation among its members and with other parties concerned with language and linguistics. It will aim to establish global standards for the creation, evaluation, dissemination and usage of lexicographic resources and solutions, and for the interoperability of lexicography with other relevant disciplines and branches of the HLT academe and industry worldwide.

We would like to thank all the dedicated members of the technical program committee, the sponsors, ELRA, and of course all authors for an inspiring and exciting workshop and proceedings.

Ilan Kernerman, Iztok Kosem, Simon Krek and Lars Trap-Jensen

Editors of the proceedings and organisers of GLOBALEX Workshop 2016

Towards a Corpus-based Valency Lexicon of Czech Nouns

Jana Klímová, Veronika Kolářová, Anna Vernerová

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
E-mail: {klimova, kolarova, vernerova}@ufal.mff.cuni.cz

Abstract

Corpus-based Valency Lexicon of Czech Nouns is a starting project picking up the threads of our previous work on nominal valency. It builds upon solid theoretical foundations of the theory of valency developed within the Functional Generative Description. In this paper, we describe the ways of treating valency of nouns in a modern corpus-based lexicon, available as machine readable data in a format suitable for NLP applications, and report on the limitations that the most commonly used corpus interfaces provide to the research of nominal valency. The linguistic material is extracted from the Prague Dependency Treebank, the synchronic written part of the Czech National Corpus, and Araneum Bohemicum. We will utilize lexicographic software and partially also data developed for the valency lexicon PDT-Vallex but the treatment of entries will be more exhaustive, for example, in the coverage of senses and in the semantic classification added to selected lexical units (meanings). The main criteria for including nouns in the lexicon will be semantic class membership and the complexity of valency patterns. Valency of nouns will be captured in the form of valency frames, enumeration of all possible combinations of adnominal participants, and corpus examples.

Keywords: corpus, lexicon, nouns, valency

1. Introduction

Nominalizations as reclassifications of their corresponding verbal clauses (Heyvaert, 2003) are in the center of attention of many researchers; both their syntactic and semantic aspects are studied across various languages and frameworks (Chomsky, 1970; Osenova, 2009; Alexiadou & Rathert, 2010; Melloni, 2011). One such aspect is argument structure (Grimshaw, 1991) or nominal “valency” (Spevak, 2014): the number, type and form of arguments that are bound to a noun. Although nominal valency still remains in the shadow of the valency of verbs, it is the matter of both theoretical and lexicographic studies which are in a close relationship. This relationship may be best exemplified by the *Explanatory Combinatorial Dictionary of Modern Russian* (Mel’čuk & Zholkovsky, 1984), a dictionary created within the theoretical framework of the Meaning-Text Theory. While many valency lexicons are primarily intended for non-native speakers (e.g., Herbst et al., 2004) is intended for learners of English and the oldest lexicon covering nouns (Sommerfeldt & Schreiber, 1977) for learners of German), nouns are also covered in lexicons created mainly with NLP applications in mind, such as FrameNet¹ (ongoing; see also Ruppenhofer et al., 2006) and NomBank 1.0² (see also Meyers, 2007). Both projects involve corpus annotation: FrameNet is based on the British National Corpus; NomBank uses the Wall Street Journal Corpus of the Penn Treebank. Corpus-based valency lexicons of Slavic languages mostly focus on verbal valency; Polish Valence Dictionary (Walenty³) also covers nouns and adjectives (cf. Przepiórkowski & Hajnicz & Patejuk & Woliński & Skwarski & Świdziński, 2014).

Valency of Czech nouns is covered by two valency lexicons that also cover verbs and adjectives, namely

a printed dictionary *Slovník slovesných, substantivních a adjektivních vazeb a spojení* (Svozilová & Prouzová & Jirsová, 2005) and an electronic lexicon built during the tectogrammatical annotation of the Prague Dependency Treebank (PDT)⁴, called PDT-Vallex⁵ (Hajič et al., 2003).

In this paper, we present our work on a new corpus-based valency lexicon of Czech nouns; the lexicographic work on the lexicon started at the beginning of 2016. First, we delimit our theoretical framework (Section 2) and specify typical and special valency behavior of Czech deverbal nouns (Section 3). Then we describe differences between our approach and existing lexical resources for Czech (Section 4, Section 5, Section 6, and Section 7). Finally, we focus on ways of searching for nominal valency patterns through the available Czech corpora, see Section 8.

2. General Framework of Functional Generative Description

Issues of valency of Czech nouns were discussed as early as the 1960s by Jirsová (1966) and Křížková (1968), and the first monograph dealing with valency of non-productively derived Czech nouns was elaborated by Novotný (1980). Valency of nouns is studied within various theoretical frameworks, e.g., the modified valency theory formulated by Karlík (2000), transformational generative grammar (Veselovská, 2001; Dvořáková-Procházková, 2008), the lexicological and “corpus-driven” approach (Čermák, 1991; Čermáková, 2009).

Focusing on deverbal nouns, our approach to noun valency is based on the theory of verbal valency developed within the framework of Functional Generative Description (FGD) by Panevová (1974 and 1975) and Sgall & Hajičová & Panevová (1986). Valency frames are presumably stored in the (mental) lexicon, and are

¹ <https://framenet.icsi.berkeley.edu>

² <http://nlp.cs.nyu.edu/meyers/NomBank.html>

³ <http://walenty.ipipan.waw.pl/>

⁴ LDC Catalog No.: LDC2006T01,

<http://hdl.handle.net/11858/00-097C-0000-0001-B098-5>

⁵ <http://hdl.handle.net/11858/00-097C-0000-0023-4338-F>

reflected in the tectogrammatical representation of sentences. Valency frames of verbs contain slots for the following types of complementation:

(a) obligatory and optional inner participants (arguments): Actor (ACT), Patient (PAT), Addressee (ADDR), Effect (EFF), Origin (ORIG);

(b) obligatory free modifications (adjuncts), especially those with the meanings of direction (e.g., *přijet někam*.DIR3 ‘to arrive somewhere’), location (e.g., *přebývat někde*.LOC ‘to dwell somewhere’), and manner (e.g., *chovat se dobře*.MANN ‘to behave well’).

This distinction between inner participants and free modifications is maintained in the description of nominal valency too. Within the treatment given to nominal valency in the FGD (Panevová, 2000; Kolářová, 2010), the meaning of a given noun is the most important factor in determining its valency frames. Following the definition of syntactic and lexical derivation given by Kuryłowicz (1936), we distinguish three groups of nouns: (i) Nouns derived from verbs by the so called syntactic derivation, i.e. nouns that differ from their source verbs only in part of speech but not in meaning. Nouns in this group denote actions (*vyrábění / výrobení* ‘manufacturing’ // *výroba* ‘production’) or states (*vyskytování se / vyskytnutí se* ‘occurring’ // *výskyt* ‘occurrence’); the forms of their participants are typical (regular).

(ii) Nouns derived by lexical derivation, i.e. nouns whose lexical meaning is unmistakably different from the lexical meaning of their source verbs; it includes names of physical entities related to actions (semantically concrete nouns) such as actor nouns (*učitel* ‘teacher’), nouns denoting a thing (*dodávka* ‘van’, *otvírák* ‘opener’) or nouns denoting a place (*stoupání* ‘slope’, *východ* ‘exit’, *čekárna* ‘waiting room’); forms of their participants can be either typical or special.

(iii) Nouns on the boundary between syntactic and lexical derivation. The meaning of these nouns is slightly different from the meaning of action or state nouns, described in group (i); however, the nouns belonging to this category, such as *pochvala* ‘praise’, are still abstract nouns. The forms of their participants can be either typical or special.

Two basic types of Czech deverbal nouns denote an action or a state and so belong to group (i): nouns derived from verbs by productive means (suffixes *-(e)nítí*, as in *honění* ‘hunting’ or *hubnutí* ‘losing weight’); and nouns derived from verbs by non-productive means or by zero suffix (such as *honba* ‘hunt’, *hon* ‘hunt’). These two types of nouns are at the center of attention in this project since they can often exhibit both typical and special valency behavior.

3. Typical and Special Valency Behavior of Czech Nouns

The valency behavior referred to as *typical* can be observed with the nouns derived by syntactic derivation (group (i) in Section 2). When determining their valency frames, the nouns are expected to inherit all participants that are present in the valency frame of their source verbs, including the “verbal” character of the participants such as Actor, Patient, and Addressee. However, the forms of the participants undergo some regular shifts that can be described in terms of rules.

The manifestation of *special* valency behavior is tied

with two basic issues (Kolářová, 2010): changes in meaning (e.g., an action → a figurative sense), and characteristic properties of valency complementation. The latter involves three phenomena: special forms of valency complementation (see below), reduction of the number of slots in the valency frame of a noun (either pure reduction or incorporation of a participant), and change of the character of valency complementation to exclusively nominal, as in (2) when compared with (1).

Czech is a highly inflectional language; valency participants of a word are primarily distinguished by their morphological category of case. Following Karlík (2000), a distinction between structural cases (NOM and ACC) and non-structural cases (GEN, DAT, LOC, and INS) is useful for the description of verbal valency. A similar distinction turns out to be important also in the nominal domain. The primary general principle (Karlík, 2000: 184) is as follows: within the process of nominalization, the forms of verbal structural cases change whereas the non-structural cases stay the same. This primary general principle explains the *typical shifts* (e.g., ACC → GEN, *lov velryb* ‘a hunt of whales’) in the surface forms of participants and makes it possible to describe the valency behavior of most Czech deverbal nouns.

Secondary general principles were formulated by Kolářová (2010); they involve various *special shifts* (e.g., ACC → prepositional phrase, *lov na velryby* ‘a hunt for whales, i.e. a whale hunt’). The term “special shift” covers the case in which the form of an adnominal participant differs from the form of the corresponding verbal participant and, at the same time, the new form does not correspond to any of the typical shifts. Special shifts frequently occur with non-productively derived nouns; however, they can also occasionally occur with productively derived nouns.

4. PDT-Vallex

Our approach to the development of a corpus-based valency lexicon extends the approach applied in the PDT-Vallex lexicon (for the differences see Section 5.1). In PDT-Vallex, the core valency information is encoded in valency frames in which the possible alternative forms of complementation are taken into account. PDT-Vallex gives information about semantic roles in the form of tectogrammatical functors of the FGD (Mikulová et al., 2006). Each PDT-Vallex entry describes a lexeme (represented by the “lemma”) and its valency frame(s). One valency frame typically corresponds to one meaning (sense) of a word (i.e., a verb, a noun, or an adjective). Although PDT-Vallex does not explicitly work with the term lexical unit, a meaning of a word with its particular valency frame corresponds to a lexical unit, understood roughly as ‘a given word in a given sense’ (Cruse, 1986).

Concerning nouns, PDT-Vallex 1.0 (included in PDT 2.0) contains 3727 entries. So far, special attention has been paid first to capturing the valency properties of nouns derived from verbs by productive means, such as the noun *balení* ‘pack(ing)’, and second to nouns occurring as nominal components in support verb constructions such as the noun *nabídka* ‘offer’ in *učinit nabídku* ‘to make an offer’. The delimitation of boundaries between particular meanings of a noun is one of the most difficult tasks in nominal valency description. For example, the noun (lexeme) *balení* ‘pack(ing)’ is

represented by three valency frames in PDT-Vallex, corresponding to three meanings of the noun, see (1)–(3). Different meanings can sometimes be distinguished by different types or forms of complementation. In (1), we encounter the semantic roles of Actor (ACT) and Patient (PAT), optionally also Effect (EFF); in (2), Material (MAT). The valency frame in (3) is empty.

- (1) *balení*₁ ‘the process of packing’:
ACT(GEN,INS,POSS) PAT(GEN,POSS) EFF^{opt}(*na* ‘on’+ACC, ...)
e.g., *balení dárků.PAT rodiči.ACT* ‘packing gifts by parents’
- (2) *balení*₂ ‘a container’: MAT(GEN)
e.g., *dárkové balení vína.MAT* ‘a gift pack of wine’
- (3) *balení*₃ ‘design’: an EMPTY valency frame
e.g., *knih v brožurkovém balení* ‘a book in a paperback binding’.

5. The Corpus-based Valency Lexicon of Czech Nouns

In order to create a resource useful to a wide audience (the general public, linguists and applications in second language education and in NLP applications) and to facilitate deeper theoretical understanding of nominal valency, we emphasize the following differences from the existing lexical resources:

5.1 Corpus-based Valency Lexicon vs. PDT-Vallex

Our corpus-based valency lexicon of Czech nouns will give a more elaborate treatment of nominal valency than PDT-Vallex. Although the current version of the nominal entries in PDT-Vallex can be exploited to a large extent, the entries should be improved in several aspects.

Extension of the list of involved nouns. PDT-Vallex covers only nouns that were encountered during the annotation of the treebanks in the Prague Dependency family (PDT, PCEDT, PDTSC)⁶. We plan to treat selected semantic classes more exhaustively, especially if the relevant nouns undergo special valency behavior.

All meanings of a noun. Only the senses which occurred in the annotated data of PDT were included in PDT-Vallex. We plan to provide valency patterns for all meanings of the treated nouns as documented in the much larger Czech National Corpus (CNC)⁷ and monolingual dictionaries.

Consistent treatment within semantic classes. PDT-Vallex was built with the intention to enable consistent annotation of each word with its valency in all of the PDT data (Hajič et al., 2003). However, consistency across whole semantic classes went beyond the main goals of the annotation, although it is crucial for the development of the theoretical understanding of valency-related phenomena.

Special forms of participants and valency frames. In PDT-Vallex, special forms of participants have been treated mostly as variants of typical forms. However, Kolářová (2014) argues that a participant in a special form cannot co-occur with the same set of forms of other

participants as the same participant expressed in a typical form. Such difference in the syntactic behavior of a deverbal noun has a certain impact on its meaning, even if it is only a slight nuance. Consequently, we expect that special forms will be treated in separate valency frames in the valency lexicon.

Participants in combinations. Participants modifying nouns can combine under certain conditions. There are some regular restrictions and rules concerning combinations of particular forms as well as their word order. Especially trivalent nouns constitute rather complex patterns when all three participants are expressed on the surface. We intend to provide all possible combinations of participants in various forms, including word order variants. Further research is necessary in order to determine to what extent can restrictions on combinatorial properties of complementation be treated in a grammar component of the lexicon and to what extent separate valency frames in the data component of the lexicon are necessary (for the two-part model of a dictionary which is divided into a grammar component and a data component, see esp. Lopatková et al., 2015).

Type of special valency behavior. The type of special valency behavior will be specified in the relevant entries (e.g., special form of a participant or reduction in the number of slots).

Frequency or stylistic evaluation of a combination / pattern. Where appropriate, frequency and a stylistic evaluation of a pattern will be indicated. In particular, as we intend to cover all forms encountered in the corpora (if they can be considered grammatical), it is necessary to indicate which of these should be considered central (productive) and which are only peripheral.

Link to the source verb and other deverbal counterparts. Every noun will be provided with a link to the verb the noun is derived from and to other nouns derived from the same verb, especially to both non-productively and productively derived counterparts. As nouns derived by non-productive means are not sensitive to aspect they will be provided by links to both perfective and imperfective verbs.

5.2 Corpus-based Valency Lexicon vs. Slovník slovesných, substantivních a adjektivních vazeb a spojení (SSSAVS)

Our corpus-based valency lexicon will differ from the SSSAVS (Svozilová – Prouzová – Jirsová, 2005) especially in following aspects:

Semantic roles and syntactic ambiguity. Concerning nouns, the SSSAVS represents a traditional way of capturing noun valency by giving only examples of particular complementation, regardless of possible combinations with other types of complementation expressed by various forms, such as *lov na medvěda* ‘hunt for bear, i.e. bear hunt’, *lov ryb* ‘hunt of fish, i.e. fishing’ (Svozilová & Prouzová & Jirsová, 2005: 130). The examples convey information about semantic requirements and syntactic forms of the arguments but do not serve as inventory of semantic roles. However, some adnominal forms may occur in constructions that are syntactically ambiguous. This is especially the case of genitives but also of other forms (e.g., in the construction *upozornění řidiče* ‘warning of the driver’ the genitive form *řidiče* ‘of the driver’ can be either ACT or ADDR).

⁶ <http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3>,
<http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4>,
<http://ufal.mff.cuni.cz/pdtsc1.0/>

⁷ <http://korpus.cz/>

Thus we suppose that all valency patterns should be supplemented with the semantic roles of participants.

“Right” as well as “left” valency. The SSSAVS only provides so-called right valency (adnominal counterparts of verbal objects and some adverbial modifications). Adnominal counterparts of verbal agents are not provided at all; the authors suppose that they are used regularly enough and so it would be redundant to supplement all nominal entries with such regular information. However, some combinations of participants are significantly influenced by the presence of an adnominal agent, so we believe that the corpus-based lexicon should provide information on both the “left” and the “right” valency of nouns.

6. Conceptual and Methodological Principles

Vernerová (2011) identifies areas that represent the crucial decisions that have to be made before any lexicographic work is begun. In our lexicon, these decisions will be made as follows:

- (a) The lexicon will target linguists as well as non-linguists. We also envisage NLP applications.
- (b) The lexicon will be organized into alphabetically ordered lexical entries with additional (optional) semantic classification added to selected lexical units (meanings), similarly as in VALLEX (Lopatková & Žabokrtský & Kettnerová, 2008).
- (c) Valency patterns will be described as combinations of syntactic and semantic phenomena.
- (d) Valency slots will be identified by semantic roles defined within the theoretical framework of FGD.
- (e) The described valency behavior and examples will be based on corpus data, in particular on PDT2.0 and PDT3.0, and on the data from the SYN family of CNC corpora.
- (f) The lexicon will provide links to the lemmas of source verbs and to productively or non-productively derived deverbal counterparts.

In addition, we specify the methodology of some partial tasks:

Criteria for the selection of lexemes to be included in the lexicon. In contrast to lexicons that include lexemes exclusively on the basis of their frequency, the lexemes involved in the lexicon will be selected with respect to the following aspects: (i) the semantic class it belongs to; (ii) whether it exhibits typical as well as special valency behavior; (iii) whether it has productively as well as non-productively derived counterparts.

Valency patterns: Utilization of existing valency lexicons. We will look up valency patterns of the source verbs in the existing valency lexicons of Czech, especially in the PDT-Vallex, VALLEX, SSSAVS (Svozilová & Prouzová & Jirsová, 2005) and *Slovesa pro praxi* (Svozilová & Prouzová & Jirsová, 1997). For valency of nouns, we will compare the patterns captured in PDT-Vallex with those present in SSSAVS.

Valency frames. The core valency information will be encoded in valency frames. We will also treat types of valency complementation that are neither participants nor obligatory free modifications but are frequent with the given noun (e.g., *vyznamenání za zásluhy*.CAUS ‘award for merits’). This kind of complementation has not yet been treated within PDT-Vallex valency frames but it is incorporated as the so-called “typical” complementation

in the valency frames in VALLEX.

Criteria for creating a new valency frame. Kolářová (2014) specifies the following reasons for creating a new valency frame:

- (i) a clear change in the meaning of a noun;
- (ii) reduction of the number of valency slots;
- (iii) a change of the character of valency complementation to exclusively nominal (e.g., Patient → Material);
- (iv) a different syntactic behavior of a noun modified by a participant in a special form, when compared with the syntactic behavior of the same noun with the participant in a typical form.

Different meanings of a noun. On the basis of the data of CNC and Czech monolingual dictionaries (especially those available to us in an electronic form), we will identify different meanings of the selected nouns (lexemes). We will focus on lexical units (meanings) with valency potential. We suppose that three basic types of nouns will occur, corresponding to the three types of derivation in Kuryłowicz’s sense (Section B.1.1). In particular, the following three prototypical “meanings” are envisaged: (i) an action or a state (the meaning which is parallel to the meaning of the source verb), (ii) an abstract result of an action (an abstract noun), and (iii) a physical object (a concrete noun).

Consistent treatment of a semantic class. We plan to treat especially the following semantic classes in detail: nouns of communication (e.g., *návrh* ‘proposal’), psychological nouns (e.g., *obava* ‘fear’), nouns of contact (e.g., *dotyk* ‘touch’) and nouns of exchange (e.g., *výdej* ‘distribution’). In order to treat the selected semantic classes consistently we will follow the treatment of verbal valency patterns and their semantic classification applied in VALLEX.

Participants in combinations. The nouns we will focus on are bivalent or trivalent and their participants can often be expressed by several forms that can co-occur in various combinations. However, not all the combinations are grammatical (the systematic restrictions will be specified in the theoretical part of the monograph); for example, this is the case of the combination PAT(POSS) + ACT(GEN) (e.g., **pacientovo.PAT ošetření lékaře.ACT* ‘patient’s treatment of the doctor’), when compared with the combination ACT(POSS) + PAT(GEN) (e.g., *lékařovo.ACT ošetření pacienta.PAT* ‘doctor’s treatment of the patient’). However, there are combinations which are grammatical though not common and their usage should be verified in CNC subcorpora, for example, double post-nominal instrumentals (e.g., *pohrdání názory.PAT veřejnosti vládou.ACT* ‘contempt for opinions of public by the government’). Combinations that we consider to be grammatical but which do not occur in CNC subcorpora will be labelled by a special mark.

Data format and software. The lexicon will be available as machine readable data in a format suitable for NLP applications.

7. A Nominal Entry Example

Such a detailed and comprehensive description of valency behavior of nouns, fulfilling all the tasks given in Section 6, undoubtedly requires considerable amount of effort. We therefore expect that the lexicon will contain only 400-500 nominal lexemes worked out in full detail,

addressing all of the issues in question, including all meanings of the nouns and a detailed analysis of possible combinations of their participants. We present here the envisaged nominal entry for the noun *vyznamenání* ‘honor’ which is an example of a noun that is present neither in the current version of PDT-Vallex nor in SSSAVS. This lexeme represents all three types of derivation of nouns (it denotes an action, an abstract result of an action, and also a physical object). It also displays special valency behavior, in particular the special shift in the form of the Patient (ACC → DAT).

Example of a nominal entry:

Noun: *vyznamenání*^{pf} ‘honor’

Semantic class: evaluation

Source verb: *vyznamenat*^{pf} ‘to honor’

1. proces vyznamenání někoho ‘the process of honoring someone’

Frame: ACT(POSS, GEN, INS) PAT(POSS, GEN)

Example: *vyznamenání veterána.PAT premiérem.ACT* ‘honoring of the veteran by the prime minister’; *vyznamenání premiérem.ACT* ‘honoring by the prime minister’; *premiérovo.ACT vyznamenání veterána.PAT* ‘the prime minister’s honoring of the veteran’; *vyznamenání veterána.PAT* ‘honoring of the veteran’; *?premiérovo.ACT vyznamenání* ‘the prime minister’s honoring’; *veteránovo.PAT vyznamenání premiérem.ACT* ‘the veteran’s honoring by the prime minister’; *veteránovo.PAT vyznamenání* ‘the veteran’s honoring’;

2. pocta, vyznamenání udělení někomu ‘honor, award’

Frame: ACT(POSS, GEN) PAT(DAT)

Type of special valency behavior: special form of PAT

Example: *vyznamenání veteránovi.PAT* ‘honor addressed to the veteran’; *?premiérovo.ACT vyznamenání veteránovi.PAT* ‘the prime minister’s honor addressed to the veteran’; *?vyznamenání premiéra.ACT veteránovi.PAT* ‘honor of the prime minister addressed to the veteran’.

3. odznak, medaile, řád ‘badge, medal, order’

Frame: EMPTY

Example: *ověnčený vyznamenáními* ‘decked with medals’

4. nejvyšší stupeň celkového prospěchu ‘honors’

Frame: EMPTY

Example: *studovat s vyznamenáním* ‘to study with honors’.

8. Nominal Valency Patterns: Searching through Czech Corpora

To exploit corpus data we will use both methods of searching, i.e. manual searching (Section 8.1) and an automatic preprocessing of corpus evidence (Section 8.2). Examples in the resulting lexicon will be extracted from CNC and/or the Araneum corpus⁸ (Benko, 2014).

8.1 Manual Searching

We will take advantage of our experience in searching for nominal valency in lemmatized and morphologically annotated linear corpora such as the SYN family of CNC subcorpora. We will use sophisticated queries that take into account word order variants and include some optional positions (e.g., adjectives modifying the participants) but exclude positions that do not match our

requirements (e.g., a verb between the noun and the participant), see the following example of a query searching for adnominal participants in the form of prepositionless genitive:

```
[lemma="..."] [tag!="[Z|R|V|J].*"]{0,2} [tag="N...2.*"]
```

However, despite carefully prepared queries, a query can often cover various dependency relations that do not match the intention of the query, so all found occurrences have to be manually checked and evaluated. This method is sufficiently precise but the whole procedure of manual searching is very time-consuming.

8.2 Automatic Preprocessing of Corpus Evidence

We will also exploit the Word Sketches (corpus-based summaries of a word’s grammatical and collocational behavior, cf. Kilgarriff and Tugwell, 2001) extracted by Sketch Engine (Kilgarriff et al., 2014). We have⁹ access to two large corpora of Czech (and their subsets): the corpus SYN (2.7 Gigawords) and the corpus Araneum Bohemicum Maximum (3.2 Gigawords). However, the Word Sketch Grammars provided for these corpora are not well suited to the analysis of valency behavior of nouns: the grammar provided for SYN does not contain relations for arguments of nouns expressed by some prepositionless cases (dative, accusative, or instrumental, so the types of valency complementation expressed by these cases are completely missing from the Word Sketch), while the grammar provided for Araneum Bohemicum extracts all nouns to the right of the headword within a single relation, listing only their lemmas (not the actual forms), so it obscures the syntactically and semantically crucial distinction between the arguments expressed by different prepositionless and prepositional cases. For example, the WordSketch of the word *dar* ‘a gift, a present’ lists the lemma *nebesa* ‘heavens, paradise’ under the relations X Y (immediately following word) and X Nn (a noun within three positions to the right). However, the lemma stands here for three different types of complementation which can be distinguished by their morphological form: genitive case *dar nebes* ‘a gift of the heavens’, prepositional case *dar z nebes* ‘a gift from the heavens’, and dative case *dar nebesům* ‘a gift to/for the heavens’. For these reasons, we are currently developing extensions of the existing Sketch Grammars which will be more suited to the analysis of noun valency.

9. Conclusion

Our corpus-based valency lexicon of Czech nouns incorporates elaborate and comprehensive theoretical description of valency behavior of Czech deverbal nouns, utilizes existing Czech valency lexicons, and exploits both Czech linear and syntactically annotated corpora. We believe that work on the lexicon will bring new theoretical findings in the field of nominal valency as well as the useful and versatile lexical resource.

⁹ Because of financial reasons, we depend on the Sketch Engine corpora and functions licensed to the Institute of the Czech National Corpus. Thus, we do not have access to another large corpus of Czech, the czTenTen Web corpus crawled in 2012.

⁸ http://ucts.uniba.sk/aranea_about/index.html

10. Acknowledgements

The research reported in the paper was supported by the Czech Science Foundation under the project GA16-02196S. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2010013 and LM2015071).

11. References

- Alexiadou, A., Rathert, M. (Ed.) (2010). *The Syntax of Nominalizations Across Languages and Frameworks*. Berlin: Walter de Gruyter. ISBN 978-3-11-024586-8.
- Baldwin, T.; Bond, F. and Hutchinson, B. (1999). A Valency Dictionary Architecture for Machine Translation. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, Chester, UK, pp. 207--217.
- Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka et al. (Eds.) *TSD 2014*. LNAI 8655. Springer International Publishing, pp. 247--56.
- Chomsky, N. Remarks on Nominalization. (1970). In R. Jacobs, P. Rosenbaum (Ed.) *Readings in English transformational grammar*. Waltham, MA: Blaisdell, pp. 184--221. ISBN 978-0878401871.
- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge, UK: Cambridge University Press. ISBN 0-521-27643-8.
- Čermák, F. (1991). Podstata valence z hlediska lexikologického. In D. Rytel-Kuc (Ed.) *Walencja czasownika a problemy leksykografii dwujęzycznej*. Wrocław: Wydawnictwo polskiej akademii nauk, pp. 15--40.
- Čermák, F., Holub, J. (2005). *Syntagmatika a paradigmatika českého slova I (Valence a kolokabilita)*. Praha: Karolinum. ISBN 8024609746.
- Čermáková, A. (2009). *Valence českých substantiv*. Praha: Lidové noviny. ISBN 978-80-7106-426-800.
- Dvořáková-Procházková, V. (2008). Argument structure of Czech event nominals. In F. Marušič, R. Žaucer (Ed.) *Contributions from Formal Description of Slavic Languages 6.5*, Bern: Peter Lang, pp. 73--90.
- Dvořák, V. (2014). Case assignment, aspect, and (non-)expression of patients: A study of the internal structure of Czech verbal nouns. In O. Spevak (Ed.) *Noun Valency*, Amsterdam: John Benjamins, pp. 8--112. ISBN 9789027259233.
- Grimshaw, J. (1990). *Argument structure*. Cambridge, MA: MIT Press.
- Hajič, J. et al. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*. Vaxjo University Press, pp. 57--68. ISBN 91-7636-394-5.
- Herbst, T. et al. (2004). *A valency dictionary of English: a corpus-based analysis of the complementation patterns of English verbs, nouns, and adjectives*. Berlin: Walter de Gruyter. ISBN 3-11-017194-5.
- Heyvaert, L. (2003). *A cognitive-functional approach to nominalization in English*. Berlin: Walter de Gruyter. ISBN 3-11-017809-5.
- Ivanová, M.; Sokolová, M.; Kyseřová, M. and Perovská, V. (2014). *Valenčný slovník slovenských slovies na korpusovom základe*. Prešov: Filozofická fakulta Prešovskej univerzity. ISBN 978-80-555-1148-1.
- Jirsová, A. (1966). Vazby u dějových podstatných jmen označujících duševní projevy. *Naše řeč*, 1966, 49, pp. 73--81.
- Karlík, P. (2000). Valence substantiv v modifikované valenční teorii. In Z. Hladká, P. Karlík (Ed.) *Čeština – univerzálie a specifika*, 2. Brno: Vydavatelství MU, pp. 181--192. ISBN 80-210-2262-0.
- Kilgarriff, A., Tugwell, D. (2001). WORD SKETCH: Extraction and display of significant collocations for lexicography. In *Proc Collocations workshop, ACL 2001*, Toulouse, France, pp. 32--38.
- Kilgarriff, A. et al. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7--36.
- Kingsbury, P.; Palmer, M. and Marcus, M. (2002). Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference, HLT-02*. March 24-27, 2002. San Diego, California.
- Klímová, J. (2005). Czech lexical database – Derivation Relations. In P. Bouillon, K. Kanzaki (Ed.) *Third International Workshop on Generative Approaches to the Lexicon*, Geneva, May 19-21 2005, Université de Genève, pp. 119--123.
- Klímová, J. (2010). Český slovtvorný systém 21. století v databázích. In S. Čmejrková, J. Hoffmannová and E. Havlová (Ed.) *Užívání a prožívání jazyka*, Praha: Karolinum, pp. 147--151. ISBN 978-80-246-1756-5.
- Kolářová, V. (2010). *Valence deverbativních substantiv v češtině (na materiálu substantiv s dativní valencí)*. Praha: Karolinum. ISBN 978-80-246-1828-9.
- Kolářová, V. (2014). Special valency behavior of Czech deverbal nouns. In O. Spevak (Ed.) *Noun Valency*, Amsterdam: John Benjamins, pp. 19--60. ISBN 9789027259233.
- Kuryłowicz, J. (1936). Dérivation lexicale et dérivation syntaxique. *Bulletin de la Société de Linguistique de Paris*. 1936, 37, pp. 79--92.
- Křížková, H. (1968). Substantiva s dějovým významem v ruštině a v češtině. In Isačenko, A. V. (Ed.) *Kapitoly ze srovnávací mluvnice ruské a české III. O ruském slovese*, Praha: Academia, pp. 81--152.
- Lopatková, M.; Žabokrtský, Z. and Kettnerová, V. (2008). *Valenční slovník českých sloves*. Praha: Karolinum. ISBN 978-80-246-1467-0.
- Lopatková, M.; Kettnerová, V.; Bejček, E.; Vernerová, A. and Žabokrtský, Z. (2015). *VALLEX 3.0 - Valenční slovník českých sloves*. Charles University in Prague, [online] <http://ufal.mff.cuni.cz/vallex/3.0/>.

- Melloni, C. (2011). *Event and result nominals: a morpho-semantic approach*. Bern: Peter Lang. ISBN 978-3-0343-0658-4.
- Mel'čuk, I.A., Zholkovsky, A.K. (1984). *Explanatory Combinatorial Dictionary of Modern Russian*. Vienna: Wiener Slawistischer Almanach.
- Meyers, A. (2007). *Annotation Guidelines for NomBank – Noun Argument Structure for PropBank*. [online] <http://nlp.cs.nyu.edu/meyers/nombank/nombank-specs-2007.pdf>.
- Mikulová, M. et al. (2006). Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report TR-2006-30, Praha: ÚFAL MFF UK.
- Novotný, J. (1980). *Valence dějových substantiv v češtině*. Sborník pedagogické fakulty v Ústí nad Labem. Praha: SPN.
- Osenova, P. (2009). *Imennite frazi v bulgarskija ezik*. Sofia: ETO.
- Panevová, J. (1974 and 1975). On verbal frames in functional generative description. *Prague Bulletin of Mathematical Linguistics*. Part I: 1974, 22, pp. 3--40. Part II: 1975, 23, pp. 17--37.
- Panevová, J. (2000). Poznámky k valenci podstatných jmen. In Z. Hladká, P. Karlík (Ed.) *Čeština – univerzálie a specifika 2*. Brno: Vydavatelství MU, pp. 173--180. ISBN 80-210-2262-0.
- Panevová, J. a kol. (2014). *Mluvnice současné češtiny 2: Syntax češtiny na základě anotovaného korpusu*. Praha: Karolinum. ISBN 9788024624976.
- Petkevič, V. (2004). Rule-based Part-of-speech and Morphological Disambiguation of the Czech National Corpus. In *Proceedings of the International Conference "Corpus Linguistics – 2004"*, St. Petersburg: St. Petersburg University Press, pp. 271--285.
- Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F. and Świdziński M. (2014). *Walenty: Towards a comprehensive valence dictionary of Polish*. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (Eds) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavík, Iceland: ELRA, pp 2785—2792.
- Ruppenhofer, J. et al. (2006). *FrameNet II: Extended theory and practice*. Berkeley, CA: Computer Science Institute, University of California.
- Sgall, P.; Hajičová, E. and Panevová, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht: Reidel. ISBN 90-277-1838-5.
- Sommerfeldt, K.E., Schreiber, H. (1977). *Wörterbuch zur Valenz und Distribution der Substantive*. Leipzig: VEB Bibliographisches Institut.
- Spevak, O. (Ed). (2014). *Noun valency*. Amsterdam: John Benjamins. ISBN 978-90-272-5923-3.
- Svozilová, N.; Prouzová, H. and Jirsová, A. (1997). *Slovesa pro praxi*. Praha: Academia.
- Svozilová, N.; Prouzová, H. and Jirsová, A. (2005). *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Praha: Academia.
- Vernerová, A. (2011) Nominal Valency in Lexicons. In J. Šafránková, J. Pavlů (Ed.) *Proceedings of the 20th Annual Conference of Doctoral Students - WDS 2011. Part I: Mathematics and Computer Science*. Praha: MATFYZPRESS, pp. 171--176. ISBN 978-80-7378-184-2.
- Vernerová, A.; Kettnerová, V. and Lopatková, M. (2014). To pay or to get paid: Enriching a Valency Lexicon with Diatheses. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavík: ELRA, pp. 2452--2459. ISBN 978-2-9517408-8-4.
- Veselovská, L. (2001). K analýze českých deverbálních substantiv. In Z. Hladká, P. Karlík (Ed.) *Čeština – univerzálie a specifika 3*. Brno: Vydavatelství MU, pp. 11--27. ISBN 80-2001-2532-8.

Ontoterminology meets Lexicography: the Multimodal Online Dictionary of Endometriosis (MODE)

Sara Carvalho¹²³, Rute Costa²³, Christophe Roche³²

¹ 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
Av. 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@fsh.unl.pt, christophe.roche@univ-savoie.fr

Abstract

With the advent of the Semantic Web and, more recently, of the Linked Data initiative, the need to operationalise lexicographic resources, i.e. to represent them in a computer-readable format, has become increasingly important, as it contributes to pave the way to the ultimate goal of interoperability. Moreover, the collaborative work involving Terminology and ontologies has led to the emergence of new theoretical perspectives, namely to the notion of Ontoterminology, which aims to reconcile Terminology's linguistic and conceptual dimension whilst preserving their core identities. This can be particularly relevant in subject fields such as Medicine, where concept-oriented and ontology-based approaches have become the cornerstone of the most recent (bio)medical terminological resources, and where non-verbal concept representations play a key role. Due to the lack of specialised lexicographic resources in the field of endometriosis, this paper aims to present the MODE project, i.e. the Multimodal Online Dictionary of Endometriosis, a multilingual resource comprising several types of data, namely video articles, a new type of scholarly communication in Medicine. It is believed that introducing a medical lexicographic resource supported by ontoterminological principles and encompassing scientific video articles may constitute a relevant window of opportunity in the research field of Lexicography.

Keywords: terminology; ontoterminology; e-lexicography; multimodal dictionary; ontology; endometriosis

1. Introduction

With the advent of the Semantic Web¹ and, more recently, of the Linked Data initiative², the notion of operationalisation, i.e. the creation of computer-readable representations, has become increasingly important, as it contributes to pave the way to the ultimate goal of interoperability.

Moreover, the collaborative work involving Terminology and ontologies – in the sense of Knowledge Engineering (KE) – has led to the emergence of new theoretical perspectives, one of them being Ontoterminology (Roche *et al.* 2009), which aims to reconcile Terminology's linguistic and conceptual dimensions whilst preserving their core identities (Roche (2012, 2015); Costa (2013); Santos & Costa (2015)).

This can be particularly relevant in subject fields such as Medicine, where concept-oriented and ontology-based approaches have become the cornerstone of the most recent (bio)medical terminological resources, and where non-verbal representations play a key role.

It is believed that ontoterminological principles may provide a relevant theoretical and methodological contribution to the research field of Lexicography by supporting the creation of specialised online lexicographic

resources, especially in domains that lack those resources, as is the case with endometriosis. Therefore, this paper aims to present the MODE project, i.e. the Multimodal Online Dictionary of Endometriosis, a multilingual resource comprising several types of data, namely video articles, a new type of scholarly communication in Medicine.

This article will thus be structured as follows: section 2 will focus on the theoretical background, specifically regarding Terminology's double dimension, the Ontoterminology approach and how both can relate to Lexicography; section 3 will provide a brief overview of endometriosis, not only as a subject field *per se*, but also in what concerns the existing specialised lexicographic resources; section 4 will be dedicated to the MODE project, with a description of its supporting principles and core structure, followed by a final section consisting of concluding remarks.

2. Terminology, ontologies and Lexicography

2.1 Terminology's double dimension

As mentioned above, this approach, which encompasses a linguistic and a conceptual dimension that are interrelated, has been described by Roche (2012, 2015), Costa (2013)

¹ Berners-Lee, Hendler, & Lassila (2001); Shadbolt, Hall, & Berners-Lee (2006).

² Berners-Lee (2006); Bizer, Heath, & Berners-Lee (2009).

and by Santos & Costa (2015). According to Roche (2015: 136), Terminology is “both a science of objects and a science of terms”. For Costa (2013), it is precisely this double dimension, as well as the study of the relationship between one and the other, that grants Terminology its place as an autonomous scientific subject.

This double dimension perspective implies, therefore, that both the experts’ conceptualisations of a given subject and the discourses produced by them must be taken into account in terminology work. In a nutshell, the cornerstone of this approach lies in the complementarity of these two fundamentally different dimensions, as two sides of the same coin.

Among the theoretical perspectives that have emerged in recent years involving Terminology and the role of ontologies, Ontoterminology is the one that best suits the objectives of the MODE project, and thus will be presented in more detail below.

2.2 Ontoterminology: a new approach to Terminology?

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

As mentioned before, even though the conceptual dimension plays a key role in Ontoterminology, due to the potential of operationalising the conceptualisations of a given subject field – thus enabling interoperability –, this does not mean that natural language should be excluded from terminology work. In fact, “to conceptualise, one must verbalise” (Roche, 2015: 149). Albeit with vagueness and inconsistencies, the discourses provide fundamental access to the expert community, especially in some areas of expertise where the main goal is knowledge stabilisation and dissemination, as is the case of endometriosis.

Consequently, both specialised texts and expert collaboration constitute invaluable resources in terminological work, provided that there is a supporting theoretical and methodological framework through which it can be possible to maximise the potential of each dimension, and mostly of the synergies resulting from their interaction.

What is important to emphasise, according to Ontoterminology, is that even though the conceptual and linguistic dimensions rely on two diverse semiotic systems that should not be confused³, both of them have their place in projects and products supported by ontoterminological

principles. As a matter of fact, this approach proposes the double semiotic triangle, an extension of Ogden and Richards’s proposition (1923) which allows a distinction between the definition of the term, written in natural language, and the definition of the concept, which may resort to either a formal or a semi-formal language (Roche, 2012). It is believed that when anchored in this approach, terminology work may contribute to further enhance the quality of specialised communication.

2.3 Ontoterminology and Lexicography: is collaboration possible?

As Terminology, in the last few decades, Lexicography has been searching for its identity as an autonomous scientific discipline in its own right, with an intense debate around the principles that should support lexicographic theory and practice (cf. Wiegand 1997, 1998; Bergenholtz & Tarp 2003; Atkins & Rundell 2008; Tarp 2008; Béjoint 2010; Hartmann 2010; Fuertes-Olivera & Bergenholtz 2011; Granger & Paquot 2012; Fuertes-Olivera & Tarp 2014). Part of this discussion pointed, understandably, towards delimiting and positioning Lexicography scientifically, as well as its branches, namely Specialised Lexicography.

In this context, a lot has been written about the need to distinguish Specialised Lexicography from Terminology. They are indeed different, first and foremost because the former studies the units of the specialised lexicon and the way they behave in discourse, whereas the latter focuses not only on the linguistic dimension, but also on a conceptual dimension that cannot be underestimated and is in fact embodied in terms⁴.

However, and despite the differences, some consider that Specialised Lexicography and Terminology are not necessarily incompatible and that both areas could benefit from collaborative work (cf. Humbley 1997; Costa 2013). Fuertes-Olivera & Tarp (2014) refer to the existing interaction between Specialised Lexicography and Terminology in the conception and production of a number of reference works, particularly within the scope of the Function Theory of Lexicography (FTL), although they do not further specify how this interaction actually takes place. As previously stated, this paper intends to show how Terminology, and particularly Ontoterminology, may contribute to the work carried out by Lexicography without undermining both research fields.

First of all, terminology work, as lexicographic practice, relies on a key premise: to have users and their respective needs in mind. In fact, the social responsibility [gesellschaftliche Verantwortung] that, according to Wiegand (1997), should characterise Lexicography as a scientific discipline could also be applied to Terminology. However, it should be noted that, in Terminology, the user may not necessarily be human – at least the primary user –, which will consequently determine the purpose, structure and content of the resource to be developed, as well as the

³ The lexical networks extracted from corpora may not always match the conceptual systems resulting from the collaboration of subject field experts – “Saying is not modelling” (Roche, 2007).

⁴ If, on the one hand, terms are in fact units of discourse, they can

also be perceived as units of representation of the concepts of a given subject field. As such, they have the capacity to exist outside of discourse, pointing towards the concept and providing access into the subject field (cf. Carvalho, Roche, & Costa, 2015).

medium.

Secondly, it is believed that the added value of Terminology in the conception and development of specialised lexicographic resources lies precisely in its double-dimensional nature, and in the fact that the conceptual dimension – substantiated in its knowledge organisation potential – may, in turn, support the linguistic dimension, namely by assisting in the drafting of natural language definitions.

The ontoterminological approach aims to take this contribution to the next level: by placing the ontoterminology at the heart of a given resource, it intends to provide a stable conceptual backbone of a subject field, built in collaboration with subject field experts, and which may become the basis of other, derived products, such as terminology databases, specialised dictionaries, thesauri, etc. The types and amount of data to be made available would then depend on the user profile, on his/her needs, as well as on the social situations and contexts, yet this conceptual core structure, which might or might not be visible to the human user, would remain the same⁵.

As described in the previous subsection, Ontoterminology does not underestimate the linguistic dimension: in fact, it values it, by allowing linguistic diversity to be registered, which is seldom the case in ontology-based approaches. Section 4 will provide an example as to how synonymy and equivalence, for instance, can – and should – have their place within this project.

To sum up, Terminology can play a role in the creation of specialised lexicographic products, both from a linguistic and a conceptual perspective. Within the framework of Ontoterminology, the latter constitutes a valuable foundation which may contribute to enhance the quality of specialised communication.

3. Endometriosis: facts and figures

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

Women with endometriosis typically have a range of pain-related symptoms, such as dysmenorrhea, dyspareunia, dyschezia, dysuria, non-cyclical pelvic pain, as well as chronic fatigue (Dunselman *et al.* 2014). A recent study conducted in 10 countries throughout the world has reported an overall diagnostic delay of 6.7 years (Nnoaham *et al.* 2011). Moreover, the World Endometriosis Research

Foundation (WERF) EndoCost study (Simoens *et al.* 2012) has shown that the costs arising from women with endometriosis treated in referral centres are substantial (an average annual total cost per woman of €9,579), an economic burden that is at least comparable to the costs of other chronic diseases, such as diabetes, Crohn’s disease, or rheumatoid arthritis.

Taking into account the estimated 10% prevalence of the disease among women of reproductive age around the world, which is significant, it is surprising to realise that there are very few specialised language-related resources dedicated to it – lexicographic or of any other nature. In fact, there is, to our knowledge, only one reference work published under the name “Dictionary of Endometriosis” (Parker & Parker 2003), yet this resource is more of an annotated bibliography and a research guide to Internet references concerning the disease. The “dictionary” section is actually a monolingual glossary, in English, with about 1,300 terms and their respective definitions, taken, according to the authors, both from the National Institutes of Health⁶ and the European Union, although it is never mentioned where exactly from the EU these definitions stem from.

An extensive search of resources on endometriosis concluded that the few that actually exist correspond mostly to the notion of glossary, perceived as a “list of designations and definitions in a particular subject field” (ISO 1087-1, 2000: 12). These lists are almost exclusively monolingual (with English as the most frequently used language), depicting a widely variable number of terms (ranging from 20 to 1,500), usually containing no sources in what concerns the definitions, and with hardly any supplementary material, namely images or videos. In addition, these resources have been built by and are destined to different types of people, and have therefore fairly distinct levels of specialisation. Some examples include the European Society for Human Reproduction and Embryology’s Guideline on the Management of Women with Endometriosis (Dunselman *et al.*, 2013) (expert > expert or > semi-expert) and the American Society for Reproductive Medicine’s Endometriosis Guide for Patients (ASRM 2012) (expert > non-expert).

4. The MODE project

As previously mentioned, the main goal of this paper is to present the Multimodal Online Dictionary of Endometriosis (MODE), a project of a multilingual resource based around the concept of <Endometriosis>⁷, which is currently at its conception stage and aims to integrate several types of data, including medical video articles⁸.

⁵ Assuming that the knowledge in that particular domain is stable enough to be represented via a semi-formal or formal conceptualisation.

⁶ That belong to the U.S. Department of Health and Human Services and integrate the National Library of Medicine, responsible for issuing and updating PubMed/MEDLINE, Medical Subject Headings (MeSH), and MedlinePlus.

⁷ In this paper, concepts will be capitalised and written between single chevrons, whereas terms will be presented in lower case and between double quotation marks (cf. Roche, 2015).

⁸ This peer-reviewed and indexed resource has been described more thoroughly in Carvalho, Roche, & Costa (forthcoming). MODE’s guiding principles have been defined within the scope of the EndoTerm project, presented in Carvalho, Roche, & Costa

Even though the inclusion of images and multimedia content is not new in medical lexicographic resources (whether in CDs and DVDs, in online editions, or more recently in apps, as is the case of the renowned Stedman's Taber's and Dorland's Medical Dictionaries, just to name a few), MODE can offer added value supported by three essential axes: the inclusion of medical video articles and the emphasis on their potential as a new type of scholarly communication in Medicine; the choice of the subject field itself, which lacks specialised resources; and finally, its ontoterminological principles, grounded in Terminology's double dimension.

This resource, aimed primarily at future experts⁹ (medical students) or experts of other, related domains (such as nursing staff, for example), can make a valuable contribution in specialised training, which is why expert collaboration plays a critical role in helping identify relevant and realistic needs in this particular subject field. As for the situations which may lead to the consultation of such a lexicographic resource, and using the terminology adopted by the FTL (cf. Tarp, 2008; Fuertes-Olivera & Tarp, 2014), it is believed that MODE's potential users will be mainly interested in acquiring knowledge about a particular subject (cognitive situation), rather than, for example, trying to solve a communication problem (communicative situation).

In the next few paragraphs, MODE's core structure, as well as a methodological proposal, will be put forward. Due to space constraints, the examples to be provided will focus on the concept of <Laparoendoscopic single-site surgery>, a relatively recent type of surgical procedure that is becoming more and more prevalent in several medical specialties, and that accounts for a significant amount of endometriosis-related surgeries (Gill *et al.* 2010).

The conceptual structure of the domain is MODE's "beating heart", providing, as stated in section 2, a backbone that supports the remaining components. As such, it constitutes the first by-product of the project, and has been built using OTE Soft ©, a concept modelling tool created by the Condillac Research Group (cf. Roche, 2015; Carvalho, Roche, & Costa 2015). Based on information provided by textual and multimedia sources, by current (bio)medical terminological resources (such as MeSH, UMLS and SNOMED CT), as well as by the feedback and validation of a team of senior expert gynaecologists, a set of concept maps have been created, one of which is shown below.

Figure 1 depicts the concept of <Laparoendoscopic single-site surgery> and aims to position it within the broader concept of <Surgical procedure> by resorting to a specific differentiation, Aristotelian-based approach. Through its analysis, it is possible to conclude that the existence of a single skin incision constitutes the essential characteristic (cf. ISO 1087-1: 2000) of this type of surgical procedure. Other characteristics comprised in the wider concept map and identified, among other sources, by a White Paper published by the Laparoendoscopic Single-Site Surgery Consortium for Assessment and Research (LESSCAR) (cf. Gill *et al.*, 2010), include: i) the type of surgery (laparoscopic, endoscopic or robotic); ii) the location of the skin incision (abdominal, thoracic or pelvic); or iii) the type of surgical approach (percutaneous intraluminal or percutaneous transluminal).

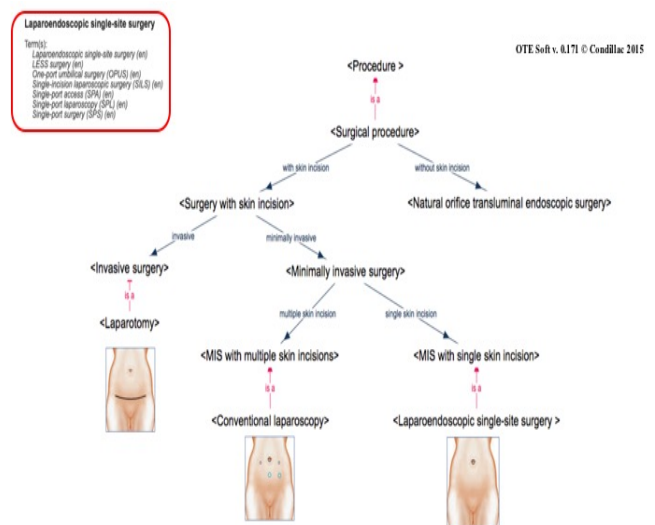


Figure 1: Concept map of <Laparoendoscopic single-site surgery>.

The project is currently at the beginning of its second stage, consisting of corpus collection and analysis, in order to see whether the selected texts contain designations that point towards the previously identified concepts. There are three working languages (English – eng, European Portuguese – pt and French – fr) involved and in this respect, the experts play a critical role, namely in advising as to the texts that are deemed representative and/or mandatory in the subject field of endometriosis. Based on their feedback, a text typology, i.e. "la réunion et la classification d'un ensemble des textes sous une même étiquette" (Costa & Silva 2008: 6) has been created, integrating 3 main types of texts: a) academic (comprising scientific articles and theses); b)

2015).

⁹ Although this goes beyond the scope of this paper, it is believed that the notions that have characterised the types of users of specialised lexicographic or terminological products, namely the distinction between subject field experts, semi-experts and non-experts or laypeople (cf. Bergenholtz & Kaufmann 1997), are becoming more and more blurred, at least in some areas, and should therefore be discussed. Within the (bio)medical domain, for instance, one could assume that a patient would belong to the last group. However, growing digital literacy has brought the patients into the driver's seat and has led them to play a more

active – and empowered – role. In fact, patient empowerment has been at the heart of the most recent healthcare policies and initiatives, particularly at a European level (<http://www.eu-patient.eu>). One of the most promising projects in this respect is the European Patient Academy (EUPATI), which provides Patient Expert Training Courses destined to increase "the capacity and capability of patients to understand and contribute to medicines' R&D" and also to improve "the availability of objective, reliable, patient-friendly information for the public". More information at <https://www.eupati.eu>.

normative (guidelines, White Papers, and standards); and c) teaching materials (textbooks, handbooks or course books). The subsequent corpus treatment and analysis are to be conducted using AntConc® and a set of candidate terms is to be presented to the experts for validation.

The next step consists of the development of the dictionary entries. At the moment, a study is being carried out regarding the layout of those upcoming entries, specifically the structure that could best suit the resource’s guiding principles, including the proposition of one entry per concept and the need for interoperability¹⁰. Thus, and as it is currently not possible to present an actual entry of the MODE, the example below resorts to CMap Tools©¹¹ – more specifically, by focusing on the central concept of this paper. This proposal includes the term in English and its synonyms, its equivalents in European Portuguese and French, as well as a definition, with the concept as core element (Check figure 2).

As mentioned before, the ontoterminological approach enables the existence of both a term and a concept definition. However, as the collection of the English corpus has not been completed up to the present moment, a natural language definition cannot be provided. Still, the designed micro-concept map containing the concept’s essential and delimiting characteristics (cf. ISO 1087-1, 2000; ISO 704, 2009) may contribute to enhance the quality of an existing natural language definition or to actually create a new one if none exists.

As regards the linguistic dimension, i.e. the term(s) designating the concept in question, a lack of terminological consensus among the expert community has been identified, with a plethora of terms coined by individual groups and organisations. In fact, more than 20 have been documented in the literature, as shown in the table below.

Abbreviated form	Full form
LESS	laparoendoscopic single-site surgery
NOTUS	natural orifice transumbilical surgery
OPUS	one-port umbilical surgery
S3	single-site surgery
SAS	single-access surgery
SAVES	single-access video endoscopic surgery
SILS	single-incision laparoscopic surgery
SIMIS	single-incision minimally invasive surgery
SIS	single-incision surgery
SLaPP	single laparoscopic port procedure
SLIT	single laparoscopic incision transabdominal surgery
SPA	single-port access
SPICES	single-port incisionless conventional equipment-utilizing surgery
SPL	single-port laparoscopy
SPLS	single-port laparoscopic surgery
SPS	single-port surgery
SSA	single-site-access laparoscopic surgery
SSL	single-site laparoscopy
SSULS	single-site umbilical laparoscopic surgery
TUES	transumbilical endoscopic surgery
TULA	transumbilical laparoscopic assisted surgery

Sources: Box et al. (2006); Gill et al. (2010); Flanesh et al. (2014); Georgiou et al. (2012); Autorino et al. (2011); Springborg & Faderl (2015); Escobar & Falcone (2014)

Table 1: Terms designating the concept of <Laparoendoscopic single-site surgery>.

In order to solve this terminological dispersion, the aforementioned LESSCAR proposed the term “laparoendoscopic single-site surgery” as the one that most accurately depicted this surgical procedure. The remaining designations can be perceived as synonyms, which, from a terminological point of view, raises the dilemma of whether apples are indeed being compared to apples, i.e. whether or not all these terms are in fact representing the same concept. A more thorough analysis of this subject, which will occur after the corpus analysis is completed, is necessary in order to confirm this hypothesis and further develop it. Therefore, our “entry” proposal contains two randomly selected terms as synonyms for the term in English.

Concerning the equivalents, the data gathering accomplished thus far has confirmed the significant discrepancy between the English-speaking corpus and the ones in French and European Portuguese, which can be explained by the predominance of English in specialised communication, particularly in the academic world¹². Moreover, in the fr and pt texts compiled so far, no equivalents of <Laparoendoscopic single-site surgery> have been found. Consequently, further research of academic texts (theses and scientific articles), as well as of teaching materials, was conducted in those two languages. The search was carried out via Google’s advanced search and the results indicate that in pt, the most frequent designation was “cirurgia laparoscópica por porta umbilical única” [single-port umbilical laparoscopic surgery], mainly within the medical specialty of Urology, whereas in fr, the term “chirurgie par accès unique” [single-access surgery] was the one most widely used. Potential synonyms have also been found in both languages, but appear to raise the same dilemmas as those mentioned above: “laparoendoscopia de incisão única” [single-incision laparoendoscopy], “cirurgia por incisão única” [single-incision surgery] (pt); “chirurgie laparoscopique par accès ombilical unique” [single umbilical access laparoscopic surgery], “chirurgie laparoscopique à trocart unique” [single-port laparoscopic surgery], “chirurgie par orifice unique” [single-orifice surgery] (fr).

Term: laparoendoscopic single-site surgery Source: LESSCAR White Paper – http://link.springer.com/article/10.1007/s00464-009-0688-8
Definition: to be included
Synonym 1: single-port access Synonym 2: single-port laparoscopy
pt: cirurgia laparoscópica por porta umbilical única Source pt: http://www.apurologia.pt/acta/2-2012/cirur-lap-port-unic-umb.pdf
fr: chirurgie par accès unique Source fr: http://ao.um5.ac.ma/xmlui/bitstream/handle/123456789/14853/Mk2085%202015%20.pdf?sequence=1&isAllowed=y
by: Sara Carvalho
e-mail: saramcarvalho@gmail.com

Figure 2: Entry proposal for MODE.

¹⁰ The ISO 1951:2007 standard, for instance, may not suit our needs, as it explicitly mentions its “lexicographical lemma-oriented approach”, hence distancing itself from “concept-oriented works”.

¹¹ A freely available software developed by the Florida Institute for Human and Machine Cognition (IHMC) and available at

<http://cmap.ihmc.us>.

¹² As this task has not yet been completed, it is not possible to present the definitive figures.

Having a conceptual framework as the basis of the MODE project can also contribute to facilitate and improve the insertion of supplementary material, such as images, diagrams, and videos, by acting as a sort of “tag”. Relying on a validated knowledge organisation proposal enables the inclusion of the aforementioned resources in a much more thorough way, which will undoubtedly be useful for a group of intended users seeking for detailed subject field knowledge. In addition, it can lead to a more effective customisation of the MODE.

Let us take the following example: as stated beforehand, the LESS technique is very often used in endometriosis-related surgical procedures, namely in hysterectomies, often seen as a last resort in cases where the disease strikes more severely. However, there are different types of hysterectomy (supracervical or partial, total and radical) and if a given video article describes, for instance, a supracervical hysterectomy using LESS, it is possible to add that video to the actual concept being depicted <LESS supracervical hysterectomy> and not to the more generic concept <LESS hysterectomy> or, going even further up, <Hysterectomy> (check Figure 3).

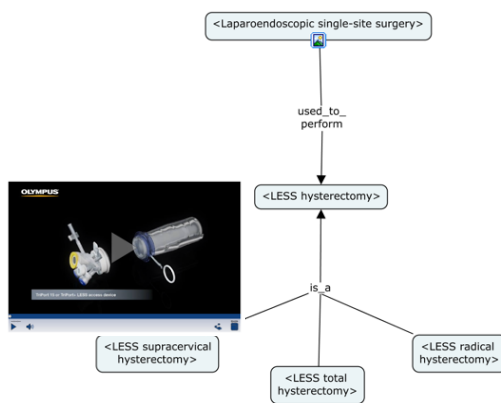


Figure 3: Micro-concept map of the different types of <LESS hysterectomy> and example of video insertion.

This issue is even more pressing when interoperability is at stake. In a study involving the creation of a dictionary for sign language, Kristoffersen and Troelsgard (2012) refer to the unsuitability, from a computational perspective, of video recordings as lemmas in a dictionary database, as they “would have to be represented by a transcription, a filename, a number, or some other sort of ID in order to be ordered or filtered” (296). A conceptual framework within the ontoterminological approach would actually enable that operationalisation, i.e. that computational representation. Furthermore, and although this is not the focus of the current project, it is also believed that the experience resulting from the inclusion of medical video articles in MODE will constitute the starting point for further projects, substantiated in the content analysis and tagging of these video articles, which may then supply inputs regarding the classification, indexing and archive of these multimedia resources.

5. Concluding remarks

Through the presentation of the MODE project, this paper intended to show that the ontoterminological approach can make a valuable contribution to the field of Lexicography. Rather than being perceived as incompatible, both areas combined provide added value to a research project and these synergies will certainly represent a window of opportunity in the conception and development of online specialised resources. As Gouws (2011: 29) points out, “looking to the future, (...) we must unlearn a great deal of what we know, and we must learn anew so that we can produce innovative reference tools, including dictionaries.”

6. 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.

7. Bibliographical References

- Adamson, G., S. Kennedy, and L. Hummelshoj. 2010. “Creating Solutions in Endometriosis: Global Collaboration through the World Endometriosis Research Foundation.” *Journal of Endometriosis* 2 (1): 3–6.
- American Society for Reproductive Medicine. 2012. *Endometriosis - A Guide for Patients*. Patient edition. ASRM.
- Atkins, S., and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Béjoint, H. 2010. *The Lexicography of English*. Oxford: Oxford University Press.
- Bergenholtz, H., and U. Kaufmann. 1997. “Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field.” *Hermes, Journal of Linguistics* 18: 91–125.
- Bergenholtz, H., and S. Tarp. 2003. “Two Opposing Theories: On H.E. Wiegand’s Recent Discovery of Lexicographic Functions.” *Hermes, Journal of Linguistics* 31: 171–96.
- Berners-Lee, T. 2006. “Linked Data.” <http://www.w3.org/DesignIssues/LinkedData.html>. Accessed March 4, 2013.
- Berners-Lee, T., J. Hendler, and O. Lassila. 2001. “The Semantic Web.” *Scientific American*. <http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html>. Accessed March 4, 2013.
- Bizer, C., T. Heath, and T. Berners-Lee. 2009. “Linked Data - The Story So Far.” *International Journal on Semantic Web and Information Systems (IJSWIS)*, no. Special Issue on Linked Data. Accessed March 4, 2013.

- Carvalho, S., C. Roche, and R. Costa. 2015. "Ontologies for Terminological Purposes: The EndoTerm Project." In *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence - Universidad de Granada, Granada, Spain, November 4-6, 2015.*, edited by T. Poibeau and P. Faber, 17–27. Granada: CEUR Workshop Proceedings.
- . (forthcoming). "Why Read When You Can Watch? Video Articles and Knowledge Representation within the Medical Domain." In *Proceedings of the 2015 TOTH Conference*. Chambéry: Équipe Condillac / Université Savoie-Montblanc.
- Costa, R., and R. Silva. 2008. "De La Typologie à L'ontologie de Textes." In *TOTH 2008. Actes de La Deuxième Conférence TOTH - Annecy - 5-6 Juin 2008*, edited by Christophe Roche, 3–16. Annecy: Institut Porphyre - Savoir et Connaissance.
- Costa, R.. 2013. "Terminology and Specialised Lexicography: Two Complementary Domains." *Lexicographica* 29 (1): 29–42.
- Dunselman, G. et al. 2014. "ESHRE Guideline: Management of Women with Endometriosis." *Human Reproduction* 29 (3): 400–412.
- Fuertes-Olivera, P., and H. Bergenholtz, eds. 2011. *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London / New York: Continuum.
- Fuertes-Olivera, P., and S. Tarp. 2014. *Theory and Practice of Specialised Online Dictionaries. Lexicography versus Terminography*. Lexicograp. Berlin / Boston: De Gruyter Mouton.
- Gill, I. et al. 2010. "Consensus Statement of the Consortium for Laparoendoscopic Single-Site Surgery." *Surgical Endoscopy* 24 (4): 762–68.
- Gouws, R. 2011. "Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries." In *eLexicography. The Internet, Digital Initiatives and Lexicography*, edited by P. Fuertes-Olivera and H. Bergenholtz, 17–29. London / New York: Continuum.
- Granger, S., and M. Paquot, eds. 2012. *Electronic Lexicography*. Oxford Lin. Oxford: Oxford University Press.
- Hartmann, R. 2010. "Has Lexicography Arrived as an Academic Discipline? Reviewing Progress in Dictionary Research During the Last Three Decades." In *Nordiska Studier I Lexikografi 10 (Proceedings of Tampere 2009 Conference of NFL)*, edited by H. Loenroth and K. Nikula, 11–35. Tampere.
- Humbley, J. 1997. "Is Terminology Specialized Lexicography? The Experience of French-Speaking Countries." *Hermes, Journal of Linguistics* 18: 13–31.
- International Health Terminology Standards Development Organisation. "SNOMED CT Browser." <http://browser.ihtsdotools.org/>. Accessed February 1 2015.
- International Organization for Standardization. 2000. "ISO 1087-1: Terminology Work - Vocabulary - Part 1: Theory and Application." Geneva: ISO.
- . 2009. "ISO 704: Terminology Work - Principles and Methods." Geneva: ISO.
- Kennedy, S. et al. 2005. "ESHRE Guideline for the Diagnosis and Treatment of Endometriosis." *Human Reproduction* 20 (10): 2698–2704.
- Kristoffersen, J., and T. Troelsgard. 2012. "Electronic Sign Language Dictionaries." In *Electronic Lexicography*, edited by S. Granger and M. Paquot, 290–312. Oxford: Oxford University Press.
- Nnoaham, K. et al. 2011. "Impact of Endometriosis on Quality of Life and Work Productivity: A Multicenter Study across Ten Countries." *Fertility and Sterility* 96 (2): 366–73.
- Roche, C. 2007. "Saying Is Not Modelling." In *Proceedings of NLPCS 2007 (Natural Language Processing and Cognitive Science), Funchal, June 2007*, 47–56. Funchal.
- . 2012. "Ontoterminology: How to Unify Terminology and Ontology into a Single Paradigm." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, 2626–30. Istanbul: European Language Resources Association (ELRA).
- . 2015. "Ontological Definition." In *Handbook of Terminology - Vol. 1*, edited by H. J. Kockaert and F. Steurs, 128–52. Amsterdam: John Benjamins Publishing Company.
- Roche, C. et al. 2009. "Ontoterminology: A New Paradigm for Terminology." In *International Conference on Knowledge Engineering and Ontology Development, Oct 2009*, 321–26. Funchal.
- Santos, C. and R. Costa. 2015. "Domain Specificity: Semasiological and Onomasiological Knowledge Representation." In *Handbook of Terminology - Vol. 1*, edited by H. J. Kockaert and F. Steurs, 153–79. Amsterdam: John Benjamins Publishing Company.
- Shadbolt, N., W. Hall, and T. Berners-Lee. 2006. "The Semantic Web Revisited." *IEEE Intelligent Systems*, no. May/June: 96–101.
- Simoens, S. et al. 2012. "The Burden of Endometriosis: Costs and Quality of Life of Women with Endometriosis and Treated in Referral Centres." *Human Reproduction* 27 (5): 1292–99.
- Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-Knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer Verlag.
- US National Library of Medicine. "UMLS Terminology Services." <https://uts.nlm.nih.gov/home.html>. Accessed May 4, 2015.
- . "MeSH Browser (2016)" <https://www.nlm.nih.gov/mesh/MBrowser.html>. Accessed March 4, 2016.
- Wiegand, H. 1997. "Über die gesellschaftliche Verantwortung der wissenschaftlichen Lexikographie." *Hermes, Journal of Linguistics* 18: 177–201.

———. 1998. *Wörterbuchforschung: Untersuchungen Zur Wörterbuchbenutzung, Zur Theorie, Geschichte, Kritik Und Automatisierung Der Lexikographie. 1. Teilband*. Berlin: Walter de Gruyter.

8. Language Resource References

“Dorland’s Illustrated Medical Dictionary.” *Elsevier*. Accessed January 10, 2016. <http://www.dorlands.com>.

Parker, J., and P. Parker, eds. 2003. *Endometriosis: A Medical Dictionary, Bibliography, and Annotated Research Guide to Internet References*. San Diego: ICON Health Publications.

“Stedman’s Medical Dictionary.” *Wolters Kluwer*.

Accessed February 20, 2016.

<http://www.stedmansonline.com>.

“Taber’s Medical Dictionary Online.” *F.A. Davis Company / Unbound Medicine*. Accessed February 3, 2016.

<http://www.tabers.com/tabersonline/>.

Verb Argument Pairing in Czech-English Parallel Treebank

Jan Hajič, Eva Fučíková, Jana Šindlerová, Zdeňka Urešová

Charles University in Prague, Faculty of Mathematics and Physics

Institute of Formal and Applied Linguistics

Malostranské nám 25, 11800 Prague 1, Czech Republic

{hajic,fucikova,sindlerova,uresova}@ufal.mff.cuni.cz

Abstract

We describe CzEngVallex, a bilingual Czech-English valency lexicon which aligns verbal valency frames and their arguments. It is based on a parallel Czech-English corpus, the Prague Czech-English Dependency Treebank, where for each occurrence of a verb a reference to the underlying Czech and English valency lexicons is explicitly recorded. CzEngVallex lexicon pairs the entries (verb senses) of these two lexicons, and allows for detailed studies of verb valency and argument structure in translation. While some related studies have already been published on certain phenomena, we concentrate here on basic statistics, showing that the variability of verb argument mapping between verbs in the two languages is richer than it might seem and than the perception from the studies published so far might have been.

Keywords: lexical resources, parallel corpus, treebank, valency, bilingual valency lexicon, Czech, English

1. Introduction

Valency, or verb argument structure, is an important phenomenon both in linguistic studies as well as in language technology applications, since verb is considered the core of a clause in (almost) every natural language utterance. Various lexicons have been built - from Propbank (Palmer et al., 2005) to Framenet (Baker et al., 1998). Various valency lexicons exist for several languages, such as Walenty (Przepiórkowski et al., 2014) for Polish, and several exist also for Czech: primarily VALLEX (Žabokrtský and Lopatková, 2007) and Verbalex (Horák, Aleš and Pala, Karel and Hlaváčková, Dana, 2013). However, there are no truly multilingual valency lexicons, and none link parallel corpora together through valency lexicons the way the CzEngVallex lexicon does, as described in (Urešová et al., 2015a) and analyzed in this paper. It thus offers an opportunity to learn not only about valency as generalized across languages, but also to study translation from a different perspective thanks to the explicit references between the parallel Czech-English corpus and the valency lexicons for the two languages.

In this paper, we briefly describe the resources and their interplay, and then analyze the CzEngVallex lexicon in more detail, showing also examples of the (mis)match of verb valency between the two languages.

2. The PCEDT parallel corpus

The Prague Czech-English Dependency Treebank (PCEDT 2.0) (Hajič et al., 2012) contains the WSJ part of the Penn Treebank (Marcus et al., 1993) and its manual professional translation to Czech, annotated manually using the tectogrammatical representation (Mikulová et al., 2005), first used for the Prague Dependency Treebank 2.0 (PDT) (Hajič et al., 2006).

2.1. PCEDT: the annotation scheme

The PCEDT contains 866,246 English tokens and 953,187 Czech tokens, aligned manually sentence-by-sentence and

automatically word-by-word. It is annotated on all three annotation layers of the PDT: morphological, analytical (surface dependency syntax) and tectogrammatical (syntactic-semantic). However, as opposed to the PDT which is annotated fully manually,¹ PCEDT has been annotated for structure and valency at the tectogrammatical representation layer manually, but for POS and morphology and surface syntax only automatically.² Both language sides of the tectogrammatical representation have been enriched with valency annotation, using two valency lexicons: PDT-Vallex for Czech and EngVallex for English. Fig. 7 shows an example of an annotated pair of aligned sentences in the PCEDT (together with visualized CzEngVallex projection, see below Sect. 3.).

2.2. PDT-Vallex: Czech valency lexicon

The PDT-Vallex (Hajič et al., 2003; Urešová, 2011b; Urešová, 2011a) has been originally developed for the PDT annotation. It contains 12,000 verb frames for about 7,000 verbs, roughly corresponding to verb senses found during the annotation of the PDT and PCEDT treebanks. For each frame, verb arguments are listed together with the obligatoriness and constraints on surface morphosyntactic realization; examples and notes are given for each entry as well. Each occurrence of a verb in the PDT (and on the Czech side of the PCEDT) is linked to one verb frame in the PDT-Vallex lexicon. The same lexicon has also been used for the annotation of spoken Czech in the Prague Dependency Corpus of Spoken Czech, or PDTSC³ (Hajič et al., 2009).

2.3. EngVallex: English valency lexicon

The EngVallex (Cinková, 2006) has been created for the English side of the PCEDT annotation. It is a semi-manual conversion of the Propbank frame files (Palmer et al., 2005) into the PDT style of capturing valency information in valency frames. The correspondence of the original Propbank

¹With the exception of certain lexical node attributes.

²The surface dependency syntax on the English side has been derived from the Penn Treebank constituent syntax annotation, using head percolation rules, and thus can be considered semi-manual as well.

³<http://ufal.mff.cuni.cz/pdtsc1.0/en/index.html>

entries and valency frames in EngVallex is not necessarily 1:1 - entries have been occasionally merged or split. It contains over 7,000 frames for 4,300 verbs.

2.4. Treebank-lexicon links and lexicon entries

From the point of view of valency in general and this paper in particular, the most important part of the annotation of the corpus and its relation to the valency lexicons is the treatment of verb arguments and adjuncts. Every (non-auxiliary) verb node in the treebank refers to one particular sense of that verb in the respective valency lexicon (PDT-Vallex or EngVallex). The nodes dependent on the verb in the annotation are obligatory or optional complementations. All actants⁴ and other obligatory complementations (we will call them collectively “arguments” for simplicity)⁵ are also recorded in the valency lexicon(s). In other words, the valency lexicon entry matches the verb-rooted subtree of the annotated tectogrammatical tree linked to it.

The “core” arguments (“actants” in the tectogrammatical terminology) are Actor (or deep subject, or first argument, ACT), Patient (deep object, or second argument, PAT), Addressee (ADDR), Effect (EFF) and Origin (in the transformational sense, such as *create a doll from wood*, labeled ORIG). Non-core arguments often deemed obligatory with certain verbs and their senses are Location (LOC), Direction-from (DIR1), Direction-to (DIR3), Manner (MANN), Beneficiary (BEN) and several others.



Figure 1: PDT-Vallex example entry of the valency frame for *respektovat* (lit. *respect, heed, honor*)

An example of a valency entry for the Czech verb *respektovat* is in Fig. 1. Since Czech is an inflective language and morphosyntactic features are essential for the description of verb arguments, they are listed in the lexicon entry as well, following the argument label (e.g., for the Patient argument in the figure, the number “4” means accusative case, and the arrows are used to specify that the argument can also be expressed as a subordinate clause, in this case using either the conjunction “že” or “když”).⁶

3. The CzEngVallex lexicon

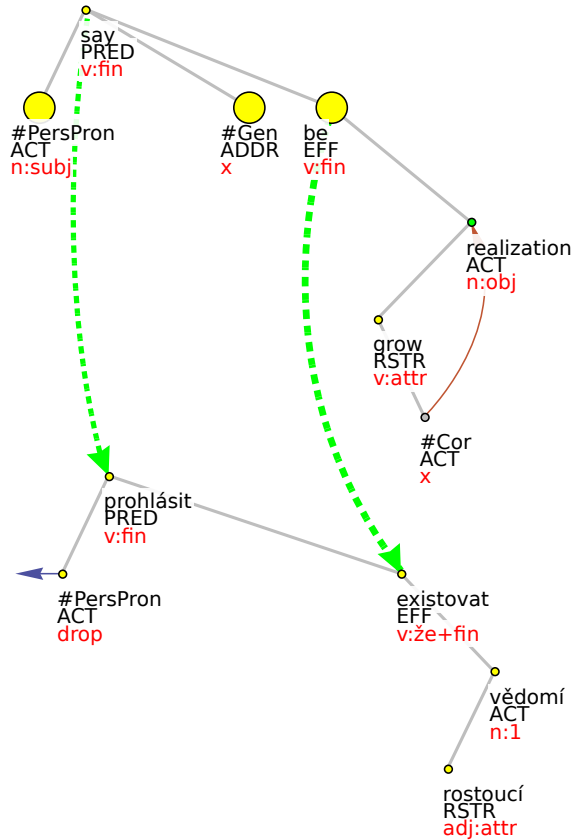
The CzEngVallex lexicon (Urešová et al., 2015a; Urešová et al., 2015b)⁷ is a bilingual valency lexicon with explicit

⁴Sometimes called “core” arguments, see below for a list.

⁵The distinction between arguments and adjuncts is often understood differently by different authors, but that is not the important point; here, our use of “argument” is wider than usual, as it gets clearer later.

⁶Frequencies in the PDT and PCEDT treebank are included as well, and so are synonyms and a human-readable description or definition of the particular verb sense, especially to distinguish entries of polysemous verbs.

⁷Available publicly for download from the <http://lindat.cz> repository, together with the monolingual valency lexicons and



En: She said there is a “growing realization”...
Cz: Prohlásila, že ... existuje “rostoucí vědomí”...

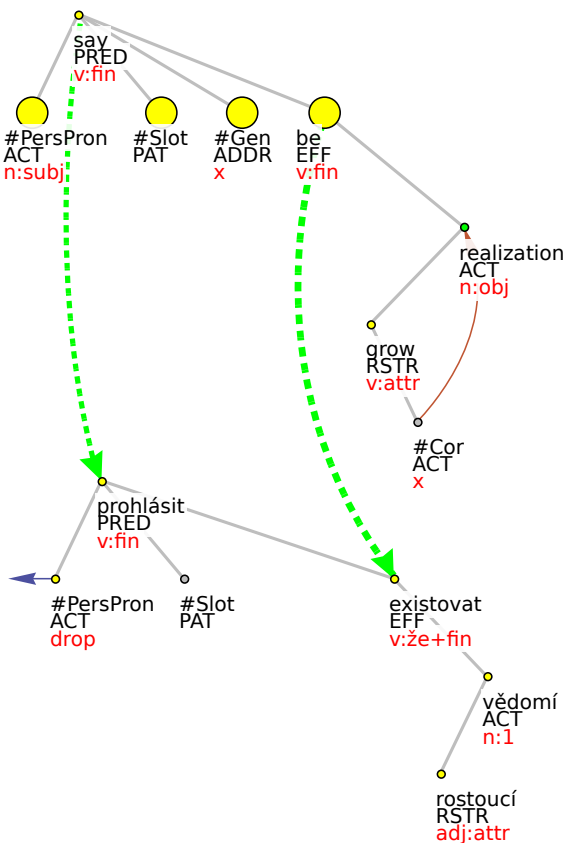
Figure 2: Verb and argument pairs suggested by the automatic preprocessing step (green arrows)

pairing of verb senses (corresponding to valency frames) and their arguments, built upon the Prague Czech-English Dependency Treebank (PCEDT), as described in the previous section. It contains 20,835 frame pairs. It should be noted that not all verbs from the PCEDT can be found in the CzEngVallex: some verbs have not at all been translated as verbs, and vice versa, and some verb translations have been so structurally different that even if translated as verbs, they have not been included in the CzEngVallex. According to (Urešová et al., 2015a), 71% of English verb tokens found in the corpus have been aligned and can be found in the CzEngVallex (for Czech verb occurrences, it is 77%). Also, due to the fact that the CzEngVallex is restricted to the parallel corpus only, it also covers only about 2/3rd of the underlying valency lexicons, i.e., PDT-Vallex and EngVallex. Exact statistics are given in Table 1 (Urešová et al., 2015a).

Language	Verb types	Frame types	PCEDT Tokens	
			verbs	aligned
English	3,292	5,010	130,514	92,747
Czech	4,218	6,930	118,189	91,656

Table 1: Alignment coverage - CzEngVallex/PCEDT

the PCEDT corpus.

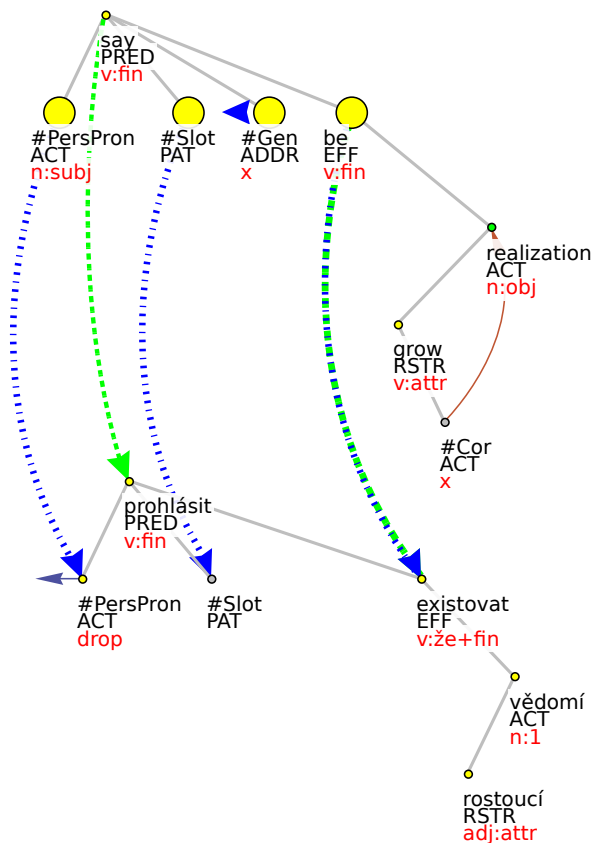


En: She said there is a “growing realization”...
 Cz: Prohlásila, že ... existuje “rostoucí vědomí”...

Figure 3: Verb and argument pairs after insertion of elided valency slots

3.1. Building CzEngVallex

CzEngVallex has been built, as it has been mentioned above, on top of the Prague Czech-English Dependency Treebank. The corpus, annotated manually for (monolingual) valency, has been first (automatically) pre-processed to align all nodes of the tectogrammatically-annotated trees, and all trees which contained at least one verb-verb pair have been extracted, re-sorted to show all pairs of trees with the same sense of the English verb in one group, and passed to the CzEngVallex annotators. Fig. 2 shows an example (fragment) of a sentence pair (Eng: *She said there is a “growing realization” ...*) containing the verb *say* and its translation (*prohlásit*, in this sentence), as displayed for the annotator, with green arrows showing pre-aligned verb and argument pairs. The main task of the annotators has been to check the pairings of both verbs and their arguments, and to add or correct them if necessary. The underlying hypothesis which has determined the design of the valency frame pairing scheme was that for each verb sense pair, the alignment of their arguments is the same (otherwise, the verb sense on one or both sides would have to be refined). This was the key point of the annotation, apart from corrections of the errors of the original automatic node alignment or corrections



En: She said there is a “growing realization”...
 Cz: Prohlásila, že ... existuje “rostoucí vědomí”...

Figure 4: Verb and argument pairs as marked by the annotator (blue arrows) for entering them into CzEngVallex

of the treebank annotation itself.⁸ In our example sentence, the annotator fills in missing (non-overt) arguments for both *say* and its Czech translation, namely, the deep object (PAT, with the lexeme represented only as #Slot, see Fig. 3). After filling in all of the elided valency slots, the annotator adds alignment links for the newly introduced arguments and for those that have not been identified by the automatic preprocessing step. In the displayed case, the nodes with ACT and PAT have been aligned and the ADDR node has been marked as non-corresponding to any Czech argument (Fig. 4, blue arrows).

Only after a careful review of the whole group of *all* PCEDT examples for the given pair of verb senses and their valency frames the alignment of the arguments has been confirmed by the annotator and the valency frame pair entered into CzEngVallex.

3.2. Annotation rules in specific cases

Due to slight inconsistencies in the handling of verb arguments and adjuncts on the two sides of the PCEDT, the annotation rules had to be gradually extended to contain con-

⁸To keep the annotation consistent, corrections in the treebank have only been suggested and passed to the treebank maintainers to include them in the next version, i.e., the underlying treebanks have not been corrected immediately.

ventions for such cases, in order to keep the CzEngVallex pairings consistent. For example, EngVallex (used for the valency annotation of the English side of the PCEDT) often includes certain adjuncts (i.e., optional free modifications in the PDT terminology) in the valency frame, while PDT-Vallex strictly does not. This is, of course, not a cause for a “true” argument mismatch, but the treatment for these had to be unified so that these cases are easily identifiable afterwards.

Similarly, certain types of verb constructions using more than one verb (typically, catenative verb or a modal) might have structurally different annotation, if only for the fact that one side of the translation only one verb is used carrying the same meaning. In these cases, the “semantic” annotation rule takes effect, i.e., the modal or catenative verb is left out and the alignment is made between the more semantically “full” verb and its single-word counterpart in the other language (node in the annotated tree). For example, *keep* and *riding (up)* are represented as two nodes in the English tree annotation, while their translation is only *klouzat* in Czech (albeit complemented by and adverbial *stále*, meaning *lit. still*); in such a case, *keep* is not considered part of the pair and alignment is made for *ride (up)* and *klouzat* and their arguments only. In addition to *keep*, *need* or *get* (when complemented by a non-finite verb) also appear often translated in the same way.

In some cases, the translation itself could be plain wrong (however unlikely it might seem after professional translation editing and fully manual tectogrammatical annotation took place on the data prior to this alignment effort). In these cases, the corpus pairing is excluded from consideration and the error reported to the treebank maintainers.

3.3. CzEngVallex format

The resulting CzEngVallex is represented as a simple stand-off file which refers back to the PDT-Vallex and Eng-Vallex lexicons, or more precisely, to the individual valency frames in them. In other words, the underlying two lexicons are not modified at all, which makes it easier to maintain them in the future (Fig. 5). The valency frames are referred to by their respective IDs, while the arguments are identified by their labels (since they are for each frame unique). Technically, all Czech frame pairs are listed for every English verb, but the relations are symmetric.

CzEngVallex is also publicly available online for quick browsing and search.⁹ This interface allows for searching for particular argument pairs aligned by CzEngVallex, resulting in a list of verbs (and their particular valency frames) where this pairing occurs. Individual verb and verb pairs can also be browsed alphabetically, in both directions (English->Czech as well as Czech->English). Moreover, each pair of valency frames displayed is complemented with all the real-usage examples from the parallel PCEDT corpus (Fučíková et al., 2015). All the displayed material (verb entry heading, valency frames, etc.) are linked through HTML links to the monolingual entries in PDT-Vallex and EngVallex, to display additional information and, in the case of PDT-Vallex, additional examples from the monolingual Czech PDT corpus.

⁹<http://lindat.mff.cuni.cz/services/CzEngVallex>

```
<frames_pairs owner="...">
<head>
...
</head>
<body>
<valency_word id=... vw_id="ev-w1">
  <en_frame id=... en_id="ev-w1f2">
    <frame_pair id=... cs_id="v-w3161f1">
      <slots>
        <slot en_funcutor="ACT" cs_funcutor="ACT"/>
        <slot en_funcutor="PAT" cs_funcutor="PAT"/>
      </slots>
    </frame_pair>
  </en_frame>
  <frame_pair id=... cs_id="v-w9887f1">
    <slots>
      <slot en_funcutor="ACT" cs_funcutor="ACT"/>
      <slot en_funcutor="PAT" cs_funcutor="PAT"/>
      <slot en_funcutor="EFF" cs_funcutor="SUBS"/>
    </slots>
  </frame_pair>
</valency_word>
</body>
</frames_pairs>
```

Figure 5: Structure of the CzEngVallex (part of *abandon* pairing)

Number of argument pairs	Number of frame pairs	Percent of all pairs
0	9	0.04%
1	593	2.85%
2	8746	41.98%
3	7939	38.10%
4	2613	12.54%
5	813	3.90%
6	103	0.49%
7	19	0.09%

Table 2: Argument pairing statistics

4. Argument matching in the CzEngVallex / PCEDT

Out of the 20,835 frame pairs recorded in the CzEngVallex lexicon, Table 2 summarizes argument alignment diversity in these frame pairs: it shows how many times a certain number of argument pairs appears in the CzEngVallex lexicon.

It should be noted that not necessarily the number of arguments on both sides is equal to the number of pairs; some pairs might in effect pair an argument with “nothing” on the other side. A study on such a “zero” alignment can be found in (Šindlerová et al., 2015).

One of the reasons for creating CzEngVallex was to have explicitly annotated corpus material for the study of translation differences in Czech and English valency, or verb argument (and in some cases, also adjunct) use. Overall statistics are given in Table 3.

An example of a well-behaved verb pair is in Fig. 6, where all three arguments match between the two languages for

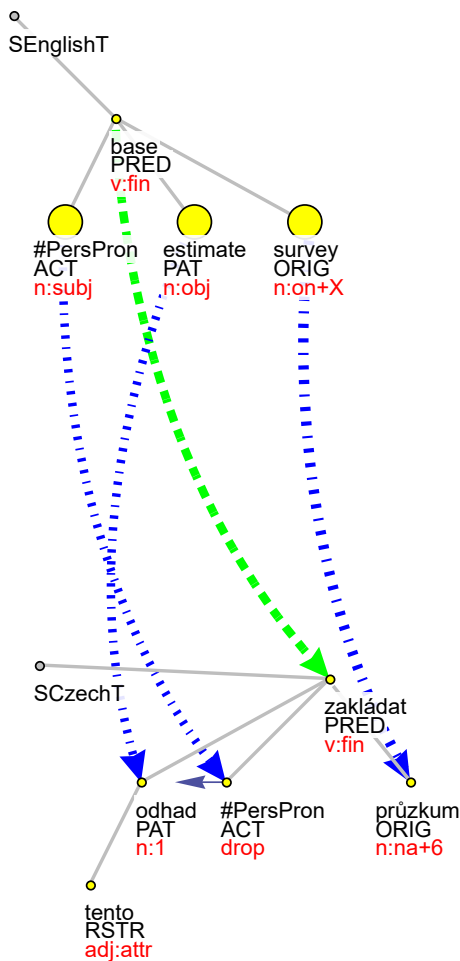
No. of argument pair differences	Number of frame pairs	Percent of all pairs
0	9302	44.646%
1	6737	32.335%
2	3313	15.901%
3	1157	5.553%
4	267	1.281%
5	49	0.235%
6	9	0.043%
7	1	0.005%

Table 3: Argument pair differences in numbers

Number of argument pairs	Number of unique pairing types	Percentage
0	1	0.04%
1	4	0.15%
2	238	9.20%
3	980	37.88%
4	935	36.14%
5	338	13.07%
6	76	2.94%
7	15	0.58%

Table 4: Argument pair differences in numbers

the verb sense pair *base*–*zakládat*.



En: He bases the estimate on a survey...
Cz: Tento odhad zakládá na průzkumu...

Figure 6: Matching arguments in verb pair *base*–*zakládat*; verb pair in green, argument links in blue.

However, quite clearly as the table shows, there are more differing pairs (over 55%) than those which match in all argument pairings.

An example of an aligned sentence with five differences in argument mapping is captured in Fig. 7.

The example with seven differences comes from the translation of the English verb “to sell” to Czech as “vyvážet”

(lit. *export*) in

- En: For example, Nissho Iwai Corp., one of the biggest Japanese trading houses, now buys almost twice as many goods from China as it.ACT sells to that country.ADDR
- Cz: Společnost Nissho Iwai Corp., jedna z největších japonských obchodních firem, dnes například kupuje dvakrát tolik zboží z Číny, než kolik.PAT do této země.DIR3 vyváží

In this case, the English entry has five argument slots, labeled ACT, PAT, ADDR, EFF, BEN and the Czech entry ACT, PAT and DIR1;¹⁰ ACT maps to PAT, ADDR to DIR3 (not included as an argument in the valency frame), and all others are unaligned (in either direction), accounting for the seven pairing differences.

Out of the frame pairs with just one argument pair, four different cases have been found. While it is not surprising that by far the most frequent pair is the expected ACT:ACT labeled argument pair, three other differing pairs have been found:¹¹

1. five frame pairs with PAT:ACT argument pair; this is apparently the relict of not shifting the English valency slot label PAT to ACT, due to its origins in Propbank which often uses Arg1 alone (such as in *the glass.Arg1 broke*, and Engvallex typically used PAT for Arg1;
2. four times no English frame argument corresponding to ACT in the Czech frame, and
3. one case of an ACT on the English side corresponding to no argument on the Czech side.

With the increasing number of arguments, there are more and more different pairings of arguments, as the combinatorics also suggest. The numbers are given in Table 4. The percentages are computed from the total number of 2,587 different (unique) pairs found in the CzEngVallex lexicon across all argument pair counts.

¹⁰Not all of them are present in the (surface form of the) example, but the alignment is not affected by argument ellipsis.

¹¹More examples and their breakout (including possible annotation errors) will be presented in the full version of the paper.

5. Mismatch classification

While complete breakout and classification of the 2,500+ mismatch types apparently needs further study, we can already provide (a coarse grained) classification. The “zero alignment” has already been mentioned and studied (Šindlerová et al., 2015), since it accounts for a large proportion of argument alignment discrepancies. However, when we step up from the investigation of individual argument alignments to the level of the whole valency frame, the situation is far richer. Nevertheless, there are certain common reasons for various types of mismatches:

- verb translation choice often combined with differing argument expression and/or representation, which can further be subdivided into several types (plain argument expression (*to drive a car*:PAT vs. *jezdít v autě*:MEANS, lit. *go in a car*), light verb constructions translated as a single verb or vice versa, such as *uzavřít smlouvu s ...*, lit. *close a contract with ...* → (*to*) *contract sb.*, “cross-language” alternation (cf. also Fig. 7 and below), other structural differences)
- treebank annotation convention and guidelines (e.g., choice of direction vs. origin), cf. Fig. 7: *is derived from the U.S.*:ORIG vs. *pochází z USA*:DIR1
- valency frame composition convention mismatch (for example: En: *spread* ACT DIR1 DIR2 DIR3 → Cz: *rozšířit se* ACT, where the direction(s) of spreading are not included in the Czech valency frame, being considered optional complementations).

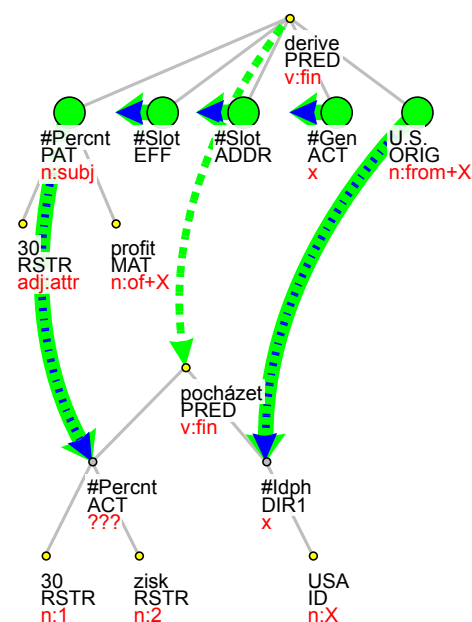
As an illustration,¹² consider the following translation:

- En: ... the change ended [the series]
- Cz: ... série skončila změnou (lit. [*the*] *series ended by-change*)

where the (deep) object (*series*) has moved to (deep) subject position in Czech (this alternation process applies to both languages; it was the translator’s choice to do so in Czech). In a slightly more complex example, we refer to another case of “cross-language” alternation (Fig. 7): the passive form “is derived” has been translated as intransitive “pocházet” (more literally translated as *come from*), where the deep subject (ACT) represents the theme, while in English this is the deep object (PAT) of “derive”: someone derives something.PAT from ... This example suggests that in translation, the choice of the translation is often not done at the more syntactically-oriented valency (or “propbanking”) level, but at a much deeper, FrameNet-like more semantically-oriented level (Baker et al., 1998); while this might not be surprising for human translators, it confirms that it has to be taken into account for MT. Interlinking all the valency/propbanking/semantic role lexicons, similarly to (Bonial et al., 2013), would give us more insight, but it must be complemented with multilingual annotation in a similar way that we have attempted here with CzEngVallex in the bilingual case.

¹²Due to the limited space in the abstract - more examples and finer grained classes in the full version of the paper.

For completeness, we should also mention our previous work on investigating how verb-noun phrasal and verb idiomatic constructions are translated (Urešová et al., 2013). We have found that only a minority of such constructions are translated as idiomatic or phrasal constructions (from English to Czech), and perhaps even more surprisingly, it also holds in the other direction, namely that idioms (in the Czech translation) are often coming from non-idiomatic constructions. The findings about translations of verb-noun idiomatic constructions has led to more focus on the representation of such constructions themselves in valency dictionaries in different languages; comparison between Czech and Polish with suggestions for improvement in representation of verb-based idiomatic constructions has been described in (Przepiórkowski et al., 2016 in print).



En: ... 30% of ... profit ... is derived from the U.S..
Cz: .. 30 % zisků ... pochází z USA.

Figure 7: Functor mismatch in 5 argument pairs

6. Related work

The predecessor to CzEngVallex, which has used machine learning methods based on a parallel corpus, has been described in (Šindlerová and Bojar, 2009), but it did not produce a manually checked and corrected resource. Another preliminary attempt at a comparison of English and Czech Valency has been using several resources (PDEV on the English side and VerbaLex on the Czech side), but it has not used a parallel corpus for linking and checking the actual usage (Pala et al., 2014). Obviously, multilingual dictionaries like FrameNet (Fillmore et al., 2003; Baker et al., 1998; Materna and Pala, 2010) inherently contain links between verb sense equivalents, but we are not aware of any work that would start from a parallel corpus, use the same methodology of valency description for both languages and that has underwent a thorough manual check.

7. Conclusions

We have described some basic statistics derived from the CzEngVallex lexicon, a bilingual valency lexicon created over the Prague Czech-English Dependency Treebank, a parallel corpus of over 50,000 sentences. Perhaps it should not be surprising that there is a large number of differences in the use of verb arguments across the two languages. The 2,587 different valency frame pairs (in the alignment of their arguments) offer a large amount of material for further studies.

Apart from studying the properties of the lexical entries themselves, we have already used the lexicon in various NLP applications, such as in word sense disambiguation using the argument and verb pairings coming from the parallel corpus as an additional features, getting an improvement over the (monolingual) baseline (Dušek et al., 2015). Since the CzEngVallex lexicon, both underlying valency lexicons (PDT-Vallex for Czech and EngVallex for English) are now publicly available online,¹³ we believe that it will be possible to get more insight into the use of verb arguments in translation, benefiting both linguistic studies as well as language technology, especially machine translation.

8. Acknowledgments

This work described herein has been supported by the grant GP13-03351P of the Grant Agency of the Czech Republic and by the LINDAT/CLARIN Research Infrastructure projects, LM2010013 and LM2015071 funded by the MEYS of the Czech Republic. It has also been using language resources developed and distributed by the LINDAT/CLARIN project. The results described in this article have been contributed to the LINDAT/CLARIN open access repository, as a item <http://hdl.handle.net/11234/1-1512> and as an user-accessible application at <http://lindat.mff.cuni.cz/services/CzEngVallex>.

9. References

- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bonial, C., Stowe, K., and Palmer, M., (2013). *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, chapter Renewing and Revising SemLink, pages 9 – 17. Association for Computational Linguistics.
- Cinková, S. (2006). From Propbank to Engvallex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy.
- Dušek, O., Fučíková, E., Hajič, J., Popel, M., Šindlerová, J., and Urešová, Z. (2015). Using parallel texts and lexicons for verbal word sense disambiguation. In *Proceedings of the Third International Conference on Dependency Linguistics, Depling 2015*, Uppsala, Sweden. Uppsala universitet, Uppsala universitet.
- Fillmore, C. J., Johnson, C. R., and L.Petruck, M. R. (2003). Background to framenet: Framenet and frame semantics. *International Journal of Lexicography*, 16(3):235–250.
- Fučíková, E., Hajič, J., Šindlerová, J., and Urešová, Z. (2015). Czech-english bilingual valency lexicon online. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 61–71, Warszawa, Poland. IPIPAN, IPIPAN.
- Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Nivre, Joakim//Hinrichs, E., editor, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.
- Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M., and Urešová, Z. (2006). Prague Dependency Treebank 2.0, LDC Catalog No. LDC2006T01.
- Hajic, J., Pajas, P., Marecek, D., Mikulova, M., Uresova, Z., and Podvesky, P. (2009). Prague dependency treebank of spoken language (PDTSL) 0.5. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.
- Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association.
- Horák, Aleš and Pala, Karel and Hlaváčková, Dana. (2013). Preparing VerbaLex Printed Edition. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013*, pages 3–11.
- Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.
- Materna, J. and Pala, K. (2010). Using Ontologies for Semi-automatic Linking VerbaLex with FrameNet. In *LREC*, pages 3331–3337.
- Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z., and Kučová, L. (2005). Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical

¹³<http://hdl.handle.net/11234/1-1512> and <http://lindat.mff.cuni.cz/services/CzEngVallex>

- Report TR-2005-28, ÚFAL MFF UK, Prague, Prague.
- Pala, K., Baisa, V., Sitová, Z., and Vonšovský, J. (2014). Mapping czech and english valency lexicons: Preliminary report. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 139–145, Brno. Tribun EU.
- Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.
- Przepiórkowski, A., Hajič, J., Hajnicz, E., and Urešová, Z. (2016, in print). Phraseology in two slavic valency dictionaries: limitations and perspectives. *International Journal of Lexicography*.
- Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., and Świdziński, M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2785–2792, Reykjavík, Iceland. ELRA.
- Šindlerová, J. and Bojar, O. (2009). Towards English-Czech Parallel Valency Lexicon via Treebank Examples. In *Eighth International Workshop on Treebanks and Linguistic Theories*, pages 185–195.
- Šindlerová, J., Fučíková, E., and Urešová, Z. (2015). Zero alignment of verb arguments in a parallel treebank. In Hajičová, E. and Nivre, J., editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 330–339, Uppsala, Sweden. Uppsala University, Uppsala University.
- Urešová, Z., Fučíková, E., Hajič, J., and Šindlerová, J. (2013). An Analysis of Annotation of Verb-Noun Idiomatic Combinations in a Parallel Dependency Corpus. Proceedings from The 9th Workshop on Multiword Expressions, Workshop at NAACL 2013.
- Urešová, Z., Dušek, O., Fučíková, E., Hajič, J., and Šindlerová, J. (2015a). Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus. In *Proceedings of the The 9th Linguistic Annotation Workshop (LAW IX 2015)*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Urešová, Z., Fučíková, E., and Šindlerová, J. (2015b). Czengvallex: Mapping valency between languages. Technical Report TR-2015-58.
- Urešová, Z. (2011a). *Valence sloves v Pražském závislostním korpusu*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.
- Urešová, Z. (2011b). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.
- Žabokrtský, Z. and Lopatková, M. (2007). Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, (87):41–60.

DEFEXT: A Semi Supervised Definition Extraction Tool

Luis Espinosa-Anke, Roberto Carlini, Horacio Saggion, Francesco Ronzano

TALN Group, Universitat Pompeu Fabra
Carrer Tànger, 122-134, Barcelona (Spain)
{luis.espinosa,firstname.lastname}@upf.edu

Abstract

We present DEFEXT, an easy to use semi supervised Definition Extraction Tool. DEFEXT is designed to extract from a target corpus those textual fragments where a term is explicitly mentioned together with its core features, i.e. its definition. It works on the back of a Conditional Random Fields based sequential labeling algorithm and a bootstrapping approach. Bootstrapping enables the model to gradually become more aware of the idiosyncrasies of the target corpus. In this paper we describe the main components of the toolkit as well as experimental results stemming from both automatic and manual evaluation. We release DEFEXT as open source along with the necessary files to run it in any Unix machine. We also provide access to training and test data for immediate use.

Keywords: lexicography, definition extraction, bootstrapping

1. Introduction

Definitions are the source of knowledge to consult when the meaning of a term is sought, but manually constructing and updating glossaries is a costly task which requires the cooperative effort of domain experts (Navigli and Velardi, 2010). Exploiting lexicographic information in the form of definitions has proven useful not only for Glossary Building (Muresan and Klavans, 2002; Park et al., 2002) or Question Answering (Cui et al., 2005; Saggion and Gaizauskas, 2004), but also more recently in tasks like Hypernym Extraction (Espinosa-Anke et al., 2015b), Taxonomy Learning (Velardi et al., 2013; Espinosa-Anke et al., 2016) and Knowledge Base Generation (Delli Bovi et al., 2015).

Definition Extraction (DE), i.e. the task to automatically extract definitions from naturally occurring text, can be approached by exploiting lexico-syntactic patterns (Reberolle and Tanguy, 2000; Sarmiento et al., 2006; Storrer and Wellinghoff, 2006), in a supervised machine learning setting (Navigli et al., 2010; Jin et al., 2013; Espinosa-Anke and Saggion, 2014; Espinosa-Anke et al., 2015a), or leveraging bootstrapping algorithms (Reiplinger et al., 2012; De Benedictis et al., 2013).

In this paper, we extend our most recent contribution to DE by releasing DEFEXT¹, a toolkit based on experiments described in (Espinosa-Anke et al., 2015c), consisting in machine learning sentence-level DE along with a bootstrapping approach. First, we provide a summary of the foundational components of DEFEXT (Section 2.). Next, we summarize the contribution from which it stems (Espinosa-Anke et al., 2015c) as well as its main conclusion, namely that our approach effectively generates a model that gradually adapts to a target domain (Section 3.1.). Furthermore, we introduce one additional evaluation where, after bootstrapping a subset of the ACL Anthology, we present human experts in NLP with definitions and distractors (Section 3.2.), and ask them to judge whether the sentence includes *definitional knowledge*. Finally, we provide a brief description of the released toolkit along with accompanying enriched corpora to enable immediate use (Section 3.3.).

2. Data Modelling

DEFEXT is a weakly supervised DE system based on Conditional Random Fields (CRF) which, starting from a set of manually validated definitions and distractors, trains a seed model and iteratively enriches it with high confidence instances (i.e. highly likely definition sentences, and highly likely not definition sentences, such as a text fragments expressing a personal opinion). What differentiates DEFEXT from any supervised system is that thanks to its iterative architecture (see Figure 1), it gradually identifies more appropriate definitions with larger coverage on non-standard text than a system trained only on WordNet glosses or Wikipedia definitions (experimental results supporting this claim are briefly discussed in Section 3.).

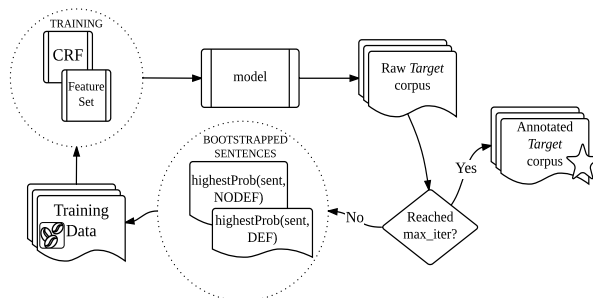


Figure 1: Workflow of DEFEXT.

2.1. Corpora

We release DEFEXT along with two automatically annotated datasets in column format, where each line represents a word and each column a feature value, the last column being reserved for the token’s label, which in our setting is DEF/NODEF. Sentence splitting is encoded as a double line break.

These corpora are enriched versions of (1) The Word Class Lattices corpus (Navigli et al., 2010) (WCL_d); and (2) A subset of the ACL Anthology Reference Corpus (Bird et al., 2008) (ACL-ARC_d). Statistics about their size in tokens and sentences, as well as definitional distribution (in the case of WCL_d) are provided in Table 1.

¹<https://bitbucket.org/luisespinosa/defext>

	WCL _d	ACL-ARC _d
Sentences	3,707	241,383
Tokens	103,037	6,709,314
Definitions	2,059	NA
Distractors	1,644	NA

Table 1: Statistics of the two enriched corpora that accompany the DEFEXT code.

2.2. Feature Extraction

The task of DE is modelled as a sentence classification problem, i.e. a sentence may or may not contain a definition. The opting for CRF as the machine learning algorithm is twofold: First, its sequential nature allows us to encode fine-grained features at word level, also considering the context of each word. And second, it has proven useful in previous work in the task of DE (Jin et al., 2013).

The features used for modelling the data are based on both linguistic, lexicographic and statistical information, such as:

- **Linguistic Features:** Surface and lemmatized words, part-of-speech, chunking information (NPs) and syntactic dependencies.
- **Lexicographic Information:** A feature that looks at noun phrases and whether they appear at potential *definiendum* (D) or *definiens* (d) position², as illustrated in the following example:
The⟨o-D⟩ Abwehr⟨b-D⟩ was⟨o-d⟩ a⟨o-d⟩ German⟨b-d⟩ intelligence⟨i-d⟩ organization⟨i-d⟩ from⟨o-d⟩ 1921⟨o-d⟩ to⟨o-d⟩ 1944⟨o-d⟩.
- **Statistical Features:** These are features designed to capture the degree of *termhood* of a word, its frequency in generic or domain-specific corpora, or evidence of their salience in definitional knowledge. These are:

- **termhood:** This metric determines the importance of a candidate token to be a terminological unit by looking at its frequency in general and domain-specific corpora (Kit and Liu, 2008). It is obtained as follows:

$$\text{Termhood}(w) = \frac{r_D(w)}{|V_D|} - \frac{r_B(w)}{|V_B|}$$

Where r_D is the frequency-wise ranking of word w in a domain corpus (in our case, WCL_d), and r_B is the frequency-wise ranking of such word in a general corpus, namely the Brown corpus (Francis and Kucera, 1979). Denominators refer to the token-level size of each corpus. If word w only appears in the general corpus, we set the value of Termhood(w) to $-\infty$, and to ∞ in the opposite case.

²The *genus et differentia* model of a definition, which traces back to Aristotelian times, distinguishes between the *definiendum*, the term that is being defined, and the *definiens*, i.e. the cluster of words that describe the core characteristics of the term.

- **tf-gen:** Frequency of the current word in the general-domain corpus r_B (Brown Corpus).
- **tf-dom:** Frequency of the current word in the domain-specific corpus r_D (WCL_d).
- **tfidf:** Tf-idf of the current word over the training set, where each sentence is considered a separate document.
- **def_prom:** The notion of Definitional Prominence describes the probability of a word w to appear in a definitional sentence ($s = def$). For this, we consider its frequency in definitions and non-definitions in the WCL_d as follows:

$$\text{DefProm}(w) = \frac{DF}{|Defs|} - \frac{NF}{|Nodefs|}$$

where $DF = \sum_{i=0}^{i=n} (s_i = def \wedge w \in s_i)$ and $NF = \sum_{i=0}^{i=n} (s_i = nodef \wedge w \in s_i)$. Similarly as with the *termhood* feature, in cases where a word w is only found in definitional sentences, we set the DefProm(w) value to ∞ , and to $-\infty$ if it was only seen in non-definitional sentences.

- **D_prom:** Definiendum Prominence, on the other hand, models our intuition that a word appearing more often in position of potential *definiendum* might reveal its role as a definitional keyword. This feature is computed as follows:

$$DP(w) = \frac{\sum_{i=0}^{i=n} w_i \in \text{term}_D}{|DT|}$$

where term_D is a noun phrase (i.e. a term candidate) appearing in potential definiendum position and $|DT|$ refers to the size of the candidate term corpus in candidate definienda position.

- **d_prom:** Similarly computed as D_prom, but considering position of potential definiens.

These features are used to train a CRF algorithm. DEFEXT operates on the back of the CRF toolkit CRF++³, which allows selecting features to be considered at each iteration, as well as the context window.

2.3. Bootstrapping

We implemented on DEFEXT a bootstrapping approach inspired by the well-known Yarowsky algorithm for Word Sense Disambiguation (Yarowsky, 1995). It works as follows: Assuming a small set of seed labeled examples (in our case, WCL_d), a large target dataset of cases to be classified (ACL-ARC_d), and a learning algorithm (CRF), the initial training is performed on the initial seeds in order to classify the whole target data. Those instances classified with high confidence are appended to the training data until convergence or a number of maximum iterations is reached.

We apply this methodology to the definition bootstrapping process, and for each iteration, extract the highest confidence definition and the highest confidence non-definition from the target corpus, retrain, and classify again. The

³<https://taku910.github.io/crfpp/>

	Iteration	P	R	F
MSR-NLP	20	78.2	76.7	77.44
W00	198	62.47	82.01	71.85

Table 3: Iteration and best results for the two held-out test datasets on the DE experiment.

number of maximum iterations may be introduced as an input parameter by the user. Only the latest versions of each corpus are kept in disk.

3. Experiments and Evaluation

As the bootstrapping process advances, the trained models gradually become more aware of the linguistic particularities of the genre and register of a target corpus. This allows capturing definition fragments with a particular syntactic structure which may not exist in the original seeds. In this section, we summarize the main conclusions drawn from the experiments performed over two held out test datasets, namely the W00 corpus (Jin et al., 2013), and the MSR-NLP corpus (Espinosa-Anke et al., 2015c)⁴. The former is a manually annotated subset of the ACL anthology, which shows high domain-specificity as well as considerable variability in terms of how a term is introduced and defined (e.g. by means of a comparison or by placing the defined term at the end of the sentence). The latter is compiled manually from a set of abstracts available at the Microsoft Research website⁵, where the first sentence of each abstract is tagged in the website as a definition. This corpus shows less linguistic variability and thus its definitions are in the vast majority of cases, highly canonical. We show one sample definition from each corpus in Table 2⁶.

3.1. Definition Extraction

Starting with the WCL_d corpus as seed data, and the ACL-ARC_d collection for bootstrapping, we performed 200 iterations and, at every iteration, we computed Precision, Recall and F-Score at sentence level for both the W00 and the MSR-NLP corpora. At iteration 100, we recalculated the statistical features over the bootstrapped training data (which included 200 more sentences, 100 of each label). The trend for both corpora, shown in Figure 2, indicates that the model improves at classifying definitional knowledge in corpora with greater variability, as performance on the W00 corpus suggests. Moreover, it shows decreasing performance on standard language. The best iterations with their corresponding scores for both datasets are shown in Table 3.

3.2. Human Evaluation

In this additional experiment, we assessed the quality of extracted definitions during the bootstrapping process. To this end, we performed 100 iterations over the ACL-ARC_d

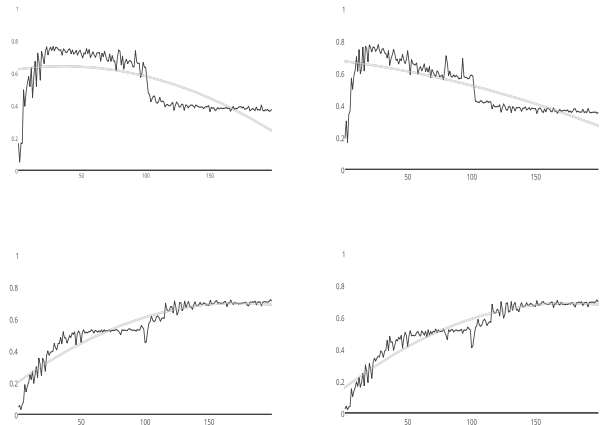


Figure 2: F-Score against iteration on the MSR-NLP (top row) and W00 (bottom row) datasets under two different confidence thresholds.

corpus and presented experts in NLP with 200 sentences (100 candidate definitions and 100 distractors), with shuffled order. Note that since the ACL-ARC corpus comes from parsing pdf papers, there is a considerable amount of noise derived from diverse formatting, presence of equations or tables, and so on. All sentences, however, were presented in their original form, noise included, as in many cases we found that even noise could give the reader an idea of the context in which the sentence was uttered (e.g. if it is followed by a formula, or if it points to a figure or table). The experiment was completed by two judges who had extensive familiarity with the NLP domain and its terminology. Evaluators were allowed to leave the answer field blank if the sentence was unreadable due to noise.

In this experiment, DEFEXT reached an average (over the scores provided by both judges) of 0.50 Precision when computed over the whole dataset, and 0.65 if we only consider sentences which were not considered noise by the evaluators. Evaluators found an average of 23 sentences that they considered unreadable.

3.3. Technical Details

As mentioned earlier, DEFEXT is a bootstrapping wrapper around the CRF toolkit CRF++, which requires input data to be preprocessed in column-based format. Specifically, each sentence is encoded as a matrix, where each word is a row and each feature is represented as a column, each of them tab-separated. Usually, the word’s surface form or lemma will be at the first or second column, and then other features such as part-of-speech, syntactic dependency or corpus-based features follow.

The last column in the dataset is the sentence label, which in DEFEXT is either DEF or NODEF, as it is designed as a sentence classification system.

Once training and target data are preprocessed accordingly, one may simply invoke DEFEXT in any Unix machine. Further implementation details and command line arguments can be found in the toolkit’s documentation, as well as in comments throughout the code.

⁴Available at http://taln.upf.edu/MSR-NLP_RANLP2015

⁵<http://academic.research.microsoft.com>

⁶Note that, for clarity, we have removed from the examples any metainformation present in the original datasets.

CORPUS	DEFINITION
WCL _d	The Abwehr was a German intelligence organization from 1921 to 1944 .
W00	Discourse refers to any form of language-based communication involving multiple sentences or utterances.
ACL-ARC _d	In computational linguistics, word sense disambiguation (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word.
MSR-NLP	User interface is a way by which a user communicates with a computer through a particular software application.

Table 2: Example definitions of all corpora involved in the development and evaluation of DEFEXT.

4. Discussion and Conclusion

We have presented DEFEXT, a system for weakly supervised DE at sentence level. We have summarized the most outstanding features of the algorithm by referring to experiments which took the NLP domain as a use case (Espinosa-Anke et al., 2015c), and complemented them with one additional human evaluation. We have also covered the main requirements for it to function properly, such as data format and command line arguments. No external Python libraries are required, and the only prerequisite is to have CRF++ installed. We hope the research community in lexicography, computational lexicography or corpus linguistics find this tool useful for automating term and definition extraction, for example, as a support for glossary generation or hypernymic (is-a) relation extraction.

5. Acknowledgements

This work is partially funded by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and Dr. Inventor (FP7-ICT-2013.8.1611383).

6. References

- Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. (2008). The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L08-1005.
- Cui, H., Kan, M.-Y., and Chua, T.-S. (2005). Generic soft pattern models for definitional question answering. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 384–391. ACM.
- De Benedictis, F., Faralli, S., Navigli, R., et al. (2013). Glossboot: Bootstrapping multilingual domain glossaries from the web. In *ACL (1)*, pages 528–538.
- Delli Bovi, C., Telesca, L., and Navigli, R. (2015). Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3:529–543.
- Espinosa-Anke, L. and Saggion, H. (2014). Applying dependency relations to definition extraction. In *Natural Language Processing and Information Systems*, pages 63–74. Springer.
- Espinosa-Anke, L., Saggion, H., and Delli Bovi, C. (2015a). Definition extraction using sense-based embeddings. In *International Workshop on Embeddings and Semantics, SEPLN*.
- Espinosa-Anke, L., Saggion, H., and Ronzano, F. (2015b). Hypernym extraction: Combining machine learning and dependency grammar. In *CICLING 2015*, page To appear, Cairo, Egypt. Springer-Verlag.
- Espinosa-Anke, L., Saggion, H., and Ronzano, F. (2015c). Weakly supervised definition extraction. In *Proceedings of RANLP 2015*.
- Espinosa-Anke, L., Saggion, H., Ronzano, F., and Navigli, R. (2016). Extasem! extending, taxonomizing and semantifying domain terminologies. In *AAAI*.
- Francis, W. N. and Kucera, H. (1979). Brown corpus manual. *Brown University*.
- Jin, Y., Kan, M.-Y., Ng, J.-P., and He, X. (2013). Mining scientific terms and their definitions: A study of the ACL anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 780–790, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Kit, C. and Liu, X. (2008). Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2):204–229.
- Muresan, A. and Klavans, J. (2002). A method for automatically building and evaluating dictionary resources. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.
- Navigli, R. and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *ACL*, pages 1318–1327.
- Navigli, R., Velardi, P., and Ruiz-Martínez, J. M. (2010). An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of LREC’10*, Valletta, Malta, may.
- Park, Y., Byrd, R. J., and Boguraev, B. K. (2002). Automatic Glossary Extraction: Beyond Terminology Identi-

- fication. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.
- Rebeyrolle, J. and Tanguy, L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, 25:153–174.
- Reiplinger, M., Schäfer, U., and Wolska, M. (2012). Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 55–65, Jeju Island, Korea, July. Association for Computational Linguistics.
- Saggion, H. and Gaizauskas, R. (2004). Mining on-line sources for definition knowledge. In *17th FLAIRS*, Miami Beach, Florida.
- Sarmiento, L., Maia, B., Santos, D., Pinto, A., and Cabral, L. (2006). Corpógrafo V3 From Terminological Aid to Semi-automatic Knowledge Engineering. In *5th International Conference on Language Resources and Evaluation (LREC'06)*, Geneva.
- Storrer, A. and Wellinghoff, S. (2006). Automated detection and annotation of term definitions in German text corpora. In *Conference on Language Resources and Evaluation (LREC)*.
- Velardi, P., Faralli, S., and Navigli, R. (2013). Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.
- Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics.

Linking and Disambiguating Swadesh Lists: Expanding the Open Multilingual Wordnet Using Open Language Resources

Lúis Morgado da Costa, Francis Bond, František Kratochvíl

Linguistics and Multilingual Studies,

Nanyang Technological University, Singapore

luis.passos.morgado@gmail.com, bond@ieee.com, fkratochvil@ntu.edu.sg

Abstract

In this paper we describe two main contributions in the fields of lexicography and Linked Open Data: a human corrected disambiguation, using the Princeton Wordnet's sense inventory (PWN, Fellbaum, 1998), of Swadesh lists maintained in the Internet Archive by the Rosetta Project, and the distribution of this data through an expansion of the Open Multilingual Wordnet (OMW, Bond and Foster, 2013). The task of disambiguating word lists isn't always a straightforward task. The PWN is a vast resource with many fine-grained senses, and word lists often fail to help resolve the inherent ambiguity of words. In this work we describe the corner cases of this disambiguation and, when necessary, motivate our choice over other possible senses. We take the results of this work as a great example of the benefits of sharing linguistic data under open licenses, and will continue linking other openly available data. All the data will be released in future OMW releases, and we will encourage the community to contribute in correcting and adding to the data made available.

Keywords: Swadesh lists, wordnet, lexicography, linked open data

1. Introduction

This work describes how we disambiguated and linked a large collection of over 1,200 Swadesh lists maintained by the Rosetta Project.¹ This was done as part of the Open Multilingual Wordnet (OMW, Bond and Foster, 2013), a linked collection of wordnets of multiple languages released under an open license, which includes the Princeton Wordnet's sense inventory (PWN, Fellbaum, 1998) extended with pronouns, determiners, interjections and classifiers (see Seah and Bond, 2014; Morgado da Costa and Bond, 2016).

We start by discussing the origin of Swadesh List and its multiple versions, as well as the intrinsic problems of disambiguating word lists. We discuss corner-cases of our disambiguation, and highlight the importance of working with disambiguated lists in research involving elicitation.

We introduce a new interface to OMW, designed to browse the OMW using lists and allowing users to use the data we collected in new and interesting ways. We also introduce the possibility of creating custom multilingual lists, and enjoy the benefits that come from using linked open linguistic data (i.e. senses and definitions in multiple languages).

With the exception of six prepositions and conjunctions, every word in the widely used Swadesh 207 list was mapped to a concept in the PWN, along with 72 other concepts that were spread across the many variant lists shared by the Rosetta Project. We started from an initial mapping provided by Huang et al. (2007), which was corrected and enhanced where necessary. Ultimately, this work produced a new extended version of the OMW, linking more than 270,000 new unique senses and raising the coverage of this resource from 150 to over 1,200 languages.

We commit to release the linked disambiguated Swadesh sense inventory under an open license, as part of OMW. This allows online search, downloads in a fixed format, and manipulation through the python Natural Language Toolkit

(NLTK: Bird et al., 2009).² This inventory can continue to be used in the same way as it once was, but brings the benefits of being defined in multiple languages, and being linked to hundreds of other languages through OMW.

2. Swadesh Lists

The Swadesh List is a classic compilation of words that change at a relatively constant rate, used in comparative linguistics studies to predict language relatedness and history by tracing the retention and relative change of vocabulary among languages (Swadesh, 1952). For this reason, words in Swadesh lists are supposed to be universal, but not necessarily the most frequent. The list has seen several versions/revisions through the years of lexico-statistics work of Swadesh (1952, 1955). Each of these versions became a list in its own right, with sizes ranging from 100 to 215 words. Through the continuous revision of these words-lists, Swadesh hoped to pinpoint a list of fundamental everyday vocabulary, present in every language, as opposed to a specialized or "cultural" vocabulary. And even though Swadesh's initial intentions were to enlarge this universal vocabulary, he acknowledged that the compilation of such list should avoid problems such as potential duplication, identical roots, sound imitation and semantic shading (ambiguity). In the end, the multiple revisions of his list kept getting shorter, until reaching 100 words.

The history of the Swadesh lists can be summed as follows: in 1952 Swadesh proposes his 200 word list (see Annex A), a selected extract from a 215 word list used in his earlier work. In this version, he tries to specify the intended meaning of these words with the use of parenthetical notes. In 1955, Swadesh publishes the original 215 word list which was used to create the 200 word list, grouping them in 23 semantic groups (see Annex B). In the same publication, Swadesh proposes his final list reduction, this time to contain only 100 words: 92 words selected from the original

¹<http://rosettaproject.org>

²<http://www.nltk.org/>

Min. No. Words	No. Languages	%
1	1211	1.000
50	1088	0.898
100	885	0.731
150	727	0.600
200	553	0.457
300	334	0.276
400	155	0.128
600	56	0.046
800	21	0.017
1000	9	0.007
2000	1	0.001

Table 1: Number of words per number of languages

215 list, and 8 new words (see Annex C). Finally, the widely used (non-official) 207 word Swadesh list (see Annex D) contains the 200 terms proposed in 1952, with the addition of 7 of the 8 new terms proposed (all except *claw*) in his final 100 word list (Huang et al., 2007).

Swadesh lists have been used in the fields of lexical-statistics and historical-comparative linguistics, from their conception until recent times. In early days, the popularity of Swadesh’s work propelled his lists into a de facto standard data-set to be collected in language description. Consequently, the large amount of collected data was eventually compiled into comparative vocabulary databases, for a number of language families. We can find examples of this in the Indo-European Lexical Cognacy Database³ and the Austronesian Basic Vocabulary Database (Greenhill et al., 2008). These came to be the standard data-sets for computational phylogenetic language change models (see, for example, Bouckaert et al., 2012).

Until today, multiple studies continue to use and produce Swadesh-like lists as seed data to study cognates and relatedness between languages, as well as to trace language history and evolution (see Serva and Petroni, 2008; Wu et al., 2015; Pagel et al., 2013). Holman et al. (2008) introduce a fundamental study, where an automatic model was used to calculate the relative stability of the items in the 100 word Swadesh list published in 1955. And they show that the 40 most stable items on the final Swadesh list are as effective in language classification models as the full 100 word list.

3. Data

All the data collected for the work presented in this paper is readily available in the Internet Archive,⁴ and was commissioned and owned by the Rosetta Project. This project is run by the Long Now Foundation, and it is a global effort, open to language specialists and native speakers, to build a publicly accessible digital library of human languages. Among other language documentation initiatives, this project maintains and openly shares a large collection of Swadesh lists in multiple languages.

A total of 1,211 lists, for 1,211 different languages, were downloaded as simple text files, along with an xml file that includes their specific meta data. This meta data includes

common information, such as license, authorship, and also the language’s code and full name. All lists dealt with are shared under a CC-BY 3.0 (Unported) license.⁵

Here is an excerpt from the list for Abui, an Alor-Pantar language spoken in Eastern Indonesia:

```
...
push: habi
rain: anui
rain: ?anuy
rain: anúy
rain: anúy
rat: rui
red: arangnabake
red: kiika
red: kiika
red: ki:ka
red: kika
ripe: kang
ripe: ma
...
```

As can be seen above, the format of these lists includes an English word and its counterpart in the target language, separated by a colon. Multiple senses can exist for a single English word. And multiple spellings are also provided for some senses. The size of lists varied greatly. Table 1 gives an account of the distribution of list sizes (incl. duplicates). As can be seen in Table 1, all 1,211 lists contained at least one word pair, and 60% of these lists contained at least 150 word pairs. We can see that a relative large portion of these lists include a few hundred pairs, and that a few languages actually included over a thousand. Upon processing and analyzing these lists, we found that duplicates and orthographic variations were quite common, explaining why some lists have a very high number of words. This can be seen above (see, for example, the repetition of the pair *red: kiika*).

Even though the lists collected were named after Morris Swadesh, we have seen in Section 2. how this is a somewhat abstract concept. Swadesh lists often also refer to lists that include words that fall outside any of the original work of Swadesh. And this was often the case for the lists we collected. Many of the lists included English words that were not included in any of the original Swadesh lists.

A closer inspection of these extra words showed that they fell into three rough classes. Most were quite general, such as *today*, *son*, *house* and *frog*. There were also many body parts, such as *finger*, *arm*, *lip*, *chin*, *forehead*. Finally, an interesting set of words was clearly focused on Australian languages, which was made evident by a very specific lexical choice of animals such as *kangaroo*, *cassowary*, *wallaby*, and *emu*, which are only found in and around this region.

4. Disambiguation

The original design of PWN includes only contentful/referential open class words: nouns, verbs, adjectives and adverbs. In this work, however, due to the nature of the

³<http://ielex.mpi.nl/>

⁴<https://archive.org/>

⁵<https://creativecommons.org/licenses/by/3.0/>

task in question, we added two expansions of PWN that include a large set of pronouns, determiners, interjections and classifiers (see Morgado da Costa and Bond, 2016; Seah and Bond, 2014).

After pre-processing the data introduced in Section 3., removing duplicates, we decided to map every English word that had translations in at least 100 languages. While ensuring this, we found that some English words appeared, inconsistently, using multiple forms. For example, the word *fly* appeared also with the form *fly v.* and *fly (v.)*. In cases like this, all words linked to any of these forms were linked to the same concept, in this case 01940403-v – “travel through the air; be airborne”.

We started with an initial mapping of the Swadesh 207 word list provided by Huang et al. (2007). We carefully rechecked this initial mapping against the cues provided in the original publications. We tried, as much as possible, to base our choices on the parenthetical notes introduced in Swadesh (1952) and the semantic grouping shown in Swadesh (1955).

Based on these, we enhanced and made a few corrections to the initial mapping provided by Huang et al. (2007). Firstly, using the expansions to PWN’s concept inventory, we were now able to map 13 pronouns for which there were no previous mappings. From the remaining data, we made only 13 corrections. We provide three of these as examples:

1. the word *squeeze* had originally been mapped to 00357023-n – “the act of gripping and pressing firmly”; but since this word is presented as *to squeeze* (Swadesh, 1952), we chose instead the verbal concept 01387786-v – “squeeze or press together”;
2. the word *day* had originally been mapped to 15155220-n – “time for Earth to make a complete rotation on its axis”; but since there is a parenthetical note stating “opposite of night rather than the time measure” (Swadesh, 1952), we corrected it to 15164957-n – “the time after sunrise and before sunset while it is light outside”;
3. the word *louse* had originally been mapped to 02185481-n – “wingless insect with mouth parts adapted for biting, mostly parasitic on birds”; but since we thought this sense was too specific (i.e. synonym of *bird louse*), we changed the mapping to the more general concept 02183857-n – “wingless usually flattened bloodsucking insect parasitic on warm-blooded animals”;

After going through the 207 mappings provided by Huang et al. (2007), we continued to map the remaining words that fell outside this list, which we collapsed into 72 other concepts. For these extra words, since little or no information was provided, we resorted to list cohesiveness and sense frequency in our disambiguation.

Through this effort, more than 270 PWN concepts received senses in at least 100 languages. The end result is an extended OMW, with more than 270,000 new unique senses and coverage for over 1,200 languages. Table 2 shows

Min. No. Concepts	No. Languages	%
1	1211	1.000
20	1151	0.950
40	1107	0.914
60	1011	0.835
80	962	0.794
100	806	0.666
120	666	0.550
140	595	0.491
160	501	0.414
180	345	0.285
200	145	0.120
220	63	0.052
240	21	0.017
250	2	0.002

Table 2: Number of concepts per number of languages

the distribution of number of concepts per number of languages. In this table we can see that all 1,211 languages received mappings to at least one concept, and that over 66% of all languages received senses to more than 100 concepts. Only two languages received mappings for more than 250 concepts, these were Orokolo (oro) and Toaripi (tqo), both from Papua New Guinea, which received sense mappings for 251 concepts each.

4.1. New and Excluded Concepts

Unfortunately, even considering an extended concept inventory from the expansion efforts mentioned above (see Section 4.), it was still insufficient to provide a complete mapping for every word. Three classes of words deserve to be mentioned here: pronouns, prepositions and conjunctions. Pronouns were first introduced to wordnets by Seah and Bond (2014), where many pronouns were introduced and marked for a number of semantic features including, for example, number, gender and politeness.

Nevertheless, while going through the word list that extended the Swadesh lists, we found occurrences for six pronouns that had not yet been accounted for. Namely, genderless third person pronouns (listed as *he/she*), dual first person pronouns (listed as *we two*), along with their inclusive and exclusive counterparts (listed as *we two (incl.)* and *we two (excl.)*), dual second person pronouns (listed as *you two*), and dual third person pronouns (listed as *they two*).

Following the same method described in Seah and Bond (2014), we added the six missing concepts to the OMW hierarchy, and linked these missing pronouns.

Concerning prepositions and conjunctions, we find a similar situation – i.e. there are no prepositions or conjunctions in the PWN to be able to map these words. But, in this case, we know of no effort done to expand wordnet inventories in this way, and we therefore excluded these two classes of words from this work.

4.2. The Problem of Ambiguity

As it has been mentioned before, disambiguating word lists isn’t a straightforward task, especially if the word lists provide little or no information that can be used to disambiguate them. Adding to this difficulty, the PWN is a vast resource

synset	lemmas	definition
00608372-v	know	perceive as familiar
00608502-v	know	be able to distinguish, recognize as being different
00595935-v	know	know how to do or perform something
00608670-v	know	know the nature or character of
00592883-v	recognize, know , acknowledge, recognise, today, ...	accept (someone) to be what is claimed or accept his power and authority
00594337-v	know	be familiar or acquainted with a person or an object
00596644-v	know , experience, live	have firsthand knowledge of states, situations, emotions, or sensations
00595630-v	know	be aware of the truth of something; have a belief or faith in something; regard as true beyond any doubt
00594621-v	know , cognize, cognise	be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about
00596132-v	know	have fixed in the mind

Table 3: PWN’s partial sense inventory for verbal concepts matching the lemma *know*

with many fine-grained senses. In this section we would like to highlight the difficulty of this task by describing a few corner cases.

Firstly, concerning pronouns, we would like to point out that we are aware that many pronouns may not be linked correctly. The reason for this comes from the rich pronominal hierarchy that was created when adding pronouns to wordnet (Seah and Bond, 2014). This pronominal hierarchy makes use of semantic features to split pronouns in multiple concepts, depending on features like number and gender, but also politeness, formality and gender speech. We will further exemplify this problem with current situation of the first person singular pronoun in English and Japanese.

In English, the concept for the pronoun *I* is marked only for three features: *first_person*, *personal_pronoun*, *singular*. But in Japanese, the same pronoun is split in multiple concepts. We can find a concept for わたし *watashi* marked for *first_person*, *personal_pronoun*, *singular*, *formal*, *polite*; a second concept for われ *ware* marked for *first_person*, *personal_pronoun*, *singular*, *formal*; a third for おれ *ore* and ぼく *boku* marked for *first_person*, *personal_pronoun*, *singular*, *informal*, *men’s_speech*; another one for わたくし *watakushi* marked for *first_person*, *personal_pronoun*, *singular*, *formal*, *polite*, *honorific*, and a few more.

The decision to split concepts by the set of semantic features they are marked for dictates that the English pronoun *I* and the Japanese pronoun わたし *watashi*, for example, are not senses of the same concept. This is simply an example, and other features are also used to further specialize other kinds of pronominal concepts.

Even though explaining the hierarchy and meaning of all these features is well beyond the scope of this work, it is important to note that, because we lack information about these above mentioned features, it is currently hard to pinpoint the correct mapping for pronouns collected. In cases where these features are not available (see the discussion about dual and genderless pronouns above), we decided to map pronouns to their English counterparts. While this will most certainly generate some noise in the mapping, we thought it was preferable to provide a mapping and correct it later than to exclude them.

Similar in spirit, we also felt it was difficult to choose be-

tween senses where a very fine grained distinction has been made in PWN. We exemplify this with the mapping of the verb *know*. Table 3 shows a partial sense inventory of PWN for verbal concepts matching the lemma *know*. As can be seen, this is a good example of a too fine-grained distinction of senses. In this case, even after excluding a few less likely choices, we are still invited to make a distinction between the meaning nuances of “familiar or acquainted with”, “have firsthand knowledge of”, “be aware of the truth of”, and “be cognizant or aware of”.

In situations like this, information on sense frequency and being consistent with the previous mapping were favored in our final choice. In this case, the concept 00594621-v – “be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about” had been chosen by Huang et al. (2007), which also happened to be the most frequent sense, so we didn’t change it. Out of other hard to disambiguate words, we highlight also *fat*, *blow*, *see*, *think* and *throw*, but many others exist.

Ultimately, what we would like to highlight is the fact that using English words as list keys is insufficient and often problematic, because it does not remove the temptation to define meanings in terms of the conceptualizations that this source language can trigger. By using language-agnostic concept-keys, the source language interference is minimized by the multilingual structure of these resources. In other words, instead of using the lemma *know* as a list key, we suggest using the equivalent, but language-agnostic concept key 00594621-v (as shown in Table 3).

5. Sharing and Visualizing the Data

Beyond the above mentioned commitment to share the processed data in subsequent OMW releases which, in turn, will also be available for manipulation through NLTK, we are also expanding the current OMW interface to allow linguists and researchers from other fields, like psychology or social sciences, to use the data described in this paper, along with the rich data already contained in OMW.

Relevant for this work, we have produced a list browser (see Figure 1), where well known vocabulary lists will be made readily available for browse and download. Currently, we include the four Swadesh lists: commonly referred to

OMW Lists

Swadesh 207 Show Languages: eng cmn jpn ind

Showing all concepts for the list: Swadesh 207

PWN 3.0	Lemmas	Definitions
00007846-n	eng: someone, somebody, soul, person, individual, mortal jpn: 員, -者, -員, 者, ひと, もの, -人, 誰か, 方, 個人, 人, -方, 人間	eng: person, singular, assertive existential pronoun, pronoun, person, singular; quantifier: assertive existential, a human being
00014742-v	eng: slumber, kip, sleep, log Z's, catch some Z's jpn: 就眠+する, ねね, 眠る, 睡る, 就眠, 寐る, ねね+する, 寝る	eng: be asleep
00015388-n	eng: animate being, brute, beast, animal, fauna, creature jpn: 毛の荒物, 生類, 獣, アニマル, 四つ脚, 珍獣, 生体, 鳥獣, 四つ足, 4つ脚, 動物, 生物, 獣畜, 4つ足, 生き物	eng: a living organism characterized by voluntary movement
00024073-r	eng: not, n't, non jpn:	eng: negation of a word or group of words
00031820-v	eng: express joy, laugh, express mirth jpn: 笑む, 咲う, 一笑, 一笑+する, 笑う	eng: produce laughter
00076400-v	eng: retch, regorge, upchuck, disgorge, be sick, spue, puke, honk, cat, purge, cast, vomit, spew, sick, barf, regurgitate, throw up, vomit up, chuck jpn: 嘔げる, 吐出す, 吐瀉+する, 吐出, 戻す, 上げる, 嘔吐, 嘔吐く, 嘔吐+する, 吐き出す, 吐瀉, 吐出+する, 吐く	eng: eject the contents of the stomach through the mouth
00101956-v	eng: ptyalise, spew, spue, ptyalize, spit jpn: 唾する	eng: expel or eject (saliva or phlegm or sputum) from the mouth
00141632-v	eng: tie jpn: 結える, 結わえ付ける, 結ぶ, 結び合せる, むすび付ける, 結わえる, 結えつける, 結う, 結び付ける	eng: form a knot or bow in

Figure 1: Excerpt from the Swadesh 207 list as show in the OMW Lists interface

as “Swadesh 200”, “Swadesh 215”, the final reduced list known as “Swadesh 100”, and the ubiquitous unofficial “Swadesh 207”.

This new interface allows the user to select any number of languages and a predefined list, for which a table-like array of data is produced, allowing to compare data across languages. We hope that this interface may help field linguists and other types of works involving elicitation, since lists can be tailor-made with a specific language selection in mind. Using lists produced in this way will guarantee that the data is pre-disambiguated, and can later be merged back and compared against other linked data.

The array of data produced can not only provide lists of lemmas in multiple languages, but also definitions where available. And since an English definition is a requirement to be a part of the OMW, lemmas in any language can always be accompanied with a definition to help disambiguate the respective sense.

This interface also has an option to produce a custom list of concepts (i.e. a list that has not been predefined). And we hope to further enrich this interface with other known lists such as the Sign Language Swadesh List⁶ (Woodward, 1993), or the Holman et al. (2008) most stable 40 word list.

6. Conclusion and Future Work

Using open data provided by the Rosetta Stone project, we have been able to link over 270,000 new senses to the Open Multilingual Wordnet. As a consequence of this, we have also greatly expanded the language knowledge this resource, which previously had data for 150 languages, but now contains data for over 1200 languages.

⁶This list modifies the Swadesh list in order to study sign languages. In particular, the proportion of indexical signs (body parts and pronouns) was reduced, as they are more likely to be similar.

To accomplish this, we have carefully disambiguated and linked over 1200 lists of words based on the work of Morris Swadesh. This disambiguation redefines the English words previously being used as Swadesh keys to a language-agnostic concept key in the Open Multilingual Wordnet.

This work has obvious practical benefits for lexicographic elicitation, setting an example on how sense disambiguated lexicons can, by linking to a language-agnostic concept key, enrich the knowledge we have of world languages. We believe that, to be able to do comparative work in the field of lexical semantics, it is important to control elicitation through an agreed upon sense inventory, as provided here. This kind of linked data, can provide enough resolution to study semantic typology (i.e. word similarity, language families, word loaning, etc.).

We have also shown that wordnets can be expanded through the use of open data. And following this trend, we want to continue linking known lexicographic lists and resources, especially when these can be sense disambiguated. Unfortunately, data in enough quantity is necessary to justify this time-consuming work.

Our next target will be to link the World Loanword Database (WoLD) (Haspelmeth and Tadmor, 2009), which provides linked mini-dictionaries (1000-2000 words) for 41 languages, with comprehensive information about the loanword status of each word. This is a well organized project, with well curated data, but disambiguating a much larger list will also have higher costs associated. To link a project of this size, since WoLD also provides definitions, we will most likely look into methods of automatic word sense disambiguation.

Nevertheless, even though the data released by the Rosetta Project is much simpler, and arguably even ill formatted (i.e. spurious repetition, spelling inconsistencies in the English

keys, etc.), we have shown with this work that only a certain amount of coherence and consistency are necessary to make data useful. The Rosetta Project is an excellent example of the benefits that come from crowd-sourced open data, which can be achieved with minimal supervision.

Concerning future work, and following the discussion introduced in Section 4.2., it would be important to do some error analysis on the mapped senses. We hope to do this in two ways. Firstly, we would like to use the multilingual structure of the OMW to automatically check the overlap between existing and newly mapped senses in languages for which we already have data. This will give us a rough estimate of the quality of the mapping, as we expect to have most of the Swadesh senses in human curated projects. We will use the results of this method to provide a confidence score to new senses added to the OMW. A second way to account for the quality of the data will be to encourage lexicographers and native speakers around the world to check, correct and enrich these lists once the data has been published. Following a crowd-sourced schema similar to the one used by Rosetta Project to produce these lists, we hope to ask subscribers of well-known listservers, such as the Linguistlist,⁷ for help correcting an enriching this data-set.

A second line of future work will focus on the further development of tools to disseminate and make this data useful for as many people as possible. As it was mentioned in Section 5., enriching the newly created interface with other lists used in research, e.g. linguistics or psychology, can help create a positive feedback of open data. Also, by providing the ability to create and save custom lists, we hope that people can be creative in the way they use these open resources.

Finally, concerning the words that were excluded during this round of linking, we hope to continue the expansion trend of wordnets, and soon include prepositions and conjunctions as two new classes of concepts. And since prepositions are an specially interesting class of word to study crosslingually, our first target will be prepositions. English prepositions are often translated as nouns in Chinese and Japanese: for example *between* is translated as 間 *aida* “space or region between” in Japanese. Towards this end, we hope to build on existing semantic taxonomies for prepositions such as Schneider et al. (2015).

7. Acknowledgements

This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13). We would also like to thank Shu-Kai Hsieh, Huang Chu-Ren and Laurent Prevot for providing the initial mapping that was used in this work.

References

Stephen Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly. (www.nltk.org/book).

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of*

the 51st Annual Meeting of the Association for Computational Linguistics, pages 1352–1362. ACL, Sofia, Bulgaria.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, and Quentin D Atkinson. 2012. Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Simon J Greenhill, Robert Blust, and Russell D Gray. 2008. The austronesian basic vocabulary database: from bioinformatics to lexicomics. *Evolutionary Bioinformatics*, 4:271.

Martin Haspelmath and Uri Tadmor. 2009. *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

Eric W Holman, Søren Wichmann, Cecil H Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica*, 42(3-4):331–354.

Chu-Ren Huang, Laurent Prevot, I-Li Su, and Jia-Fei Hong. 2007. Towards a conceptual core for multicultural processing: A multilingual ontology based on the swadesh list. In *Intercultural Collaboration*, pages 17–30. Springer.

Luis Morgado da Costa and Francis Bond. 2016. Wow! What a useful extension! Introducing non-referential concepts to wordnet. In *Proceedings of the International Conference on Language Resources and Evaluation*. Slovenia.

Mark Pagel, Quentin D Atkinson, Andreea S Calude, and Andrew Meade. 2013. Ultraconserved words point to deep language ancestry across eurasia. *Proceedings of the National Academy of Sciences*, 110(21):8471–8476.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A hierarchy with, of, and for preposition supersenses. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 112–123.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Maurizio Serva and Filippo Petroni. 2008. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos. *Proceedings of the American philological society*, 96(4):452–463.

⁷<http://linguistlist.org/>

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

James Woodward. 1993. Intuitive judgments of Hong Kong signers about the relationship of sign language varieties in Hong Kong and Shanghai. *CUHK Papers in Linguistics*, 4:88–96.

Ren Wu, Yuya Matsuura, and Hiroshi Matsuno. 2015. On generating language family-trees based on basic vocabulary. In *ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications*, pages 272–275.

Annex A: The 200 word list proposed by Morris Swadesh in 1952, including parenthetical explanations

I, thou, he, we, ye, they, this, that, here, there, who?, what?, where?, when?, how, not, all, many, some, few, other, one, two, three, four, five, big, long, wide, thick, heavy, small, short, narrow, thin, woman, man (male human), person, child (young person rather than as relationship term), wife, husband, mother, father, animal, fish, bird, dog, louse, snake, worm, tree, woods, stick (of wood), berry (of fruit), seed, leaf, root, bark (of tree), flower, grass, rope, skin (person's), meat (flesh), blood, bone, fat (organic substance), egg, tail, feather (larger feathers rather than down), hair, head, ear, eye, nose, mouth, tooth (front rather than molar), tongue, foot, leg, hand, wing, belly, guts, neck, back (person's), heart, liver, to drink, to eat, to bite, to suck, to spit, to vomit, to blow (of wind), breathe, to laugh, to see, to hear, to know (facts), to think, to smell (perceive odor), to fear, to sleep, to live, to die, to kill, to fight, to hunt (game), to hit, to cut, to split, to stab (or stick), to scratch (as with fingernails to relieve itch), to dig, to swim, to fly, to walk, to come, to lie (on side), to sit, to stand, to turn (change one's direction), to fall (drop rather than topple), to give, to hold (in hand), to squeeze, to rub, to wash, to wipe, to pull, to push, to throw, to tie, to sew, to count, to say, to sing, to play, to float, to flow, to freeze, to swell, sun, star, water, to rain, river, lake, sea (ocean), salt, stone, sand, dust, earth (soil), cloud, fog, sky, wind, snow, ice, smoke (of fire), fire, ashes, to burn (intrans.), road (or trail), mountain, red, green, yellow, white, black, night, day (opposite of night rather than the time measure), year, warm (of weather), cold (of weather), new, old, good, bad (deleterious or unsuitable), rotten (especially log), dirty, straight, sharp (as knife), dull (knife), smooth, wet, dry (substance), right (correct), near, far, right (hand), left (hand), at, in, with (accompanying), and, if, because, name

Annex B: The 215 word list organized by semantic groups, published by Morris Swadesh in 1955

- (a) **personal pronouns:** *I, thou, we, he, ye, they*
- (b) **interrogatives:** *who, where, what, when, how*
- (c) **correlatives** *and, if, because*
- (d) **locatives:** *at, in, with*

- (e) **location:** *there, far, near, right (side), here, that, this, left(side)*
- (f) **position and movement:** *come, sit, give, fly, stand, hold, fall, swim, turn, walk, throw, pull, float, flow, lie, push*
- (g) **manipulations:** *wash, split, tie, hit, wipe, cut, rub, dig, scratch, squeeze*
- (h) **time periods:** *year, day, night*
- (i) **numerals:** *one, two, three, four, five, six, seven, eight, nine, ten, twenty, hundred*
- (j) **quantitatives:** *all, few, many, some*
- (k) **size:** *wide, thick, long, thin, narrow, big, small, short*
- (l) **natural objects and phenomena:** *ice, salt, star, sun, wind, sky, cloud, rain, water, sea, smoke, snow, sand, stone, mountain, ashes, earth, dust, lake, fog, river, fire*
- (m) **plants and plant parts:** *bark, leaf, grass, tree, root, flower, woods, seed, berry (fruit), stick*
- (n) **animals:** *worm, snake, louse, fish, dog, animal, bird*
- (o) **persons:** *person (human being), woman, child, man*
- (p) **body parts and substances:** *blood, ear, hand, tongue, tooth, foot, egg, back, tail, meat (flesh) eye, feather, skin, bone, head, mouth, nose, wing, heart, fat, guts, belly, neck, hair, liver, leg*
- (q) **body sensations and activities:** *drink, die, hear, see, sleep, live, eat, know, bite, fear, think, breathe, vomit, smell*
- (r) **oral activities:** *laugh, sing, suck, cry, spit, speak*
- (s) **colors:** *black, green, red, white, yellow*
- (t) **descriptives:** *old, dry, good, new, warm, rotten, cold, sharp, right (correct), straight, smooth, bad, wet, dull, dirty*
- (u) **kinship:** *brother, sister, father, mother, husband, wife*
- (v) **cultural objects and activities:** *sew, rope, shoot, hunt, cook, count, play, clothing, work, dance, spear, stab, fight*
- (w) **miscellaneous:** *name, other, not, burn, blow, freeze, swell, road, kill*

Annex C: The 100 word list proposed by Morris Swadesh in 1955

all, ashes, bark, belly, big, bird, bite, black, blood, bone, burn, cloud, cold, come, die, dog, drink, die, ear, earth, eat, egg, eye, fat (grease), feather, fire, fish, fly, foot, give, good, green, hair, hand, head, hear, heart, I, kill, know, leaf, lie, live, long, louse, man, many, meat (flesh), mountain, mouth, name, neck, new, night, nose, not, one, person (human being), rain, red, road (path), root, sand, see, seed, sit, skin, sleep, small, smoke, stand, star, stone, sun, swim, tail, that, this, thou, tongue, tooth, tree, two, walk, warm (hot), water, we, what, white, who, woman, yellow, say, moon, round, full, knee, claw, horn, breast

Annex D: The widely used 207 word list with Princeton Wordnet 3.0 Mappings

Synsets starting with the numeral 7 are part of the expanded wordnet and are not in yet included in PWN.

English	PWN3.0	English	PWN3.0	English	PWN3.0	English	PWN3.0
I	77000015-n	stick/ club	04317420-n	smell	02124748-v	sand	15019030-n
thou	77000021-n	fruit	13134947-n	fear	01780202-v	dust	14839846-n
he	77000046-n	seed	11683989-n	sleep	00014742-v	earth	14842992-n
we	77000002-n	leaf	13152742-n	live	02614181-v	cloud	09247410-n
you	77000019-n	root	13125117-n	die	00358431-v	fog	11458314-n
they	77000031-n	bark	13162297-n	kill	01323958-v	sky	09436708-n
this	77000061-n	flower	11669335-n	fight	01090335-v	wind	11525955-n
that	77000079-n	grass	12102133-n	hunt	01143838-v	snow	11508382-n
here	08489497-n	rope	04108268-n	hit	01400044-v	ice	14915184-n
there	08489627-n	skin	01895735-n	cut	01552519-v	smoke	13556893-n
who	77000095-n	meat	07649854-n	split	02030158-v	fire	13480848-n
what	77000091-n	blood	05399847-n	stab	01230350-v	ashes	14769160-n
where	77000084-n	bone	05269901-n	scratch	01250908-v	burn	00377002-v
when	77000104-n	fat	05268965-n	dig	01309701-v	road	04096066-n
how	77000090-n	egg	01460457-n	swim	01960911-v	mountain	09359803-n
not	00024073-r	horn	01325417-n	fly	01940403-v	red	04962784-n
all	02269286-a	tail	02157557-n	walk	01904930-v	green	04967191-n
many	01551633-a	feather	01896031-n	come	01849221-v	yellow	04965661-n
some	01552634-a	hair	05254393-n	lie	01547001-v	white	04960729-n
few	01552885-a	head	05538625-n	sit	01543123-v	black	04960277-n
other	02069355-a	ear	05320899-n	stand	01546768-v	night	15167027-n
one	13742573-n	eye	05311054-n	turn	01907258-v	day	15164957-n
two	13743269-n	nose	05598147-n	fall	01972298-v	year	15201505-n
three	13744044-n	mouth	05301908-n	give	02199590-v	warm	02529264-a
four	13744304-n	tooth	05282746-n	hold	01216670-v	cold	01251128-a
five	13744521-n	tongue	05301072-n	squeeze	01387786-v	full	01211531-a
big	01382086-a	finger nail	05584265-n	rub	01249724-v	new	01640850-a
long	01433493-a	foot	05563266-n	wash	00557686-v	old	01638438-a
wide	02560548-a	leg	05560787-n	wipe	01392237-v	good	01123148-a
thick	02410393-a	knee	05573602-n	pull	01609287-v	bad	01125429-a
heavy	01184932-a	hand	02440250-n	push	01871979-v	rotten	01070538-a
small	01415219-a	wing	02151625-n	throw	01508368-v	dirty	00419289-a
short	01436003-a	belly	05556943-n	tie	00141632-v	straight	02314584-a
narrow	02561888-a	guts	05534333-n	sew	01329239-v	round	02040652-a
thin	02412164-a	neck	05546540-n	count	00948071-v	sharp	00800826-a
woman	10787470-n	back	05558717-n	say	00979870-v	dull	00800248-a
man	10287213-n	breast	05554405-n	sing	01729431-v	smooth	02236842-a
person	00007846-n	heart	05388805-n	play	01072949-v	wet	02547317-a
child	09918248-n	liver	05385534-n	float	01904293-v	dry	02551380-a
wife	10780632-n	drink	01170052-v	flow	02066939-v	correct	00631391-a
husband	10193967-n	eat	01168468-v	freeze	00445711-v	near	00444519-a
mother	10332385-n	bite	01445932-v	swell	00256507-v	far	00442361-a
father	10080869-n	suck	01169704-v	sun	09450163-n	right	02031986-a
animal	00015388-n	spit	00101956-v	moon	09358358-n	left	02032953-a
fish	02512053-n	vomit	00076400-v	star	09444783-n	at	<i>excluded</i>
bird	01503061-n	blow	02100632-v	water	14845743-n	in	<i>excluded</i>
dog	02084071-n	breath	00001740-v	rain	15008607-n	with	<i>excluded</i>
louse	02183857-n	laugh	00031820-v	river	09411430-n	and	<i>excluded</i>
snake	01726692-n	see	02150948-v	lake	09328904-n	if	<i>excluded</i>
worm	01922303-n	hear	02169702-v	sea	09426788-n	because	<i>excluded</i>
tree	13104059-n	know	00594621-v	salt	07813107-n	name	06333653-n
forest	09284015-n	think	00629738-v	stone	09416076-n		

The Role of Computational Zulu Verb Morphology in Multilingual Lexicographic Applications

Sonja Bosch, Laurette Pretorius

University of South Africa

PO Box 392, UNISA, Pretoria, South Africa 0003

E-mail: boschse@unisa.ac.za pretol@unisa.ac.za,

Abstract

Performing cross-lingual natural language processing and developing multilingual lexicographic applications for languages with complex agglutinative morphology pose specific challenges that are aggravated when such languages are also under-resourced. In this paper, Zulu, an under-resourced language spoken in Southern Africa, is considered. The verb is the most complex word category in Zulu. Due to the agglutinative nature of Zulu morphology, limited information can be computationally extracted from running Zulu text without the support of sufficiently reliable computational morphological analysis by means of which the essential meanings of, amongst others, verbs can be exposed. The central research question that is addressed in this paper is as follows: How could ZulMorph (<http://gama.unisa.ac.za/demo/demo/ZulMorph>), a finite state morphological analyser for Zulu, be employed to support multilingual lexicography and cross-lingual natural language processing applications, with specific reference to Zulu verbs?

Keywords: Zulu verb morphology, multilingual lexicography, semantic interoperability

1. Introduction

Web-scale knowledge graphs¹, based on the Resource Description Framework (RDF) data model, form an essential part of the Semantic Web. Representing growing amounts of information, in many different languages and then providing computational infrastructure to perform cross-lingual information gathering, is one of the challenges of the Multilingual Semantic Web (MSW), specifically for under-resourced languages. Exposing any information encapsulated in running text, as RDF triples in the MSW requires the transformation/extraction of such information into RDF triples according to Linked (Open) Data (LOD) principles (Heath & Bizer, 2011). This essentially means that information needs to be transformed into simple statements of the form (*subject, predicate, object*) that can then, in turn, be linked together to form more complex concepts or information networks. The *subject* is the resource described by the statement, which is uniquely identified by its so-called URI (Uniform Resource Identifier). The *object* represents the content of the statement and can be either a simple string or a resource with its own URI. The *predicate* provides the semantic link between the subject and the object and describes the meaning of the relation between them. The predicate has its own URI and is also referred to as a property.

Also the broad field of language and linguistic resources has not remained untouched by the Semantic Web and its standards². Growing volumes of lexicons, corpora, dictionaries, etc., in many languages are being published in the Linguistic Linked (Open) Data (LLOD)³ cloud and are playing and increasingly important role in the realisation of the MSW (for example, Garcia, 2015; Declerck et al., 2015). For under-resourced languages to

be part of the MSW, both information in such languages and language resources for these languages should be exposed in the LOD and the LLOD, respectively (Pretorius, 2014). In this paper we focus on the hypothetical situation where certain information is only available in one language, viz. Zulu. Zulu is a Bantu language spoken in Southern Africa. It is an agglutinative, morphologically complex language and in terms of the availability of language resources and language technology Zulu is considered an under-resourced language.

In order to support the exposition of this information in Zulu in RDF, the relevant concepts and the relations between them need to be identified as a first step. For the purposes of our discussion we will also assume that we have already identified the concepts that constitute the subject and the object of the RDF triple and that the challenge is to find the relation (predicate or verb) that relates them semantically, and then to render the relationship in such a way that its meaning is known and can be accessed in other languages, in our case English. This is also important for cross-lingual information representation and semantic interoperability, as we hope to demonstrate.

The verb is the morphologically most complex word category in Zulu. This means that in order to identify the relation between two concepts as accurately as possible, the morphological analysis of the Zulu verb in question has to be available together with some indication of the English meaning. ZulMorph is a rather extensive, mature, finite state morphological analyser that arguably constitutes one of the most complete computational models of Zulu morphology that has been developed up to now⁴. It is proposed in this work as a useful tool towards

¹ <http://lod-cloud.net/>

² https://www.w3.org/2001/sw/wiki/Main_Page

³ <http://linguistic-lod.org/lod-cloud-Feb2016.php>

⁴ isiZulu.net (2016) is a Zulu-English online dictionary that offers users bidirectional lookups and automatic morphological decomposition of Zulu words. Morphologically complex Zulu words are translated into literal English translations. Although

supporting relation-mining in Zulu for the MSW. The most important contribution of this paper is the detailed exposition of the morphology of the Zulu verb and its extensions, its continued accurate modelling in ZulMorph and the systematic addition of semantic support in English without which the high quality verb analyses that ZulMorph produces, remain unexposed, both to users that do not already know Zulu, and more important, to the vast knowledge graphs and algorithms of the MSW.

In the first part of the paper we provide an overview of the salient features of Zulu verb morphology, showing that it is, amongst others, characterised by sequences of suffixes, including so-called verbal extension morphemes, which play an important role in Zulu (verb) semantics. The basic meaning of a verb root may be modified to a greater or lesser degree by the suffixing of such extension morphemes, e.g. *-hamb-* (go/travel); *-hamb-is-* (send off), *-hamb-isis-* (travel fast); *-hamb-el-* (visit); *-hamb-is-an-* (accompany). Both morphological and semantic challenges that pertain to verbal extensions are discussed.

As second part, we then provide a brief overview of ZulMorph, an existing online finite state morphological analyser for Zulu⁵, with specific reference to how verb morphology is modelled computationally: We explore different ways of handling verbal extensions, keeping in mind the challenges that we mentioned above. The central issue discussed here concerns the interplay between the morphology (structure) and the lexicon (lexical semantics). Our first attempt in this paper to quantify the occurrence and range, and (computationally) capture the significance of verbal extensions in Zulu, on the basis of existing Zulu dictionaries, is novel.

The third part of the paper considers how ZulMorph can be enhanced so as to allow the representation of cross-lingual verb semantics in relation to, for example, English, as a kind of pivot language. We demonstrate the possible use of ZulMorph in this regard with examples from bilingual (Zulu-English) e-lexicography for English-speaking language learners of Zulu. Finally, and as a proof of concept, we show, also by means of an example from a selected Zulu Wikipedia article, we believe for the first time, how a Zulu morphological analyser may be used to provide cross-lingual support for information extraction from Zulu text.

The **paper** is concluded with a short summary of the contribution, as well as plans and ideas for future work.

2. Zulu Verb Morphology

The morphological composition of the verb is considerably more complex than that of any other word category in Zulu. A number of slots, preceding and also following the verb root may contain numerous

morphemes with functions such as derivations, inflection for tense-aspect and marking of nominal arguments. Examples are cross-reference of the subject and object by means of class- (or person-/number-) specific markers, locative affixes, morphemes distinguishing verb forms in clause-final and non-final position, negation morphemes and so forth. In this paper we concentrate on the so-called verb extension morphemes.

In the inflectional morphology of Zulu the basic meaning of a verb root in Zulu may be modified by suffixing one or more extension morphemes to the verb root⁶, e.g.

- (1a) *-bon-a* > *-bona* (see)
-verb.root-terminative
- (1b) *-bon-is-a* > *-bonisa* (show)
-verb.root-caus.ext-terminative
- (1c) *-bon-an-a* > *-bonana* (see each other/greet one another)
-verb.root-reciproc.ext-terminative
- (1d) *-bon-is-an-a* > *-bonisana* (show each other)
-verb.root-caus.ext-reciproc.ext-terminative

It is significant that the verb root *-bon-* may use 29 different combinations of verb extensions of which 7 feature as headwords in Doke and Vilakazi (1964:83-85):

- (2a) *-bonakalisa* (make visible, bring into view, disclose, reveal, indicate);
- (2b) *-bonakala* (appear, come into vision, be visible obvious, evident, be revealed, found out);
- (2c) *-bonana* (see each other/greet one another)
- (2d) *-bonelela* (treat with consideration, treat leniently)
- (2e) *-bonela* (see for, perceive for, convey greetings, copy, imitate, plan out, prepare ahead, improve)
- (2f) *-bonisela* (look after for)
- (2g) *-bonisa* (cause to see, show, direct, inform, explain).

In the outer matter, Doke and Vilakazi (1964:ix) indicate that separate entries have been made for “verbal derivatives” (extended verb stems) that “convey some meaning or idiomatic usage not deducible from the inherent significance of the derivative form”, e.g.

- (3a) *-hamba* (travel, move along)
- (3b) *-hamb-el-a* (visit, be on good terms with)

In other cases, where the “inherent significance of the derivative form” is easily deducible from the basic verb stem, the derivative forms are listed in brackets after the entry of the basic form, e.g.

- (4) *-pikiza* (wriggle about, waggle) (pass. *pikizwa*; ap. *pikizela*; caus. *pikizisa*)

To give an idea of the productive use of verb extensions in Zulu, it can be mentioned that of a total of 8031 basic verb

isiZulu.net would go a long way to support multilingual lexicography, the authors are not aware of the use of this online dictionary as component in any natural language processing applications.

⁵ Currently only one word at a time can be analysed in the online demo version of ZulMorph.

⁶ For purposes of convenience a verb root followed by one or more extensions, is called an extended root in this paper.

roots, 22 of them use between 20 and 30 combinations of one or more verb extensions as entered by Doke and Vilakazi (1964).

2.1 Morphological Challenges

Within a rule-based approach to morphology, the following are examples of morphological challenges (morphotactics and morphophonological alternation rules) that are encountered with regard to verb extensions:

a) Some basic verb roots resemble extended verb roots, e.g. the verb root *-hlangan-* (come together; unite; connect) in which the morpheme *-an-* resembles the reciprocal extension. In this case it is not an extension but part of the verb root.

b) Rule-based palatalisation occurs in the formation of passives when the final syllable of a verb root begins with a bilabial consonant, also when such a verb root is separated from the passive extension *-w-* by another extension, e.g.

(5a) *-boph-a* (tie, fasten, button up) >
 -verb.root-terminative
-bosh-w-a (be tied, fastened, buttoned up)
 -verb.root-pass.ext-terminative

(5b) *-boph-el-a* (tie for, imprison for) >
 -verb.root-appl.ext-terminative
-bosh-el-w-a (be tied for, be imprisoned for)
 -verb.root-appl.ext-pass.ext-terminative

Occasionally however, idiosyncrasies occur when bilabials appearing elsewhere in the verb root are palatalised, e.g.

(6) *ezisetshenziswa* (that are used):
-sebenz-is-w-a >
 -verb.root-caus.ext-pass.ext-terminative
-setshenz-is-w-a
 (not *-sebenziswa** as expected)

c) The order of extension suffixes is not always fixed. For instance the passive extension usually follows other extensions, e.g.

(7a) *-akh-el-w-a* (be built for)
 -verb.root-appl.ext-pass.ext-terminative

In some cases, the reciprocal follows the passive extension, e.g.

(7b) *-akh-el-w-an-a* (be built for each other)
 -verb.root-appl.ext-pass.ext-recip.ext-terminative
 (cf. van Eeden, 1956:657)

2.2 Semantic Challenges

In most grammatical descriptions of the Bantu languages, verb extensions are considered to be inflectional suffixes since “they do not change the word category to which a word belongs, but add a regular, predictable meaning to

the word” (Kosch, 2006:109). The predictable meanings of extended verb roots can be summarised as follows⁷:

passive > (be, being)
 applicative > (for, on behalf of)
 causative > (cause to, help)
 intensive > (expresses intensity)
 neuter > (cause or assist to perform an action)
 reciprocal > (each other)

Not all verb roots may take all extensions arbitrarily since there are restrictions on the combinations of certain meanings (Poulos & Msimang, 1998:183). The following examples are ungrammatical (*) because the neuter extension is incompatible with the meaning of the two verbs and therefore signifies a semantic restriction:

(8a) *-ephuka* (get broken; die suddenly) > *-ephuk-ek-a**
 (8b) *-shona* (sink, go down, die etc.) > *-shon-ek-a**

Exceptions occur when the meaning of an extended verb root is lexicalised, and therefore becomes unpredictable to a large extent. Kosch (2006:106) singles out a suffix such as the causative which is prone to lexicalisation in combination with certain verb roots. The result is an unpredictable meaning and a display of derivational properties, e.g.

(9a) *-bon-a* (see)
 -verb.root-terminative
-bon-is-a (show)
 -verb.root-caus.ext-terminative

(9b) *-lum-a* (bite, suffer sharp pain, itch)
 -verb.root-terminative
-lum-is-a (cause to bite/itch; give a bite of food to/share with)
 -verb.root-caus.ext-terminative

The applicative extension is also used to indicate “in a direction” when followed by a noun indicating location, e.g.

(10) *-gijim-el-a ezintabeni* (seek shelter in the mountains)
 -verb.root-appl.ext-terminative

3. ZulMorph: An Overview

3.1 Finite State Approach

The ZulMorph finite state computational morphological analyser for Zulu was developed with the Xerox finite state toolkit (Beesley & Karttunen, 2003), but has also been successfully compiled with Foma (Hulden, 2009). The two central problems of morphology, viz. *morphotactics* (rules for morpheme sequencing) and *morphophonological alternation rules* (rules for spelling

⁷ Also cf. de Schryver (2010:178).

and sound changes) are computationally modelled by and implemented as finite state transducers, which are then composed to form one single transducer, which constitutes the morphological analyser. For modelling the morphotactics **lexc** with its cascading continuation classes of morpheme lexicons (Beesley & Karttunen, 2003:210) is provided and for the alternations rules, **xfst**, a language for using the extensive Xerox finite state calculus, is used. An important and useful construct offered by the mentioned toolkits is that of so-called *flag diacritics*. Flag diacritics provide a light-weight approach to feature-setting and feature-unification operations for enhancing modelling accuracy and runtime efficiency. Specific uses are to enforce separated dependencies and mark idiosyncratic morphotactic behaviour (see Beesley & Karttunen, 2002) for a comprehensive exposition). In **lexc** and **xfst** flag diacritics are so-called multicharacter symbols with a distinctive spelling:

@operator.feature.value@ and
 @operator.feature@ where the operators are
 P (positive (re)setting), N (negative (re)setting), R (require test), D (disallow test), C (clear feature) and U (unification test). The features and values are specified by the user.

In ZulMorph flag diacritics are used extensively to, amongst others, model the Zulu noun class system (Bosch & Pretorius, 2002; Pretorius & Bosch, 2003), long distance dependencies (Pretorius & Bosch, 2008), part of speech information and a wide variety of other morphotactic constraints that apply in Zulu. In this paper the focus is on their use for annotating each basic verb root with its valid and attested extension sequences, as discussed in Section 3.2.

As discussed in Section 2, the various verbal extensions are not compatible with all verb roots, and there are no hard and fast rules that determine the possible combinations, i.e. roots with extensions, as well as extensions with one another. Such information is not available elsewhere - not even paper dictionaries provide complete information on combinations and sequences for all verb roots. The inclusion of such comprehensive “idiosyncratic” information about verb roots and their (semantically) valid extensions in ZulMorph further emphasises its role as one of the most comprehensive computational models of Zulu morphology yet.

It is well-known that the coverage of a finite state morphological analyser such as ZulMorph is determined by (i) the accurate and complete modelling of the morphological structure of the language, and (ii) the comprehensiveness of the noun stem and verb root lexicons. Only valid Zulu surface forms of which the noun stems or verb roots are present in the respective lexicons, can be analysed correctly. For such a morphological analyser to be maximally useful, these stem and root lexicons need to be maintained and extended as new words enter the language. Various approaches are possible in this regard. For example, the use of a so-called guesser variant of the morphological analyser (outside the scope of this paper) and the regular application of the analyser to

new Zulu corpora⁸ to identify new stems and roots for inclusion into the analyser. This remains ongoing work.

However, this distinction between morphology and the root lexicon becomes somewhat fuzzy in the case of the Zulu verb and its extensions in that the attested extension sequences are marked on the basic roots and thereby become part of the “lexicon”.

Simplistically speaking, we model the extensions (the morphology) and the roots (the lexicon) as separate **lexc** continuation classes with the verb roots in the LEXICON VRoot and the short list of possible extensions are in their own continuation class, LEXICON VExtNew, as shown in the **lexc** script fragment in Section 3.2.1. In order to model sequences of extensions, we merely allow the iteration of the extension continuation class.

However, we also have the obligation to address the challenges that were mentioned in Section 2. The morphological challenges are at present treated as follows: Challenge (a) concerns the common ambiguity of human language for which no real solution exists except to deal with it through semantic context-based disambiguation at a later stage of processing – at morphological level such limited over-generation will thus occur; Challenge (b) is non-rule-based and is met by hand-crafting the analyser to accurately model all the individual known cases; Challenge (c) is taken care of by the above-mentioned simplistic iteration model, which is inherently prone to over-generation. However, this challenge may also be viewed as a “semantic” challenge since, as was discussed in Section 2, the sequences of extensions are semantically determined.

Semantic challenges arise from the fact that extensions and their sequencing are semantically determined and may not be valid for all verb roots. This means that mere iteration (as above) is semantically not sufficiently accurate. Moreover, the semantics of an extension (sequence) is either predictable or lexicalised. Modelling approaches in these cases are discussed in the next section.

3.2 Modelling the Verb and Adding Semantics

The modelling of the verb and its extensions in ZulMorph are presented by means of a simple **lexc** example. In four steps we systematically extend the example to cover the following four aspects:

- Simple iteration (unattested (new) extension sequences);
- Attested extension sequences and the verb roots with which they may occur;
- Predictable meaning;
- Lexicalised meaning.

3.2.1. A **lexc** script for simple iteration

The **lexc** script fragment⁹ below will accept any arbitrary

⁸ Zulu is a resource-scarce language and the availability and development of high quality free and open corpora remain a challenge.

⁹ The detailed explanation of the **lexc** language and the example script fall outside the scope of the article. The interested reader is referred to Beesley and Karttunen (2003).

(finite) sequence of extensions that is included in LEXICON VExtNew, even sequences that are semantically not plausible. This implementation is useful for the purposes of mining new sequences of extensions from a corpus. The example script is based on the basic root *-bon-*, discussed in examples 1, 2 and 9.

The basic root resides in the LEXICON VRoot and the extensions in LEXICON VExtNew, which is cyclic and will continue to process extensions until none is found. The next expected morpheme is in LEXICON VerbTerm.

```

Multichar_Symbols
@U.CL.15@ @U.SYL.POLY@ @R.Verb.ON@ ^BR ^ER [ATT]
@P.Basic.ON@ @R.Basic.ON@ @D.Basic@
...
LEXICON BeginVRootMarker
0:^BR VRoot;

LEXICON VRoot
bon@P.Basic.ON@ VPSClass15;

LEXICON VPSClass15
@U.CL.15@@U.SYL.POLY@ EndVRootMarker;

LEXICON EndVRootMarker
[VRoot]:^ER VExt;

LEXICON VExt
@R.Basic.ON@ VExtNew;
@D.Basic@ VExtAttested;

LEXICON VExtNew
! Recursion to cater for unknown extension sequence
akal[NeutExt]:akal VExtNew;
an[RecipExt]:an VExtNew;
ek[NeutExt]:ek VExtNew;
el[ApplExt]:el VExtNew;
is[CausExt]:is VExtNew;
isis[IntensExt]:isis VExtNew;
elel[IntensExt]:elel VExtNew;
elez[IntensExt]:elez VExtNew;
w[PassExt]:w VExtNew;
iw[PassExt]:iw VExtNew;
!
@R.Verb.ON@ VerbTerm;

LEXICON VerbTerm
! Addition of verb final morpheme

```

3.2.2. Annotating the verb root with its attested extension sequences

ZulMorph contains 8031 basic roots and 28477 verb roots with attested extension sequences, bringing the number of entries in the verb root lexicon of ZulMorph to approx. 36000. From the extensive data harvested from paper dictionaries, including Doke and Vilakazi (1964), 133 different extension sequences were identified, with the first 30 most frequent sequences representing more than 98% of all attested extensions. Statistics were also accumulated about with the number of extension per basic verb root. The basic verb root with the most number of extensions, viz. 30, is *-fan-* (resemble). The basic root *-bon-* of our examples in Section 2 has 28 extension sequences. Moreover, Zulmorph contains 6153 basic verb roots that have at least one attested extension and 1878

that have no extensions.

To show how verb roots are annotated with their attested extension sequences, we extend the example as follows:

For each of the 113 extension sequences in the comprehensive list of attested extension sequences, we define two unique flag diacritics @P.ExtEL.ON@, @R.ExtEL.ON@, ..., @P.ISANISIS.ON@ and @R.ISANISIS.ON@. We then extend the LEXICON VRoot in Section 3.2.1 as follows:

```

LEXICON VRoot
bon@P.Basic.ON@ VPSClass15;
bon@P.ExtW.ON@ VPSClass15;
bon@P.ExtAKAL.ON@ VPSClass15;
bon@P.ExtEL.ON@ VPSClass15;
bon@P.ExtAN.ON@ VPSClass15;
bon@P.ExtIS.ON@ VPSClass15;
bon@P.ExtISIS.ON@ VPSClass15;
bon@P.ExtELEL.ON@ VPSClass15;
bon@P.ExtAKALEL.ON@ VPSClass15;
bon@P.ExtAKALIS.ON@ VPSClass15;
bon@P.ExtAKALISIS.ON@ VPSClass15;
bon@P.ExtAKALISW.ON@ VPSClass15;
bon@P.ExtAKALISEL.ON@ VPSClass15;
bon@P.ExtWAN.ON@ VPSClass15;
bon@P.ExtANEL.ON@ VPSClass15;
bon@P.ExtANIS.ON@ VPSClass15;
bon@P.ExtELW.ON@ VPSClass15;
bon@P.ExtELEL.ON@ VPSClass15;
bon@P.ExtELAN.ON@ VPSClass15;
bon@P.ExtELIS.ON@ VPSClass15;
bon@P.ExtELELW.ON@ VPSClass15;
bon@P.ExtELELAN.ON@ VPSClass15;
bon@P.ExtISW.ON@ VPSClass15;
bon@P.ExtISEK.ON@ VPSClass15;
bon@P.ExtISEL.ON@ VPSClass15;
bon@P.ExtISAN.ON@ VPSClass15;
bon@P.ExtISELW.ON@ VPSClass15;
bon@P.ExtISELEL.ON@ VPSClass15;
bon@P.ExtISELAN.ON@ VPSClass15;

```

We also add LEXICON VExtAttested to the *lexc* fragment in Section 3.2.1. By way of illustration we show only three of the 113 entries. The tag [ATT] on the analysis side of an entry indicates that the sequence is an attested one. Since complete extension sequences are modelled with single entries no iteration is necessary.

```

LEXICON VExtAttested
...
el[ApplExt]@R.ExtEL.ON@[ATT]:el@R.ExtEL.ON@
VerbTerm;
is[CausExt]@R.ExtIS.ON@[ATT]:is@R.ExtIS.ON@
VerbTerm;
is[CausExt]el[ApplExt]@R.ExtISEL.ON@[ATT]:isel
@R.ExtISEL.ON@ VerbTerm;
...

```

3.2.3. Adding predictable meaning

By semi-automatically adding basic meanings to the 8031 basic verb roots and by including the predictable meanings of the 10 extensions in LEXICON VExtNew, we are able to provide a first approximation of the meaning of each of the ~36000 entries in the LEXICON

VRoot. Keeping in mind that the extensive Princeton Wordnet for English has 11529 verbs, the ZulMorph coverage of the Zulu extended verb root semantics is quite significant and can already be used, as alluded to in Section 1.

Adding the mentioned predictable meaning is illustrated by further extending the example:

Replace `bon` with `bon[[see]]` on the analysis side of all the *-bon-* entries in Section 3.2.2. so that, for example,

```
bon@P.ExtEL.ON@ VPSClass15;
```

becomes

```
bon[[see]]@P.ExtEL.ON@:bon@P.ExtEL.ON@
  VPSClass15;
```

Also replace

```
el[ApplExt]:el VExtNew;
```

with

```
el[ApplExt][[for,on_behalf_of]]:el VExtNew;
```

in LEXICON VExtNew and replace

```
el[ApplExt]@R.ExtEL.ON@[ATT]:el@R.ExtEL.ON@
  VerbTerm;
```

with

```
el[ApplExt][[for,on_behalf_of]]@R.ExtEL.ON@[ATT]
]:el@R.ExtEL.ON@ VerbTerm;
```

in LEXICON VExtAttested.

An analysis (only partially shown here) of a verb with root *-bon-* and extension *-el-* will then render

```
...bon[[see]]el[ApplExt][[for,on_behalf_of]]...
```

3.2.4. Adding lexicalised meaning

Adding lexicalised meaning is the most resource intensive part of endowing ZulMorph verb analyses with accurate lexical semantics since it has to be added manually. The process is as follows: For each basic verb root and a particular extension sequence for which a lexicalised meaning is available, the meaning of the *basic* root is replaced by the lexical meaning of the *extended* root. The entry will also be marked as such so that the regular meanings of the extensions are no longer displayed. We use *-bonisa-* as example.

We introduce the flag diacritics `@P.Lex.ON@`, `@R.Lex.ON@` and `@D.Lex@`, and in LEXICON VRoot replace

```
bon@P.ExtIS.ON@ VPSClass15;
```

with

```
bon[[show]]@P.Lex.ON@@P.ExtIS.ON@:bon@P.Lex.ON@
@P.ExtIS.ON@ VPSClass15;
```

We extend LEXICON VExt as follows:

```
LEXICON VExt
@R.Basic.ON@      VExtNew;
@D.Basic@D.Lex@  VExtAttested;
@D.Basic@R.Lex.ON@ VExtLexicalised;
```

and add a LEXICON VExtLexicalised (in which none of the 113 extension sequences has its meaning provided).

An analysis (only partially shown here) of a verb with root *-bon-* and extension *-is-* will then render¹⁰

¹⁰ The somewhat counter-intuitive position of the English

```
...bon[[show#]]is[CausExt]...11
```

In summary, by annotating each entry in the verb root lexicon with its meaning (either predictable or lexicalised) and by providing the meanings of the 113 extension sequences, the morphological analysis of any Zulu verb will contain sufficient semantic information to support a basic notion of semantic linking/interoperability - a possibility that did not exist before.

4. Cross-lingual Verb Semantics

4.1 Bilingual E-lexicography

We explore the possible use of ZulMorph in the context of e-lexicography for language learners, in particular for English-speaking language learners of Zulu. Bothma (2011:72) emphasises that user needs should be on the forefront when decisions are made on the implementation of information technologies in e-lexicography. The latter should enhance access to information in terms of user needs. Given the complex morphology of verbs in Zulu, as described earlier on, the language learner when confronted with a Zulu text, needs inflected verb forms to be normalised to a root form with its accompanying meaning, e.g.

```
yaqala:
qal[VRoot]
start/begin
```

```
kuhlangana:
hlangan[VRoot]a[VT]
come together; unite; connect
```

In the case of suffixed verb extensions, the user also has the need to have quick access to the predictable or “regular” meanings of verb extensions, and ultimately to the unpredictable lexicalised meanings of verb roots and their extensions, as in Figure 1.

```
ezi setshenziswa:
ezi[RC][10]sebenz[[be_used]][VRoot]is[CausExt]w[
PassExt][ATT]a[VT]
ezi[RC][8]sebenz[[be_used]][VRoot]is[CausExt]w[P
assExt][ATT]a[VT]
zivumelekile:
zi[SC][10]vum[[agree]][VRoot]el[ApplExt]ek[NeutE
xt]ile[VTPerf]
zi[SC][10]vum[[be_allowed#]][VRoot]el[ApplExt]ek
[NeutExt][ATT]ile[VTPerf]
zi[SC][8]vum[[agree]][VRoot]el[ApplExt]ek[NeutE
xt]ile[VTPerf]
zi[SC][8]vum[[be_allowed#]][VRoot]el[ApplExt]ek[
NeutExt][ATT]ile[VTPerf]
```

Figure 1: Examples of analyses

meaning is post-processed for human consumption.

¹¹ # denotes lexicalised meaning

The normalisation process, as required by the user, could be facilitated by a web service of ZulMorph, providing cross-lingual support for information extraction from Zulu text.

4.2 Cross-lingual Support for Information Extraction

As a proof of concept and by way of illustration the following sentence from the article on Shaka in the Zulu Wikipedia¹² is considered:

UShaka uzalwa indlovukazi uNandi kaBhebhe nenkosi uSenzangakhona kaJama.

Shaka is born (by) queen Nandi of Bhebhe and king Senzangakhona of Jama.

A possible hypothetical (*subject, predicate, object*) triple that could be extracted from this short text is

(Shaka , is_born by , queen Nandi).

From the ZulMorph analyses in Figure 2 the lexicalised meaning “be_born_by” and the predictable meaning “bear,give_birth”, together with the passive extension “be,being”, are obtained. At this point there are various possibilities for linking (only briefly mentioned and considered as future work):

Firstly, “be_born_by” is semantically equivalent to “isChildOf”, which is already a property in various vocabularies/ontologies in the SW with URI’s such as <http://purl.org/saws/ontology#isChildOf>, <http://purl.org/mont/mont.owl#isChildOf> and

<http://purl.org/vocab/relationship/childOf>.

By further including, for example, such URIs in ZulMorph as part of the verb semantics, the linking is accomplished once and for all.

Secondly, the Princeton WordNet (PWN) contains the concept “bear, birth, deliver, give birth, have”, with URI <http://wordnet-rdf.princeton.edu/wn31/200056644-y>, which is semantically equivalent to the predictable meaning “bear,give_birth”. It would then require some additional logic to combine the PWN URI with the passive extension to obtain the triple

(queen Nandi , give birth , Shaka).

```

ushaka:
u[NPrePre][1a]shaka[[Shaka]].1a-2a[NStem]

uzalwa:
u[SC][1]zal[[be_born_by*]][VRoot]w[PassExt][ATT]
a[VT]
u[SC][1]zal[[bear,give_birth]][VRoot]w[PassExt][
[be,being]][ATT]a[VT]

indlovukazi:
i[NPrePre][9]n[BPre][9]dlovukazi[[queen]].9-10[N

```

¹² <https://zu.wikipedia.org/wiki/Shaka>

```

Stem]

unandi:
u[NPrePre][1a]nandi[[Nandi]].1a-2a[NStem]
kabhebhe:
ka[PosSKA]u[NPrePre][1a]bhebhe[[Bhebhe]].1a-2a[N
Stem]

nenkosi:
na[AdvPre]i[NPrePre][9]n[BPre][9]khosi[[king]].9
-6[NStem]

usenzangakhona:
[NPrePre][1a]senzangakhona[[Senzagakhona]].1a-2a
[NStem]

kajama:
ka[PosSKA]u[NPrePre][1a]jama[[Jama]].1a-2a[NStem
]

```

Figure 2: ZulMorph analyses of the words in the sentence¹³

5. Conclusion and Future Work

In this paper we made the point that exposing Zulu verb semantics through computational morphological analysis can play an important support role in making Zulu and the information encoded in it available in the MSW. We substantiated our claim by explaining the complexity and the challenges of Zulu verb morphology - specifically with respect to verbal extensions to the root, their modelling and implementation in ZulMorph and the comprehensive coverage that has been achieved. The large number (8031) of basic verb roots, the comprehensive list of 113 attested extension sequences and the verb root lexicon with 36000 entries, representing comprehensive attested information about the extended verb roots in Zulu, provide a solid basis for principled verbal lexical semantics as part of the morphological analysis of the verb, even in new unattested cases.

In summary, ZulMorph constitutes an NLP component or tool that could serve as a starting point for exposing information encoded in Zulu as Linked Data in the MSW. In terms of future work, two topics for further investigation are the use of new Zulu corpora for the continued improvement of ZulMorph and the inclusion of dereferenceable URIs as part of the verb semantics, as briefly mentioned in Section 4.2.

Important areas of future application include multilingual HLT-oriented and e-lexicography, morphological analysis as a service in multilingual and cross-lingual contexts, and multilingual semantic interoperability.

¹³ Due to length restrictions, the tags used in the morphological analysis are not provided. The interested reader is referred to Bosch and Pretorius (2011) for a partial list.

7. References

- Beesley, K.R., Karttunen, L. (2003). *Finite State Morphology*. Stanford: CSLI Publications.
- Bosch, S.E., Pretorius, L. (2002). The significance of computational morphological analysis for Zulu lexicography. *South African Journal of African Languages*, 22(1), pp. 11-20.
- Bosch, S.E., Pretorius, L. (2011). Towards Zulu corpus clean-up, lexicon development and corpus annotation by means of computational morphological analysis. *South African Journal of African Languages*, 31(1), pp. 138-158.
- Bothma, T.J.F.D. (2011). Filtering and adapting D and information in an online environment in response to user needs. In: P.A. Fuertes-Olivera and H Bergenholtz (eds), *e-Lexicography – The Internet, Digital Initiatives and Lexicography*. London: Continuum.
- Declerck, T., Wand-Vogt, E. and Mörth, K. (2015). Towards a Pan European lexicography by means of Linked (Open) Data. In: I. Kosem, M. Jakubiček, J. Kallas and S. Krek (eds.), *Proceedings of eLex 2015*, Ljubljana: Trojina, Institute for Applied Slovene Studies.
- De Schryver, G-M. (2010). Revolutionizing Bantu lexicography – A Zulu case study. *Lexikos* 20, pp. 161–201.
- Doke, C.M. and Vilakazi, B.W. 1964. *Zulu-English Dictionary*. Johannesburg: Witwatersrand University Press.
- Gracia, J. (2015). Multilingual dictionaries and the Web of Data. In: *Kernerman Dictionary News*, 23, pp.1-4. Tel Aviv: KDictionaries.
- Heath, T., Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, pp.1-136. Morgan & Claypool.
- Hulden, M. (2009). Foma: a finite-state compiler and library. In: *Proceedings of the EACL 2009, Demonstrations Session*, pp. 29-32, Athens: Association for Computational Linguistics.
- isiZulu.net. (2016). Available at <https://isizulu.net/>
- Kosch, I.M. (2006). *Topics in Morphology in the African Language Context*. Pretoria: Unisa Press.
- Poulos, G., Msimang, C.T. (1998). *A linguistic analysis of Zulu*. Pretoria: Via Afrika.
- Pretorius, L. (2014). The Multilingual Semantic Web as virtual knowledge commons: The case of the under-resourced South African languages. In: P. Buitelaar and P. Cimiano (eds), *Towards the Multilingual Semantic Web*, Springer, 2014.
- Pretorius, L., Bosch, S.E. (2003). Finite-State Computational Morphology: An Analyzer Prototype For Zulu. *Machine Translation*, Special issue on finite-state language resources and language processing, 18, pp.195-216.
- Pretorius, L., Bosch, S. (2008). Containing overgeneration in Zulu computational morphology. *Southern African Linguistics and Applied Language Studies*, 26(2), pp. 209-216.
- Van Eeden, B.I.C. (1956). *Zoeloe Grammatika*. Stellenbosch: Universiteitsuitgewers en Boekhandelaars (Edms.) Beperk.

CombiNet. A Corpus-based Online Database of Italian Word Combinations

Valentina Piunno

Roma Tre University
Via Ostiense 236 - 00146 Rome, Italy
valentina.piunno@uniroma3.it

Abstract

This paper introduces *CombiNet* dictionary, an on line corpus-based lexicographic tool representing combinatorial properties of Italian lexemes, developed by Roma Tre University, University of Pisa and University of Bologna. The lexicographic layout of *CombiNet* is designed to include different sets of information, such as i) syntactic configurations and ii) syntactic function of word combinations, iii) degree of lexical variation associated with specific types of multiword units. In fact, *CombiNet* records word combinations showing different degrees of lexicalizations and paradigmatic variability, which is a novelty in lexicography. This investigation intends to tackle several issues associated with *CombiNet*, and in particular it aims at a) showing procedures and methods used to create and compile *CombiNet*'s entries, b) describing particular types of combinatorial phenomena emerged from the analysis of corpus-based data, c) illustrating the lexicographic layout that has been elaborated for word combinations representation, d) describing the advanced research tool *CombiNet* is equipped with, a useful device for lexicographic investigations as well as for lexicological analysis.

Keywords: word combinations, corpus-based, Italian lexicographic database

1. Introduction

This contribution is carried out within a research project dealing with word combinations in Italian and aiming at the realization of a corpus-based online combinatory dictionary¹. Recent investigations² have shown the importance of word combinations in the lexicon of languages, and have introduced specific methods for their lexicological analysis and lexicographic representation. In particular, Italian lexicon is extremely rich of regular combinatorial phenomena, some of which still need further investigation.

This paper describes *CombiNet* dictionary, a lexicographic tool representing some of the most frequent Italian word combinations. *CombiNet*'s interface is already online, and the database is subject to a constant implementation and is continuously updated. This contribution will describe the methods that have been applied for the extraction of the most representative word combinations of Italian and the procedures developed to represent the dictionary's entries.

¹ *CombiNet* is a Research Project funded by the Italian Ministry for Education, University and Research (MIUR), developed by Roma Tre University, University of Pisa and University of Bologna. Coordinators: Raffaele Simone (from 2013 to 2014) and Alessandro Lenci (from 2015 to 2016). Project title: PRIN Project 2010-2011 (n. 20105B3HE8) "*CombiNet* - Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary".

² Among others, Goldberg (2006), Simone (2007). As far as Italian language is concerned, cf. Voghera (1994, 2004), Simone (2006), Simone & Masini (2007), Simone *et al.* (2015), Masini (2012), Piunno (2013, in press), Piunno *et al.* (2013). As far as lexicographical works are concerned, cf. Simone (2009), Lo Cascio (2013).

The paper is structured as follows: the next section summarizes the main features of *CombiNet* dictionary and is devoted to the description of data extraction methods and to the presentation of the dictionary's lexicographic layout. The last section describes *CombiNet* as a tool for lexicological investigations; some examples of possible queries and related results will be provided.

2. CombiNet

CombiNet is an online dictionary of Italian word combinations containing data extracted from corpora according to statistical association measures. The dictionary represents different types of combinatorial phenomena, which have been classified on the basis of the most recent theoretical achievements in linguistics.

Before *CombiNet*, other dictionaries of Italian word combinations have been published³. In fact, the interest in word combinations is a growing phenomenon in Italian lexicographic studies. However the lexicographic classification and representation of combinatorial phenomena may vary considerably from one dictionary to another, according to the different parameters taken into accounts, such as⁴: i) the theoretical classification of combinatorial units and its possible representations in the entries, ii) types of recorded combinatorial phenomena, iii) the lexicographic layout, iv) the intended target.

The lexicographic layout of *CombiNet* has been conceived taking into account the general outline of different types of combinatorial dictionaries of European languages; lexicographic entries have been shaped to include the several types of word formats (syntactic

³ Cf. in particular Lo Cascio (2013) and Urzì (2009).

⁴ Cf. Piunno *et al.* (2013) for the analysis and comparison of several European combinatory dictionaries.

configurations) that have been identified in the lexicographic and lexicological literature concerning Italian language.

CombiNet differs from other lexicographic works on Italian combinations and it can be considered as a *unicum*, being characterized by the following features:

- i) its theoretically-based classification of combinatorial phenomena is clearly identifiable in the meta-language or in the lexicographic entry layout;
- ii) it is the first Italian combinatory dictionary containing word combinations which are not lexically specified and allow paradigmatic variability;
- iii) its layout is original and is specifically oriented to record the most representative word combinations and the most productive syntactic patterns;
- iv) it includes information about the argument/adjunct structure of verbal entries;
- v) it has been designed to extract linguistic data for quantitative and qualitative lexicological analysis;
- iii) it has been designed to have a digital interface and to be freely available on line.

2.1 Lexicographic Entries

Different types of words have been selected as lexical entries of *CombiNet* dictionary, i.e. nouns, verbs and adjectives. The lemmatization process has followed the following procedure. Firstly, a set of lexemes has been selected from the ones included in the *Senso Comune* base knowledge⁵, and organized on the basis of their frequency of occurrence in *La Repubblica* corpus⁶. Secondly, the most frequent lexical entries have been chosen according to their combinatorial properties and selected for the inclusion in the database: thus, lexemes allowing a higher number of word combinations have been selected as entries. Up to now, nearly 400 entries have been completed, but only a set of about 200 has been uploaded in the database, which is constantly updated.

2.2 Data Extraction

Word combinations included in the entries have been extracted from two Italian corpora (*La Repubblica* corpus and *Paisà* corpus⁷), according to statistical association measures. Two different extraction technologies have been applied to gather data: the *EXTra* software (Passaro & Lenci, 2016) and the *LexIt* tool (Lenci *et al.*, 2012).

The *EXTra* software (Passaro & Lenci, 2016, Castagnoli *et al.*, 2015; 2016) has been created for the extraction of different types of word combinations, on the basis of predefined Part-of-Speech patterns, characterized by i) a specific syntactic order and ii) a restricted number of syntactic slots. A set of specific PoS sequences has been collected on the basis of the most recent lexicological works on Italian word combinations

and Multiword Expressions (hereinafter MWEs)⁸. We have produced a list of the most frequent MWEs patterns differentiated according to their specific function (nominal, adjectival, adverbial, verbal). For example, we have identified nearly 40 PoS patterns creating Italian MWEs with an adjectival function (that is to say, having an adjectival sequence as their output), such as:

Syntactic configuration	Examples
Adj + Adj	<i>bianco sporco</i> 'off white' (<i>lit.</i> white dirty)
Adj + Conj + Adj	<i>vero e proprio</i> 'real' (<i>lit.</i> true and appropriate)
Past Part. + Noun	<i>fatto in casa</i> 'homemade' (<i>lit.</i> made in home)
Noun + Prep + Noun	<i>chiavi in mano</i> 'turnkey' (<i>lit.</i> keys in hand)
Prep + Noun	<i>a colori</i> 'colour' (<i>lit.</i> at colours)
Prep + Adj + Noun	<i>di seconda mano</i> 'second hand' (<i>lit.</i> of second hand)

Table 1: Examples of PoS patterns of Italian Multiword Adjectives

This tool allowed us to collect different sets of information about extracted combinations, such as i) the log likelihood ratio, ii) the absolute frequency of occurrence in the corpora, iii) the morpho-syntactic features of extracted words (e.g. the presence of the article).

LexIt is an online resource able to collect information about the distributional features of Italian nouns, verbs and adjectives (Lenci *et al.*, 2012). *LexIt* represents the argument structure of lexical items as a syntactic and semantic frame structure. This tool is able to extract the most important distributional properties of lexical units, through specific measures of associations (e.g. Local Mutual Information) (Lenci *et al.*, 2014).

It is worth noting that data extracted from corpus strictly reflect the corpus nature and sometimes are not truly representative of real language use. For example, data extracted from *La Repubblica* sometimes reflect the fact that the corpus is based on newspaper texts, thus sometimes revealing misleading information about Italian language combinatorial features⁹.

2.3 Data Representation

CombiNet lexicographic layout is designed to represent the following information about the word combinations where the entry occurs:

⁸ Voghera (1994, 2004), Simone & Masini (2007), Masini (2012), Piuanno (2013, 2015, in press).

⁹ Thus, for example, the log-likelihood of *trovare un accordo* ('to reach an agreement) is strangely lower than that of *trovare un cadavere* ('to find a corpse').

⁵ <http://www.sensocomune.it>

⁶ <http://sslmit.unibo.it/repubblica>

⁷ <http://www.corpusitaliano.it>

- a) the *word formats* (namely, the specific syntactic configurations concerning word combinations), represented as phrasal structures: e.g. [Noun + Preposition + Noun]
- b) the *function* of word combinations, that is to say the "output" of a word format: e.g. [Noun + Preposition + Noun]_{NOUN}
- c) the combinatorial profile (or 'type'), namely the nature of the multiword expression (e.g. collocations, multiword lexemes, light verb and support verb constructions, binomial constructions, idiomatic expressions, interjections, proverbs)
- d) the degree of lexical variation associated with a specific type of word combination.

One of the hallmarks of *CombiNet* lies in (d), which is a novelty in lexicography. In fact, the dictionary represents word combinations showing different degrees of lexicalizations, as well as combinatorial phenomena characterized by paradigmatic variability.

According to the most significant studies on word combinations¹⁰, we believe that word combinations can be distinguished into two different groups: *completely filled sequences* vs *partially filled sequences*.

The former are fully lexically specified (they are stable combinations in terms of lexical features), syntactically fixed and do not allow any lexical or syntactic variation. On the contrary, the latter are characterized by a lower degree of fixedness and cohesion (and as such, they show a lower degree of lexical specification). Moreover, partially filled combinations are represented as containing "empty positions", which can be filled according to specific morpho-syntactic and semantic restrictions. They can be represented as syntactic patterns having a fixed slot and a variable one, the latter subjected to some specific semantics restrictions.

(1) [*fixed slot* + VARIABLE LEXEME_{Semantic Restriction}]

(2) [*dare per* + ADJ/PAST|PRES_PART] = 'consider'
(*lit.* to give for)

dare per favorito ('odds-on-favourite', *lit.* to give for favourite)

dare per morto ('give up for dead', *lit.* to give for dead)

dare per buono ('consider valid', *lit.* to give for good)

dare per spacciato ('give up for dead', *lit.* to give for dead)

dare per assodato ('take for granted', *lit.* to give for ascertain)

dare per scomparso ('consider as missing', *lit.* to give for missing)

(3) [NOUN_{Device} + *alla mano*] = 'x_{NOUN} ready to be used'
(*lit.* at the hand)

armi *alla mano* ('weapons ready', *lit.* weapons at the hand)

pistola *alla mano* ('guns ready', *lit.* gun at the hand)

documenti *alla mano* ('documents at the ready', *lit.* documents at the hand)

carte *alla mano* ('papers at the ready', *lit.* papers at the hand)

statistiche *alla mano* ('statistics at the ready', *lit.* statistics at the hand)

dati *alla mano* ('data at the ready', *lit.* data at the hand)

(4) [*avere* + DET + NOUN + *facile*] = 'to be inclined to do something connected with x_{NOUN}'
(*lit.* to have + x_{NOUN} + easy)

avere il bicchiere *facile* ('to be likely to drink alcoholic drinks', *lit.* to have the glass easy)

avere il grilletto *facile* ('to be trigger-happy', *lit.* to have the trigger easy)

avere la lacrima *facile* ('to cry at the drop of a hat', *lit.* to have the tear easy)

avere la pistola *facile* ('to be likely to use the gun', *lit.* to have the gun easy)

avere la battuta *facile* ('to be quick on the draw', *lit.* to have the joke easy)

Thus, partially filled profiles are represented in *CombiNet* as lists of word combinations i) sharing similar syntactic configurations, ii) having variable slots iii) slots can be often filled selecting from a range of semantically connected lexemes, iv) having similar semantic properties and restrictions, iv) iii) occurring in similar syntactic contexts. Both "completely filled" and "partially filled" word combinations are recorded in the dictionary. Thus, for example, the nominal entry *mano* 'hand' will include word combinations characterized by variable degrees of lexical specification. Thus, we will find both completely filled combinatorial profiles (e.g. collocations *dorso della mano* 'back of the hand'; multiword lexemes, *stretta di mano* 'handshake'; idiomatic expressions *avere le mani in pasta* 'to have a finger in the pie') and partially filled word combinations (cf. example (3)).

Combinatorial profiles are distinguished through a set of printing marks:

i) Underlined typeface marks word combinations showing a high degree of lexicalization and cohesion,

ii) Topographic position in the entry: the entry is divided into different fields, and each field is devoted to include specific types of word combinations (e.g. completely or partially filled units),

iii) Specific tags (proverbs and idiomatic expressions are signalled as such through specific labels).

¹⁰ With particular reference to the works of Fillmore *et al.* (1988), Goldberg (2006), Simone (2007).

3. The Lexicographic Layout

As in conventional dictionaries, *CombiNet* provides different types of information and each field of the entry is devoted to a specific purpose. Each entry is associated with a word class category (i.e. noun, verb or adjective) and a subcategory (e.g. masculine or feminine for nouns or adjectives, transitive or intransitive for verb). Furthermore, entries are subdivided into different broad sense blocks, each one containing a brief definition of the entry.

The screenshot shows the entry for 'mano' with the following structure:

- Entry:** Home / Lemma / mano
- Word class:** mano sostantivo f. s.
- Subcategory:** Indietro, Mostra tutte, + Crea nuova accezione
- Senses:**
 1. Estremità dell'arto superiore.
 2. Stile, impronta di qcn.
 3. Nel gioco delle carte, singola fase di una partita.

Figure 1: *CombiNet*'s entry

Sense blocks of verbal entries also contain a 'syntactic frame', representing the argument /adjunct structure of the verb. Each verbal sense contains almost one syntactic frame.

The screenshot shows the entry for 'aprire' with the following structure:

- Entry:** Home / Lemma / aprire
- Word class:** aprire¹ verbo transitivo
- Subcategory:** Indietro, Mostra tutte, + Crea nuova accezione
- Senses:**
 - 1.a Disgiungere le parti unite di un oggetto. SOGG aprire OGG (Avv)
 - 1.b Allargare, stendere (anche *fig.*). SOGG aprire OGG (Avv)
 - 2.a Cominciare, avviare. SOGG aprire OGG (Avv)

Annotation: Syntactic frame

Figure 2: *CombiNet*'s verbal entry *chiamare* 'to call'

Each sense block of an entry registers its own combinatorial types and examples, in a four columns layout, where each column provides a specific lexicographic purpose:

- Column 1: *Categoria* ('Category');
- Column 2: *Struttura* ('Structure');
- Column 3: *Dati Primari* ('Primary Data');
- Column 4: *Dati Secondari* ('Secondary Data').

The screenshot shows the entry for 'mano' with the following structure:

- Entry:** Home / Lemma / mano
- Word class:** mano sostantivo f. s.
- Subcategory:** Indietro, Nascondi tutte, + Crea nuova accezione
- Senses:**
 1. Estremità dell'arto superiore

Annotation: Syntactic frame

Figure 3. *CombiNet*'s four column layout

The first column is devoted to represent the "output" category (that is to say the syntactic function of a word combination) or, in some cases, a specific morpho-syntactic or functional property of the word combination. The dictionary includes a set of thirteen "pre-packaged" categories so far, such as noun, adjective, adverb, preposition, etc. For example, the word combination *bagaglio a mano* 'hand luggage' is a multiword lexeme performing the same syntactic function as the noun, and accordingly, it is recorded as a noun.

The second column contains the "structure" of a word combination, namely the PoS sequence or syntactic configuration associated with a multiword unit. For example, the combination *bagaglio a mano* ('hand baggage', *lit.* 'baggage at hand') is recorded as [Noun Preposition Noun]. The following table represents just some examples of word combinations including the nominal entry *mano* 'hand', which are represented as categories and structures.

Category	Structure	Examples
Noun	Noun + Adjective	<i>mano destra</i> 'right hand' (<i>lit.</i> hand right)
Noun	Adjective + Noun	<i>ultima mano</i> 'last hand' (<i>lit.</i> last hand)
Noun	Noun + Prep + Noun	<i>mano di vernice</i> 'coat of paint' (<i>lit.</i> hand of paint)
Noun	Noun + Prep + Noun	<i>stretta di mano</i> 'handshake' (<i>lit.</i> grasp of hand)
Prep	Prep + Noun + Prep	<i>per mano di</i> 'at the hands of' (<i>lit.</i> for hand of)
Verb	Verb + (Det) + Noun	<i>alzare la mano</i> 'raise the hand' (<i>lit.</i> raise the hand)

Table 2: Categories and structures of word combinations including the nominal entry *mano* 'hand'

The third and the fourth columns are devoted to the representation of the real word combinations extracted from corpora, distinguished on the basis of lexical variability. The third column ('Primary Data') records examples of completely filled combinations, while the last column ('Secondary Data') represents partially filled ones, as in the following Figure.

The screenshot shows the entry for 'mano' with the following structure:

- Entry:** Home / Lemma / mano
- Word class:** mano sostantivo f. s.
- Subcategory:** Indietro, Nascondi tutte, + Crea nuova accezione
- Senses:**
 1. Estremità dell'arto superiore

Annotation: Syntactic frame

Figure 4: *CombiNet*'s entry layout

4. *CombiNet* as a Tool for Lexicological Investigations

CombiNet differs from similar lexicographic works in that it is not just a combinatory dictionary, but a tool aiming at the analysis of word combinations from both a lexicological and a lexicographic point of view. Indeed, *CombiNet* is equipped with an integrated query system, allowing several types of lexicographic investigation and gathering resources for lexicological quantitative and qualitative analysis.

4.1 Simple Query

SIMPLE QUERY can be used to find out:

- i) a particular entry,
- ii) words beginning or ending with a specific set of letters (the user has to supply a character string, also using a wild card),
- iii) an entry belonging to a specific word class (i.e. noun, verb, adjective) or to a particular grammatical category (e.g. transitive verbs, feminine nouns, etc.).

Figure 5: Simple Query interface

All the above queries can be simultaneously activated, so as to narrow the results. SIMPLE QUERY provides the user with a list of recorded entries.

4.2 Advanced Query

CombiNet is equipped with an ADVANCED QUERY tool allowing different types of search combinations.

As for SIMPLE QUERY, also ADVANCED QUERY allow the user to search for the entry's attributes, such as i) a specific part of speech, ii) any grammatical information associated with the headword.

ADVANCED QUERY is also useful to find out a particular sense block or the information about verbal argument/adjunct structure. Furthermore, the user can explore the combinatorial properties of an entry, and in particular the information about categories, structures or specific combinatorial profiles (included completely and partially filled patterns).

Also in this case, more parameters can be combined together in a multifunctional query.

Figure 6: Advanced Query interface

All types of queries give both qualitative and quantitative results, and the user can select the ones he is interested in. For example, it is possible to find all fixed multiword combinations containing a specific lemma, as in the following Figure, showing word combinations with the entry *mano* 'hand'.

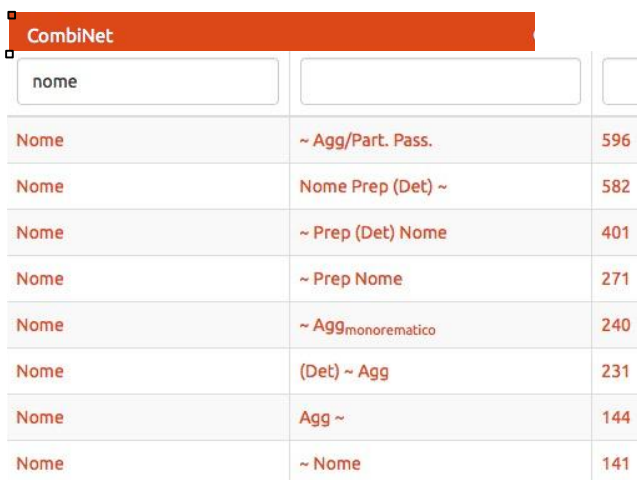
Singolo esempio	Indice statistico	Frequenza assoluta
(pl.) a piene mani [generosamente]	356.091.809.616	414
(pl.) a mani nude [senza armi o senza guanti]		
(pl.) a mani vuote		
(pl.) in buone mani		
(pl.) Mani Pulite		
(pl.) nelle mani di		

Figure 7: Example of possible Advance Research results

The results include various types of multiword items: adverbs (e.g. *a piene mani*, *a mani nude*, *a manu vuote*),

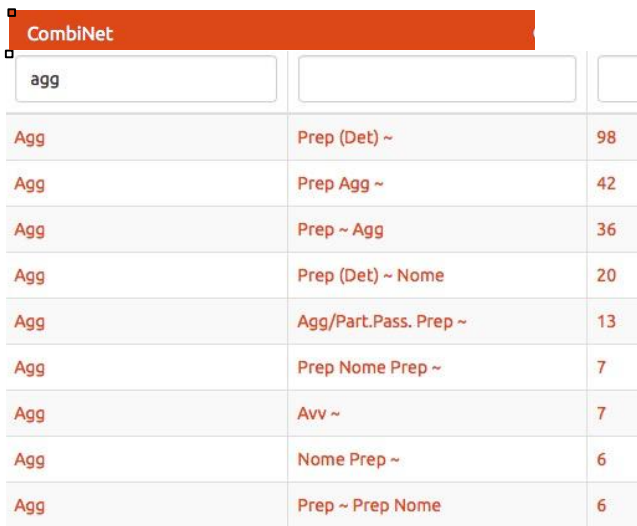
adjectives (e.g. *in buone mani*), nouns (e.g. *Mani Pulite*), as well as prepositions (e.g. *nelle mani di*). It is worth noting that this type of results also allow to search for statistical information associated with the word combination: each combination displays a statistical index (the log-likelihood ratio) and the absolute frequency as extracted from corpora through EXTra or LexIt.

Furthermore, *CombiNet* is not only able to represent the combinatory frequency of word sequences, but it also extracts quantitative information on categories, structures and combinations. As a result, it also allows to identify the most frequent and productive word combination formats in Italian. The following Figures show quantitative information about word combinations pattern recorded in *CombiNet*.



CombiNet		
nome		
Nome	~ Agg/Part. Pass.	596
Nome	Nome Prep (Det) ~	582
Nome	~ Prep (Det) Nome	401
Nome	~ Prep Nome	271
Nome	~ Agg _{monorematico}	240
Nome	(Det) ~ Agg	231
Nome	Agg ~	144
Nome	~ Nome	141

Figure 8: *CombiNet*'s syntactic patterns for Italian nominal multiword



CombiNet		
agg		
Agg	Prep (Det) ~	98
Agg	Prep Agg ~	42
Agg	Prep ~ Agg	36
Agg	Prep (Det) ~ Nome	20
Agg	Agg/Part.Pass. Prep ~	13
Agg	Prep Nome Prep ~	7
Agg	Avv ~	7
Agg	Nome Prep ~	6
Agg	Prep ~ Prep Nome	6

Figure 9: *CombiNet*'s syntactic patterns for Italian adjectival multiword

The diagrams above show that most used syntactic patterns for Italian nominal multiword are [Noun + Adjective]_{NOUN} and [Noun + Preposition + Noun]_{NOUN}, and that the most productive syntactic configurations used as adjectives are [Preposition + Noun]_{ADJ} or [Preposition + Adjective + Noun]_{ADJ}.

Finally, *CombiNet* is able to collect the partially filled patterns (Figure 11) - together with their semantic restrictions- included in the entries. This section, however, is still in need of improvement and refinement through the continuous feeding of data.



CombiNet

Home / Pattern parzialmente riempiti

Pattern parzialmente riempiti

Visualizzo 1-20 di 241 elementi.

Contenuto
[~ Det N _{cibo}]
[~ Det N _{finanziario}]
[~ Det N _{ruolo}]
[N _{prodotto} umano <i>dell</i> ~]
[N <i>di</i> ~]
[~ Det Nome]
[~ Det Nome _{professione}]
[~ <i>senza</i> N]

Figure 11: Examples of *CombiNet*'s partially filled patterns

5. Conclusion

The starting point of our investigation was the identification of mere Italian word combinations and the analysis and the classification of different combinatorial types. However, the development of our investigation and the design of *CombiNet* shed light on a rich multitude of phenomena and issues which deserve further investigation, thus highlighting new research perspectives.

CombiNet has turned out to be a valuable device for the description and the characterization of Italian lexicon as a collection of different lexical structures:

1. It should suffice it to mention phenomena such as partially filled combinations and the role of semantic restrictions for slot filling. Combinatorial phenomena need a further distinction related to their response to lexical variability: partially filled units sharing specific semantics and morpho-syntactic properties are to be represented in the dictionary as productive patterns or "semantic types" (Bybee 1985, Bybee and Thompson 1997).
2. Secondly, the analysis of word formats of Italian also raises the general problem of the identification of possible word formats in other languages.
3. Thirdly, many combinatorial phenomena have not property been considered by the literature, such as word combinations containing a compulsory negated pattern (e.g. *non vedere l'ora* 'to look forward', *lit.* not to see the

hour), or intensification patterns (e.g. *innamorato pazzo* 'smitten', *lit.* in love mad).

Finally, the collection of combinatorial sequences of *CombiNet*'s entries shed light on another possibly important lexicological issue. Italian lexicon is not simply monorhematic, but it is extremely rich of combinatorial phenomena (Simone *et al.*, 2015). It will suffice to mention the fact that each *CombiNet*'s entry contains on average about 50 combinatorial examples.

This necessarily impacts also on the automatic processing of language data and requires an implementation of tagging and extraction technologies.

6. Acknowledgements

I am extremely grateful to Raffaele Simone for his precious comments on a previous draft of this paper. Any mistakes or inaccuracies that might still remain in the text are my sole responsibility.

7. Bibliographical References

- Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., Mazzoleni, M. (2004). Introducing the La Repubblica corpus: a large, annotated, TEI(XML)-compliant corpus of newspaper Italian". In Lino, M.T., Xavier, M.F., Ferreira, F., Costa, R., Silva, R. (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Paris: ELRA, 1771-1774.
- Bybee, J.L. 1985. *Morphology: A Study into the Relation between Meaning and Form*, Amsterdam: John Benjamins.
- Bybee, J., Thompson, S. (1997). Three Frequency Effects in Syntax, *Berkeley Linguistic Society*, 23, 65-85.
- Castagnoli, S., Lebani, G. E., Lenci, A., Masini, F., Nissim, M., Piuanno, V. (2015). Towards a corpus-based online dictionary of Italian Word Combinations Automatic Knowledge Acquisition for Lexicography, *COST ENeL WG3 meeting*, Herstmonceux Castle, 13 August 2015. http://www.elxicography.eu/wp-content/uploads/2015/10/Paper_Castagnoli_et_al_ENeL_final.pdf
- Castagnoli, S., Lebani, G. E., Lenci, A., Masini, F., Nissim, M., & Passaro, L.C. (2016). POS-patterns or Syntax? Comparing methods for extracting Word Combinations. In Corpas Pastor, G. (Ed.). *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives (Full papers) - Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües (Trabajos completos)*. Geneva, Switzerland: Tradulex, 101-114.
- Fillmore, Ch.J., Kay, P., O'Connor, M.C. (1988). Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language*, 64, 501-38.
- Goldberg, A (2006). *Constructions at work*, Oxford: Oxford University Press.
- Lenci, A., Lapesa, G., Bonansinga, G. (2012). LexIt: A Computational Resource on Italian Argument Structure. *LREC 2012*, 3712-3718.
- Lo Cascio, V. (2013). *Dizionario Combinatorio Italiano*. Amsterdam/Philadelphia: John Benjamins.
- Passaro, L.C., Lenci, A. (2016). Extracting Terms with EXTra, *Paper presented at EUROPHRAS 2015 -Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*. Malaga, Spain, 29 June -1 July 2015.
- Piuanno, V. (2013). *Modificatori sintagmatici con funzione aggettivale e avverbiale*, PhD Thesis, Roma: Roma Tre University.
- Piuanno, V. (2015), Sintagmi Preposizionali come Costruzioni Aggettivali, *Studi e Saggi di Linguistica*, 53, 1, 65-98.
- Piuanno, V. (in press) Multiword Modifiers in Romance languages. Semantic formats and syntactic templates. *Yearbook of Phraseology*, Berlin: Mouton de Gruyter.
- Piuanno, V., Masini, F., Castagnoli, S. (2013), *Technical report: Studio comparativo dei dizionari combinatori dell'italiano e di altre lingue europee*. TRIPLE, Roma Tre University and University of Bologna.
- Simone, R. (2006). Nominales sintagmáticos y no-sintagmáticos. In De Miguel E. *et al.* (Eds.), *Estructuras léxicas y estructuras del léxico*. Frankfurt am Main: Peter Lang. 225-246.
- Simone, R. (2007). Constructions and categories in verbal and signed languages. In Pietrandrea, P. *et al.* (Eds.), *Verbal and Signed Languages. Comparing Structures, Constructs, and methodologies*, Berlino-New York: Mouton-De Gruyter, 198-252.
- Simone, R. (2009). *Grande dizionario analogico della lingua italiana*. Torino: UTET.
- Simone, R., Masini, F. (2007), "Support nouns and verbal features: a case study from Italian", In Grezka, A., Martin-Berthet, F. (Eds.), *Verbes et classes sémantiques*, Special issue of *Verbum*.
- Simone, R., Masini, F., Piuanno, V., Castagnoli, S. (2013). Combinazioni di parole in italiano: risorse lessicografiche e proposte di tipologia", Paper presented at *XLVII Congresso Internazionale SLI*, Salerno 26-28 September 2013.
- Simone, R., Piuanno, V., Cominetti, F. (2015). Le categorie lessicali preesistono ai dati o originano dai dati?, Paper presented at *XLIX Congresso internazionale di Studi della SLI Tipologia e 'dintorni'. Il metodo tipologico alla intersezione di piani d'analisi*, Malta 24-26 September 2015.
- Urzi, F. (2009). *Dizionario delle combinazioni lessicali*. Lussemburgo: Convivium.
- Voghera, M. (1994). Lessemi complessi: percorsi di lessicalizzazione a confronto. *Lingua e Stile*. 29. 2: 185-214.
- Voghera, M. (2004). Polirematiche. In Grossmann, M. - Rainer, F. (Eds.) *La formazione delle parole in italiano*. Tübingen: Max Niemeyer Verlag. 56-69.

Creating seed lexicons for under-resourced languages

Ivett Benyeda, Péter Koczka, Tamás Váradi

Research Institute for Linguistics of the Hungarian Academy of Sciences

H-1068 Benczúr utca 33., Budapest

{benyeda.ivett, koczka.peter, varadi.tamas} @nytud.mta.hu

Abstract

In this paper we present methods of creating seed dictionaries for an under-resourced language, Udmurt, paired with four thriving languages. As reliable machine readable dictionaries do not exist in desired quantities this step is crucial to enable further NLP tasks, as seed dictionaries can be considered the first connecting element between two sets of texts. For the language pairs discussed in this paper, detailed description will be given of various methods of translation pair extraction, namely Wik2Dict, triangulation, Wikipedia article title pair extraction and handling the problematic aspects, such as multiword expressions (MWUs) among others. After merging the created dictionaries we were able to create seed dictionaries for all language pairs with approximately a thousand entries, which will be used for sentence alignment in future steps and thus will aid the extraction of larger dictionaries.

Keywords: under-resourced languages, dictionary extraction, seed dictionaries, comparable corpora

1. Introduction

In this paper we will present a method of creating seed dictionaries for four language pairs: Udmurt–Russian, Udmurt–Finnish, Udmurt–English and Udmurt–Hungarian. The research demonstrated in this paper is part of a project whose aim is to support small Finno-Ugric languages in generating on-line content. The goal of this project is to create bilingual dictionaries and parallel corpora for six small Finno-Ugric (Udmurt, Komi-Permyak, Komi-Zyrian, Hill Mari, Meadow Mari and Northern Sámi) languages paired with four thriving ones which are important for these small communities. For creating these sources a seed dictionary is essential in the process. In this paper we are focusing on the Udmurt language and demonstrate the process of creating seed dictionaries for language pairs where Lang1 is Udmurt, of which a detailed introduction is given in section 2., and Lang2 is {English, Finnish, Hungarian, Russian}.

As reliable machine readable dictionaries are not available for Udmurt in sufficient size, we had to create these lexicons ourselves. The lack of parallel corpora for these language pairs makes the process challenging. We created comparable corpora for the above language pairs ranging from 96 133 tokens (Udmurt–Hungarian) to 225 914 tokens (Udmurt–English) in size.

So-called seed dictionaries play a significant role in extracting bilingual information from parallel and especially from comparable corpora. Seed dictionaries can be considered the first connecting elements between two sets of texts, allowing the extraction of parallel sentences from comparable corpora among others. Context similarity methods, the standard approach to bilingual lexicon extraction from comparable corpora (e.g. (Fung and Yee, 1998)), crucially rely on seed lexicons so the quality of these dictionaries is critical even if they are created automatically without supervision. Although bilingual dictionaries are easily accessible for widely-spoken languages (which can be used easily as a seed lexicon), it is still a challenge even to get a small set of bilingual dictionaries for endangered languages as these are rarely available in digital format and

their quality is often questionable.

Fortunately we could download dictionaries for two of the language pairs. The first step was processing these sources. Using these lexicons we could create dictionaries for Udmurt–Russian and Udmurt–Finnish language pairs. This method is discussed in section 3.

An additional source was the Wiktionary dictionary. The Wik2dict tool made us able to extract translation pairs for language pairs which are in our interest.

For creating more translation pairs from Wiktionary we used the triangulation method (Ács, 2014). This technique uses a pivot translation to get additional word pairs. “Triangulation is based on the assumptions that two expressions are likely to be translations if they are translations of the same word in a third language” – (Ács, 2014). As these word pairs were created automatically we consider the output less reliable and these pairs will be processed later with Wikipedia title dictionary.

As the first and second lexicons were made manually we considered the entries from them reliable and these were not validated by experts.

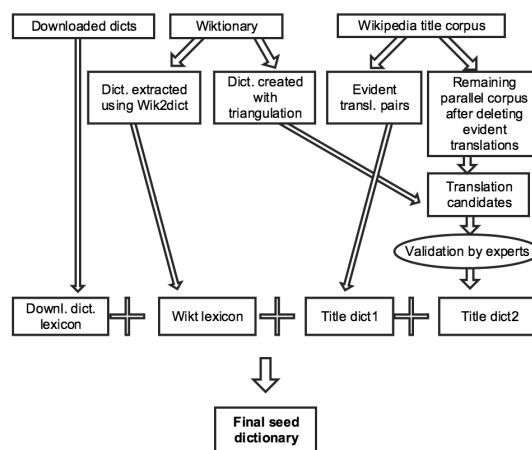


Figure 1: The process of creating seed dictionaries.

The third source of seed dictionary building was the Wikipedia article title pairs. This parallel corpus was processed in two steps. In the first step the evident translation pairs were extracted. This method resulted in another lexicon. The details of this process can be found in section 4.2. As these pairs were made by Wikipedia users we also considered the output reliable (these contains title pairs where both title are one word long or a one word long title is paired with a multi-word expression (MWE)).

After these translations were deleted from the parallel corpora the remaining pairs were processed and additional translations were extracted. These translation candidates were merged with another lexicon which was extracted using the triangulation method and these translation pairs were processed together. The result of this step was another lexicon which was validated by experts.

At the end of the dictionary building all of the created small lexicons were concatenated and this resulted one dictionary with approximately 1000 entries for each language pairs. After filtering out duplicates these dictionaries could be used as seed dictionaries.

2. The sociolinguistic situation of Udmurt

The language in centre of this paper is Udmurt, among the so-called thriving languages (Russian, English, Finnish and Hungarian) which are also mentioned. While the thriving languages are well known, Udmurt might need some introduction. Even if Udmurt is considered as the most visible and one of the bigger of the Finno-Ugric languages of the Russian Federation (Pischlöger, 2014), it is, unfortunately, still classified by the UNESCO as definitely endangered (UNESCO Atlas 2014)¹. The sociolinguistic situation of the language is clearly supporting this classification. According to the 1989 Russian Census, 747.000 people declared themselves to be of Udmurt origin and of these people circa 70% (520.000 people) speak Udmurt as their mother tongue (Winkler 2001). The 2002 Census showed a significant drop in both the number of speakers and people who identified themselves to be of Udmurt origin, 637.000 people with around 73% claimed to be able to speak the language (464.000 people) (from Perepis 2002)². The most recent Russian Census shows even more alarming numbers, only 59%, 324.000 people of the self-identified Udmurts (550.000) could speak the language to a certain degree, but not exclusively fluently. Younger people, especially in urban areas, are prone to Russification, the generation that has the most access to new technology. Udmurts living in scattered settlements usually form a majority in said communities and thus preserved their language very well, but given the location and infrastructural features of these villages, along with the demographic composition of the community (younger people tend to give up village life and move to urban areas where Russian is the language of everyday life) the speakers there are unlikely to have a significant web presence.

For Udmurt, there are prescriptive rules and a standardized orthography (Winkler, 2001), which makes it possible to publish Udmurt language materials, including mass

media (TV and radio broadcast, books, newspapers, etc.) and most importantly, from a language revitalization point of view, Web 2.0 and especially the Social Network Sites (Pischlöger, 2014) can increase the visibility of the language and provide material for research. While Social Network Sites have a more relaxed atmosphere and hardly any sign of linguistic purism, Wikipedia articles, Wikipedia being another exceptional example of a community driven Web 2.0 project, are expected to be well written, following the orthographic rules of the language.

Udmurt, being a Finno-Ugric language, is heavily agglutinating. This means that morphological analysers have to deal with rather complex constructions and while there is a well performing tool available for years³, unfortunately it is not open source. There are initiatives to create a HFST-based analyser for Udmurt, among many other Uralic languages, at Giellatekno⁴ in Tromsø, but the development of such tools is very laborious.

3. Extracting word pairs from existing lexicons

As it was mentioned in the previous paragraphs Udmurt is a severely under-resourced language. Considering this, it is no surprise that we only have Wikipedia texts as comparable corpora for the mentioned language pairs. Unfortunately we have not found any translation texts (which would be suitable for parallel corpora) in electronic form. For processing the texts of Wikipedia article pairs it is necessary to have a reliable and relatively large seed dictionary. Although there are some existing e-dictionaries, these are quite small and we decided to expand them. We also used Wiktionary entries to have more translation pairs which also resulted in a few additional dictionary entries. As a first step, we extracted translation pairs from downloaded lexicons which were in different formats. Using these resources we created additional lexicons with a few hundred entries.

Sources used for creating bilingual seed lexicons:

- Small downloaded dictionaries from the web
We could download 90 translation word pairs for Udmurt–Finnish from Goldendict⁵ and 1466 pairs for Udmurt–Russian. Another 136 translations could be downloaded from Apertium^{6,7}, another relevant site.
- Word pair extraction from Wiktionary
Using the Wikt2dict tool we extracted translation pairs for three of the language pairs.
- Extracting additional word pairs from Wiktionary using the triangulation method
The Triangulating method is also based on Wiktionary,

³<http://www.morphologic.hu/urali/>

⁴<http://giellatekno.uit.no>

⁵<http://yoshkarola.bezformata.ru/listnews/slovari-dlya-goldendict/>

⁶<https://svn.code.sf.net/p/apertium/svn/nursery/apertium-udm-rus/apertium-udm-rus.udm-rus.dix>

⁷<https://svn.code.sf.net/p/apertium/svn/incubator/apertium-fin-udm/apertium-fin-udm.fin-udm.dix>

¹<http://www.unesco.org/culture/languages-atlas/>

²<http://www.perepis2002.ru>

but it deals with extracting more translations using so-called pivot elements. Using this we were able to extract another set of translations.

Language pair (L1-L2)	E-dict	Wikt2dict	Wikt triang.
Udmurt-English	-	102	1202
Udmurt-Finnish	90	-	1213
Udmurt-Russian	1602	276	811
Udmurt-Hungarian	-	11	723

Table 1: The number of translation word pairs in the extracted lexicons

As Wikipedia texts are highly varied in their topics, utilizing a similarly comprehensive seed dictionary is essential. As it can be seen in the table above, we were able to download existing lexicons for Udmurt-Finnish and Udmurt-Russian, and extract a number of translation pairs from Wiktionary for Udmurt-English, Udmurt-Russian and Udmurt-Hungarian using the Wikt2dict tool. The other approach based on Wiktionary is the so-called triangulating method. Using this we could extract approximately a thousand of word translations.

4. Using Wikipedia title corpus to extract translation word pairs

While there are no extensive parallel corpora for language pairs formed with Udmurt, we can still find a minuscule parallel subset of the Wikipedia articles, their titles. Wikipedia title translation pairs can be easily extracted using the so-called interwiki links, or otherwise called Wikipedia interlanguage links (ILL). This resource has very valuable translation texts since these translations are manually made by Wikipedia contributors (Hara et al., 2008). Unfortunately processing them is not as obvious as it seems at first sight. While it is quite often the case that both of the titles are one word long, sometimes one of the languages appear to have a multiword expression. When titles in both languages are single words, they can be directly treated as a bilingual dictionary entry. In some cases this could be true for a number of title-based translation pairs even where we find multiword expressions, phrases or sentence fragments, for which reason we can consider a subset of the title pairs a comparable corpora.

4.1. Preprocessing title pairs

4.1.1. Language Identification

Although these title pairs are made manually by Wikipedia users or editors, allowing them to be considered a reliable and valuable source, there are some pairs which are of hardly any use when it comes to bilingual dictionary building. This is the case, for example, if the text is not in the expected language as it often happens with articles about plants and animals where one can find the scientific, latin name instead of the generally used term in a given language. Since it is quite frequent that the pair of the Udmurt title in the other language (in our case these are the English, Finnish, Hungarian and Russian titles) is the

Latin name, we decided to filter out these using a language identification tool. As expected, these language identification tools are well performing if the input is longer, but title texts have a tendency to be rather short, which causes this identification and filtering process to become more difficult and less reliable without the use of any precautionary measure. This means that if we used LANGID⁸ in a way when everything was filtered out from the corpora which were not written in the given language (according to langid) not taking into account the possibility of falsely identified texts, a remarkable number of good translation candidates were left out. Because of this reason we decided to filter out candidates where texts in L2 were written in Latin language. This technique allows the more careful, more precise filtering of titles that are of no interest.

4.1.2. Filtering Out Stopwords

For L2 titles we used stopword lists to make the output better. This was done using the stopword lists of PYTHON’s NLTK⁹ module. For Udmurt, we had to avoid using any stopword lists. Using the highest frequency words from an Udmurt Wikipedia based frequency list, the resulting output had an easily noticeable drop in quality as the list used was noisy, contaminated with strings that cannot be considered stopwords.

4.2. Extracting translation pairs where the correspondence is evident

As it was mentioned above, we considered a part of this resource as a dictionary. Following the pre-processing and modifying the corpora to be case-insensitive, the next processing step was the extraction of word pairs. Extracting the pairs where the title1 and title2 are one word long, we created a dictionary from this title corpora. If only one of the pair is one word long and its translation is longer we consider it also as a dictionary item and the longer translation is handled as an MWU.

Udmurt (L1)	English (L2)
дунай	danube
донецк	donetsk
койык	moose
тӧдьыгышлы	lily of the valley
соборной мечеть	mosque

Table 2: Examples from the lexicon

After the extraction of the dictionaries files were created containing only reliable data. After this process only longer title pairs remained in the corpus.

4.3. Extracting other word pairs from the remaining comparable corpora

4.3.1. The handling of multi-word expressions

This process is quite robust and it is based on word translation co-occurrence. The script for processing these is able to handle multi word units using an n-gram model.

⁸<https://github.com/saffsd/langid.py/tree/master/langid>

⁹<http://www.nltk.org/>

L1-L2	Whole title corp.	Extracted dict.
Udm-Eng	2701	1172
Udm-Hun	1428	589
Udm-Rus	2519	265
Udm-Fin	1663	795

Table 3: Dictionary sizes

Our observation is that the longest multi-word expression in these small corpora is three word long. So bigrams and trigrams were used in this process. This multi-word unit extraction method is quite simple. It counts how many times the bi- or trigram occurs in the text and how many times these words are found in other contexts. If it is repeated that these are occurring together, these are handled as multi-word units and marked and concatenated with underscores in the corpora.

4.3.2. Expanding the remaining title corpora with other word pairs and finding translation candidates

As the output candidates of triangulation method are not always reliable, it seems to be a reasonable idea to use these candidates with the remaining title parallel corpora helping to choose the valid translations. To make the next process easier we deleted the words which were already in the extracted corpora. If the L1 word, which is in the existing dictionary, can be found in the longer parallel L1 title text, and it is also the case with the translation word and L2 text, these are deleted. For example, if the extracted dictionary contains the pair *ӧуӧ – language*, and the remaining parallel title corpora contains entries like the pair *ӧуӧ кыӧӧ – russian language*, the output of this process will result in the pair *ӧуӧ – russian*.

After this step each L1 word in the actual entry is paired with each L2 word in the same entry. These translation candidates will be scored using a method discussed in the next paragraph.

L1 title	L2 title
каӧӧкуспо телефон код	telephone numbering plan

Table 4: An example entry

L1	L2
каӧӧкуспо	telephone
каӧӧкуспо	numbering
каӧӧкуспо	plan
телефон	telephone
телефон	numbering
телефон	plan
код	telephone
код	numbering
код	plan

Table 5: Candidates created from the previous entry

4.3.3. Calculating scores for translation candidates

Bharadwaj G., Tandon and Varma (Rohit Bharadwaj et al., 2010) used a method for calculating scores which were based on translation co-occurrences. Although the scores in our work are also based on translation co-occurrences among the candidates, there are some plus weights which make the method a bit more complex.

The created candidates are stored in a DICT TYPE in Python (a DICT TYPE is a hash-table). The keys of this dict are the Udmurt (L1) words. Each key have a list value which stores tuples¹⁰ (the tuple contains the L2 translation candidate and its actual score).

KEY: UDMWORD VALUE [(L2TRANSL, SCORE), (L2TRANSL, SCORE), ...]

The first idea is that a candidate is more likely to be reliable if it can be found multiple times in this corpus. Each time when an L1 word is paired in the corpus with an L2 translation it gets plus one score (if this translation pair has not existed it will be created). As in these language pairs it is mostly true that good translations are in the same position L2 candidates which are in the same position as the L1 candidate get another plus 1 score. It is also reasonable that if the title pairs are one word long (because other words were deleted as they existed in our previously created dictionary) it is much more probable that they are good translations. In this case they get another plus score.

4.3.4. Choosing best translations and defining a threshold of candidate scores

As we wanted to have the most reliable translations, the threshold was quite high at the beginning which resulted in a rather small output. The solution to get more good translations was not just lowering the threshold, as it resulted in the reduced quality of results. Because of this reason we decided to run the extraction and scoring method iteratively several times. First time, the threshold is rather high, resulting in only a few translations. Following this we stored these pairs in a list and deleted them from the parallel corpora as described in paragraph 4.3.2. This means that the parallel corpora gets smaller in each iteration and the list of extracted translations grows. The threshold is lowered in each iteration and as the pairs in the parallel corpora are always shorter because of the deletion of good translation word pairs (and as one to one word translations get plus scores) we will get more translations in each iteration. The script ran 9 times and the first threshold was 20 which was decreased by 2 at each iteration. At the end of the process the threshold got as low as 2. This means that if the score of the candidate was above 2 it was moved to the created dictionary.

4.3.5. Results and evaluation of the method

Using all the processes combined we managed to extract an additional lexicon.

¹⁰Tuple is a container datatype in PYTHON which is able to store two values.

L1-L2	Number of pairs	Precision
Udm-Eng	68	79,10%
Udm-Fin	45	90,90%
Udm-Hun	40	92,30%
Udm-Rus	59	63,79%

Table 6: Size and quality of the resulting dictionaries

The validation of the lexicons were done manually by experts. Since the word pairs were not lemmatized, the translations were considered good regardless of the suffixes that may have appeared on either word, meaning, for example, if the Udmurt word was in plural, but its translation was in singular, this pair was still considered valid.

5. Final size of the seed dictionaries

As the quality of downloaded dictionaries are good, we consider the outputs of Wik2dict reliable similarly to the first lexicon extracted from the Wikipedia title corpora. The only output that needed to be evaluated was the extracted translation word pairs from the remaining parallel corpora following the first lexicon extraction. After the evaluation we merged the mentioned reliable dictionaries with the evaluated new dictionary. After this step the created big dictionary could contain duplicates for avoiding this we deleted duplicated translations.

L1-L2	D.	W2D	WT1	WT2 good	C.
Udm-Eng	0	102	880	53	1034
Udm-Fin	90	0	496	40	626
Udm-Rus	1602	276	259	37	2172
Udm-Hun	0	11	497	36	543

Table 7: Final dictionaries, where D is the size of downloaded, W2D the Wik2dict, WT1 all Wikipedia titles, WT2 the validated Wikipedia titles and C is the combined, final dictionary

Using the aforementioned methods we could create seed dictionaries for all the language pairs which will allow the extraction of more translations from comparable corpora and additionally aid to parallelize these texts and create parallel corpora for further research.

6. Summary and future plans

The research presented in this paper is part of a bigger project whose aim is to support small Finno-Ugric communities in generating online content. As the role of bilingual dictionaries and parallel corpora is huge in machine translation (Bender et al., 2003), cross-language information retrieval (Grefenstette, 1998) and also language learning (Kilgarriff et al., 2013) creating these sources is a very important step in order to support the digital presence of these small languages. Since we are processing comparable corpora seed dictionaries are essential in our work. In this paper we introduced a method which enabled us to create seed dictionaries for Udmurt–English, Udmurt–Finnish, Udmurt–Hungarian and Udmurt–Russian language pairs.

In our work we used open-source software (Wik2dict, Triangulation method) and downloadable sources (Wiktionary, free bilingual dictionaries, Wikipedia) and created seed dictionaries for the mentioned language pairs which will be used for extracting parallel fragments from comparable corpora for creating parallel texts. An additional goal is to extract more translation word pairs from comparable sources in order to create large lexicons which will be uploaded to Wiktionary at the end of the project.

7. Acknowledgements

The research reported in the paper was conducted with the support of the Hungarian Scientific Research Fund (OTKA) grant #107885.

8. References

- Ács, J. (2014). Pivot-based multilingual dictionary building using wiktionary. In Chair, N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).
- Bender, O., Och, F. J., and Ney, H. (2003). Maximum entropy models for named entity recognition. In *Conference on Computational Natural Language Learning*, pages 148–152, Edmonton, Canada, May.
- Fung, P. and Yee, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1, COLING '98*, pages 414–420, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Grefenstette, G. (1998). *Cross-Language Information Retrieval*. Springer US.
- Hara, T., Erdmann, M., Nakayama, K., and Nishio, S. (2008). A bilingual dictionary extracted from the wikipedia link structure. *Database Systems for Advanced Applications*, pages 686–689.
- Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., and Volodina, E. (2013). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.
- Pischlöger, C. (2014). Udmurtness in web 2.0: Urban udmurts resisting language shift. In Hasselblatt, C. & Wagner-Nagy, B., editor, *Finnisch-Ugrische Mitteilungen*, volume 38, pages 143–161. Buske.
- Rohit Bharadwaj, G., Tandon, N., and Varma, V. (2010). An iterative approach to extract dictionaries from wikipedia for under-resourced languages.
- Winkler, E. (2001). *Udmurt*. Lincom Europa, München.

GDEX for Japanese: Automatic Extraction of Good Dictionary Example Candidates

Irena Srdanović¹, Iztok Kosem^{2,3}

¹ Juraj Dobrila University of Pula, Faculty of Humanities; I.M. Ronjgova, Pula 52100, Croatia

² Centre for Applied Linguistics, Trojina Institute, Trg republike 3, 1000 Ljubljana, Slovenia

³ Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana, Slovenia

E-mail: irena.srdanovic@gmail.com, iztok.kosem@trojina.si

Abstract

The GDEX tool, devised to assist lexicographers in identifying good dictionary examples, was initially created for the English language (Kilgarriff et al., 2008) and proved very useful in various dictionary projects (c.f. Rundell & Kilgarriff, 2011). Later on, GDEX configurations were developed for Slovene (Kosem et al., 2011, 2013) and other languages. This paper employs similar methods to design GDEX for Japanese in order to extract good example candidates from Japanese language corpora available inside the Sketch Engine system. Criteria and parameters, which were adapted to Japanese language needs, were based on the configuration for Slovene as well as the default language independent configuration available in the Sketch Engine. A number of different configurations were devised and compared in order to identify optimal values for good example identification. The paper also explores a language-learner oriented approach to good example extraction by taking into account different difficulty levels of lexemes based on the Japanese Language Proficiency Test list of words and levels. For this purposes, additional configurations were devised, which are tailored to individual levels and thus useful for language learners and lexicographers of learner's dictionaries.

Keywords: dictionary examples, corpus lexicography, Japanese language, difficulty levels, automatic extraction

1. Introduction

Examples are a crucial part of a dictionary since they illustrate to a user how a word is used in a particular meaning, pattern or situation. They are an additional support to definitions (Atkins & Rundell, 2008), and enable a language learner to remember and understand a new word and its common and correct usage more easily. Collecting good examples, which should be natural, typical, informative and intelligible to learners (Atkins & Rundell, 2008), requires a significant effort and cost in lexicographic projects. The development of new technologies such as mobile telephones and the Internet resulted in an even greater demand for up-to-date, concise and instant information by dictionary users. At the same time, recent technologies have also provided lexicographers with necessary data, means and tools for addressing the new needs in the dictionary-making process. One such important tool is GDEX (Good Dictionary EXamples; Kilgarriff et al., 2008), which is part of the Sketch Engine corpus query system (Kilgarriff et al., 2004) and aims to assist lexicographers with identification of good dictionary example candidates.

This paper describes the development of GDEX for Japanese with the aim to facilitate the identification and extraction of good example candidates from Japanese language corpora. The process of development of GDEX configuration for the Japanese language included employing methods already used in the development of GDEX configurations for other languages (c. f. Rundell and Kilgarriff, 2011; Kosem et al. 2011; Kosem et al., 2013; Kosem 2015). The paper also explores a language-learner oriented approach to good example extraction by taking into account different difficulty levels of lexemes based on the Japanese Language Proficiency Test list of words and

levels, which makes the results useful for language learners and learner's dictionaries as well. Similar approach has already been used by Hmeljak Sangawa et al. (2009) whose aim was to devise a corpus-based example resource for language learners. They used JLTP data to annotate a 100-million-word sample of the JpWaC web corpus with an additional information on difficulty levels, and then automatically extracted individual sentences to create a monolingual corpus of example sentences. In contrast, our research aims at designing GDEX for Japanese with a wider application, not only for language learning but mainly for Japanese language lexicography. Therefore, we already apply majority of important filtering in the general configuration. Moreover, we utilize some of the advantages of the newer morphological annotation tool set (MeCab and UniDic), and the latest GDEX configuration methodology, used for certain other languages, which for example identify typical collocational relations in example candidates.

2. GDEX (Good Dictionary EXamples) and Its Functionalities

GDEX is a tool in the Sketch Engine designed to help the lexicographers and other potential users identify dictionary examples by ranking sentences according to how likely they are to be good example candidates. Thus, the tool is very useful in helping to avoid time-consuming searching of good examples in hundreds or thousands corpus sentences.

The ranking of example sentences is done automatically using various syntactic and lexical features specified in a configuration file. As mentioned in existing research that developed GDEX configurations (e.g. Kosem et al. 2011), these features (classifiers in the configurations) often include sentence length, word length, presence of absence

of certain words (e.g. low frequency words, polysemous words) and/or characters and symbols, whole sentence form, position of keyword in the sentence etc. Configuration classifiers and their parameters can be included/excluded, adjusted or combined according to the characteristics of a particular language or specifics of the intended GDEX use.

GDEX can be used in the Sketch Engine in two ways: in Concordance¹ or via the TickBox Lexicography (TBL) function in Word Sketch. TBL provides clickable boxes next to each collocate, which are then used to export wanted collocates and selected number of their ranked examples. A recently added and very helpful feature of Concordance view is the option to show GDEX score for individual concordances. Similarly, TBL view is useful not only because it provides only the first X example candidates (default setting is 6), but as it enables comparison of outputs of two different configurations (see Figure 2). It is worth noting that in TBL, GDEX first randomly selects 300 corpus sentences for each collocation and then ranks them; this is necessary as processing and ranking a large number of concordances can take a long time.

The GDEX tool was initially created for the English language (Kilgarriff et al., 2008) and proved to be very useful for lexicographers in various dictionary projects (c.f. Rundell and Kilgarriff, 2011). Later on, it was created for Slovene (Kosem et al., 2011; Kosem et al., 2013; Gantar et al., in print) and other languages. An overview of configurations for different languages shows that many classifiers are used almost in every configuration, only the values are adapted to the language or corpus size. Yet, there are classifiers found only in certain configurations, e.g. keyword position in the sentence. This classifier was used in the first version of GDEX for English, then also in GDEX for Slovene (Kosem et al., 2011), while in the improved GDEX for Slovene (Kosem et al. 2013; Gantar et al. in print), used for automatic extraction, the classifier was kept only in configuration for verbs.

The Sketch Engine also provides the default GDEX configuration that is supposed to be language-independent. In essence, it is a simplified version of GDEX for English, devised to fit various languages. It contains only three absolute classifiers (whole sentence, blacklist for illegal characters, and minimum token frequency) and three penalty classifiers (optimal sentence length, penalty for rare words, and penalty for rare characters). Clearly, such configuration cannot meet the needs of lexicographers for a particular language, but can be useful for certain general purposes or as a point of departure for developers of GDEX configurations.

3. Designing GDEX for Japanese

3.1 Japanese Language Resources Used

The Sketch Engine is hosting various Japanese language

corpora and each of these corpora could be used with the GDEX configuration to withdraw examples from it. The most preferable for lexicographic use is the large-scale web corpus JpTenTen11 [SUW] of 10 billion tokens (Pomikálek & Suchomel, 2012; Srdanović et al., 2013), because of its size, a consistent morphological annotation in short unit words (SUW), and carefully planned design methods. A sample corpus of almost three hundred million tokens, called JpTenTen11 [SUW, sample], was built from it, which is more suitable for the development of GDEX configurations due to its smaller size². The same corpus was sampled and annotated using long unit word annotation (LUW), which could be, in practice, a very good resource for lexicographic purposes, since it avoids splitting a word into morphemes. However, the LUW version of corpus is still not a preferable solution for lexicographers, at least for describing middle and low frequency words, due to its limited size. Therefore, we decided to develop the GDEX for Japanese for use with corpora with the SUW annotation. The aforementioned corpora use the morphological annotator MeCab with the UniDic electronic dictionary, which are also used for BCCWJ (Balanced Corpus of Contemporary Written Japanese; cf. Maekawa et al., 2013).

Each of these corpora, using a sketch grammar for Japanese (c.f. Srdanović et al. 2008, 2013), provide a lexicogrammatical profile of words and their collocates summarizing their usage within the Word Sketch functionality.

Finally, the SkE system enables building new corpora or adding other corpora to the platform, after which, preferably with some adjustments, each of the corpora can use the GDEX configuration.

3.2 Sample List of Words

For the purposes of the evaluation of test GDEX configurations for Japanese, we used a randomly extracted list of lemmas (Kilgarriff et al., 2010; Srdanović et al., 2011), shown in Table 1. The list was created by taking a sample from the 30,000 commonest nouns, verbs and adjectives in the Japanese web corpus JpWaC, in a ratio of roughly 2:1:1, where a number of lemmas were selected for high (top 2999 words), mid (3000-9999) and low (10,000-30,000) frequency groups. Within these constraints, the sampling was random, but for the purposes of GDEX testing, we selected 5 nouns, 4 verbs and 4 adjectives (for high frequency group 3 nouns, 2 verbs, and 2 adjectives, and for the mid and low frequency groups one lemma of each word class per frequency group).

The testing of configurations was conducted both in the TBL output in the Word Sketch, and in the Concordance feature. Namely, in Concordance, one can also observe GDEX scores of individual corpus sentences, making it easier to monitor the influence of individual parameters, or to validate the correctness of their syntax for that matter.

¹ The setting can be activated in View Options.

² Testing GDEX configurations on corpora that are over a billion

words in size can be very time-consuming as sometimes the results takes several minutes to load.

	Sample lists		
	Nouns	Verbs	Adjectives
High	急 <i>kyuu</i> 'sudden' 研究 <i>kenkyuu</i> 'research' 完成 <i>kansei</i> 'completion'	生まれる <i>umareru</i> 'to be born' 扱う <i>atsukau</i> 'to treat'	よろしい <i>yoroshii</i> 'good/fine' 素晴らしい <i>subarashii</i> 'great'
Mid	欠席 <i>kesseki</i> 'absence'	資する <i>shi suru</i> 'contribute'	黒い <i>kuroi</i> 'black'
Low	方角 <i>hougaku</i> 'direction'	駆け込む <i>kakekomu</i> 'rush into'	腹立たしい <i>haradatashii</i> 'irritating'

Table 1: Random lemmas used for testing various GDEX configurations for Japanese.

3.3 Preparation Steps and Findings

3.3.1 Japanese with Language Independent GDEX

The first step was to observe behaviour of corpus example candidates offered by the currently available GDEX configuration file, which is language independent (default GDEX). For each word from the sample list, the first 20 GDEX examples in the jpTenTen11 [SUW, sample] corpus were examined.³



Figure 1: Japanese examples using default GDEX configuration.

Figure 1 shows a sample of Japanese corpus sentences evaluated using default GDEX. The evaluation revealed several potential issues of sentences offered by the default configuration file when applied to Japanese data, which resulted in the following initial guidelines for the first version of GDEX for Japanese: a) to exclude sentences that do not have sentence final punctuation mark in the end, b) to penalize examples with numbers, c) to exclude sentences with an open or close bracket only, d) to exclude sentences

³ The following options were used in the Concordance window to be able to extract the best set of GDEX examples: Sort good dictionary examples, Show GDEX score in concordance, Allow multiple line selection.

⁴ The latest version of GDEX for Slovene was developed for the

with \cdot , $!$, — and other similar symbols, e) to reduce sentences containing letters of the Latin alphabet, f) to penalize sentences that contain various types of brackets, such as $()$, $【】$, g) to penalize sentences ending with $は$, $こと$ に $。$, $、$ etc. for incompleteness, h) to penalize sentences that start with various conjunctions and anaphoric expressions, such as $または$, $そして$, $そんな$ etc., due to potential vagueness.

3.3.1 Japanese with Language Independent vs. Slovene GDEX

In order to identify other potential classifiers and settings of the first version of GDEX for Japanese, we compared the results of default GDEX configuration and the latest version of GDEX for Slovene (Kosem et al., 2013) on Japanese corpus data.⁴ We used Tickbox Lexicography and a side-by-side comparison of examples offered by two different GDEX configurations (see Figure 2).

The length of candidate examples proved to be the most significant difference between the two configurations, and also one of the biggest shortcomings of both configurations. Default GDEX configuration offered very short corpus sentences, and in some cases, only a part of a sentence appeared as an example. On the other hand, GDEX for Slovene offered much longer corpus sentences and there were even cases when candidate examples consisted of more than two sentences. In addition, several categories used in the Slovene configuration file (e.g. the classifier penalizing pronouns relies on tags from the Slovene tagset) needed an adjustment to the Japanese language in order to be of any use.

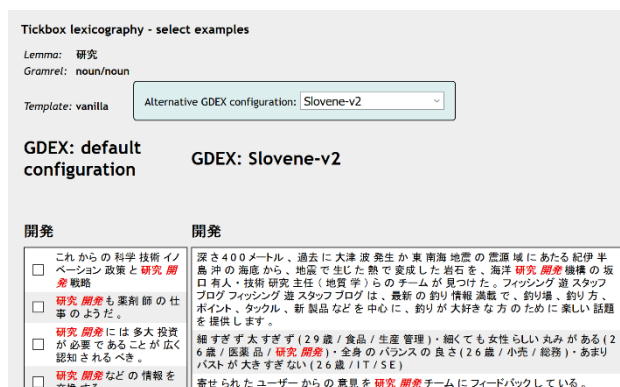


Figure 2: Comparison of two configurations, default and Slovene, in Tickbox Lexicography, when applied to Japanese corpora.

3.4 Development of GDEX for Japanese

The development of GDEX for Japanese aims to respond to two needs: a general lexicographic need for identifying

purposes of automatic extraction; in fact, four different configurations were developed, one per word class (noun, verb, adjective, adverb). For the purpose of this research, we used the configuration for verbs, since it contains the largest set of useful classifiers.

good examples in Japanese language corpora, and the use of GDEX examples for language learning purposes and compilation of learner's dictionaries. This section describes the development of GDEX configuration for Japanese that aims to respond to the first need. The next section presents GDEX for Japanese aimed especially at language teaching and learning situation (Learner's GDEX for Japanese henceforth).

GDEX for Japanese was devised in a number of stages that were circularly repeated until satisfactory results were achieved:

1. Selecting various classifiers (e.g. blacklists for signs and letters, allowed sentence length, preferred sentence length, minimum token frequency, Japanese-specific symbols and other specified items, etc.).
2. Quantitatively determining the values of classifiers and their weight. Using weights, classifiers were divided into absolute, which penalized candidate sentences so heavily they were ranked near the bottom of the list, and penalty classifiers with various degree of penalty (or bonus points). Also, nearly absolute classifiers were group together, so sentence meeting the criteria of only one of the listed classifiers, already incurred the penalty.
3. Evaluating the configuration in the word sketches of selected sample lemmas.
4. Devising an improved configuration.
5. Evaluating the two configurations by comparing their results.
6. Devising the next improved configuration based on the findings, evaluating and comparing the results of the last configuration and the newly devised one, and so on.

The final GDEX for Japanese configuration consists of the following major classifiers:

- Mandatory features of a candidate example: the Japanese full stop 。 or a question mark ? , which need to appear once only (one of them, not both) in a corpus sentence to avoid corpus hits that contain only a part of a sentence on the one hand, and to avoid corpus hits that include more than one sentence on the other.
- Defining the preferred and allowed sentence length. The initial preferred sentence length was from 10 to 25 tokens (penalty classifier), while allowed sentence length (absolute classifier) was set to between 8 and 30 tokens.⁵
- Penalizing symbol signs and Latin characters to avoid corpus sentences with noise and inappropriate form. When creating a blacklist of symbols, we took an advantage of the narrow MeCab tagset used in combination with the UniDic electronic dictionary and SUW annotation and its fine annotation of various symbols and

their types (e.g. Sym.ch, Supsym.g, Supsym.aa.e, and Supsym.aa.g). In addition, the blacklist specified some additional illegal characters, as well as spam characters and strings.

- Penalizing rare characters including certain types of brackets in Japanese. The web data with brackets appear to be less appropriate for good dictionary examples, as they contain unnecessary information.
- Excluding all types of brackets (absolute classifier) that appear as only open or only closed in the candidate sentence examples.
- Penalizing words that are longer than 7 characters. Most typically, Japanese words consists of one, two, three or four characters. Longer words are mainly borrowed words, which are written in syllabic script katakana. We intend to further explore this classifier in the future.
- Penalizing sentences containing proper nouns (names, surnames, geographical names). Here again, MeCab tagset for proper nouns is used: N.prop.g, N.prop.n.g, N.prop.n.f, N.prop.n.s, N.prop.p.g, N.prop.p.c. Penalizing sentences for every lemma with frequency below 10,000 in the sample corpus. This value needs to be adapted according to the size of the corpus.
- Rewarding sentences containing top ten collocates of the collocation (a classifier for a second collocate, see Kosem, 2015). The purpose of this classifier is to give priority to sentences containing typical patterns of a particular collocation and thus obtain more typical examples.

During the process of configuration and evaluation, altogether 20 different configurations were devised and tested. The initial configurations focussed on adding new classifiers, and the later configurations on fine-tuning classifier settings, i.e. parameter values and weights.

3.5 Evaluation of GDEX for Japanese

Evaluation of GDEX for Japanese was part of the configuration development process since each time the configuration was improved and thus a new version devised, a comparison of previous and new configurations was conducted. Here we summarize some of the findings. Initially, after inspecting the tag for period punctuations Supsym.p, the configurations included several types of possible sentence-ending punctuation, such as . and . , as well as exclamations. The evaluation of the results showed that exclamations needed to be removed (! , !), as often the corpus sentences containing them were too marked. Also, the period punctuations other than *kuten* brought up examples with dates and years, as well as ordered lists, which is why we decided to remove them as much as possible.

Evaluation of the candidate sentences with brackets showed that brackets used for citations do not need to be

⁵ This category is tuned for SUW annotation and is needs to be

adjusted if used on the LUW annotated data.

penalized unless they contain only closed or only open bracket. Here, Japanese tagset was helpful in specifying classifier parameters as it has different tag types for different types of the brackets: Supsym.bo for opening brackets and Supsym.bc for closing brackets.

Figure 3 shows the final version of GDEX for Japanese (Japanese-v1u) compared to default GDEX configuration. The collocation *jouhou wo atsukau* ('to deal with an information') of the verb *atsukau* ('to treat') is examined. The first two candidate examples offered by default GDEX contain an exclamation and are too marked, the fourth example is too long, consists of two sentences and contains obstructive elements in the form of Latin strings. The eighth example is not complete and has noise at the beginning. The ninth example has an open bracket, but not a closing one, and is also incomplete. The candidate examples of GDEX for Japanese are all well-formed, informative and clear. Nonetheless, even more clarity could be achieved with avoiding context dependant conjunctions at the sentence beginning. The typical collocation 個人情報 *kojin jouhou* ('personal information') appears in three out of ten examples, which well grasps typical usage but needs to be further examined for diversity.

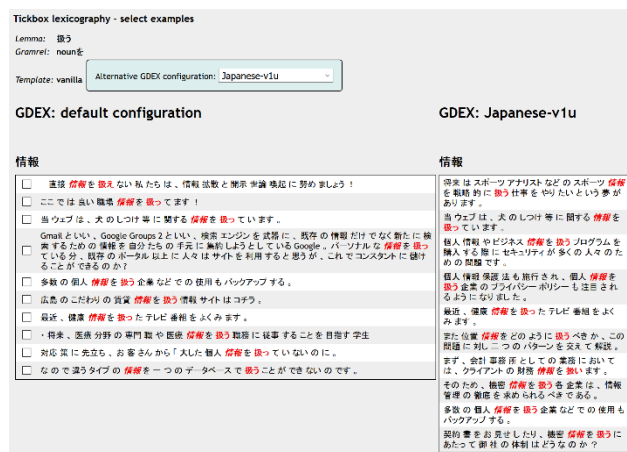


Figure 3: Comparing default GDEX and GDEX for Japanese configurations.

One of the issues we want to explore further is the influence of numbers in candidate example sentences, i.e. their suitability for dictionaries. The Japanese tagset provides a tag for numbers N.num, which is often accompanied with nominal counters annotated as nouns or as suffixes (N.c.count, Suff.n.g, Suff.n.count). Namely, we noticed several examples containing numbers to be less useful. However, classifier for penalizing numbers needs to be designed carefully to avoid penalizing perfectly good candidate sentences (e.g. ones where numbers denote a year).

For the sake of clarity the candidate examples should not require any additional context, a list of words or strings of

words that appear at the beginning and at the end of Japanese sentences needs to be compiled and implemented into the configuration.⁶ As observed, this will also to some extent help in avoiding too informal sentences.

4. Learner's GDEX configuration(s) for Japanese

4.1 JLPT Levels

The Japanese language proficiency test (JLPT) word list is the most widely used vocabulary list for Japanese as a foreign language. The list is used as a standard in creation of Japanese language textbooks as well as tests for measuring the proficiency level of learners. The old version of the list divides the Japanese vocabulary into 4 levels, where 1 is the most difficult and 4 is the least difficult level (Japan Foundation and Association of International Education Japan, 2004). The newest version of the word list uses 5 levels, 5 being the least difficult level. Since there is no official new vocabulary list available, for the purposes of this research, we used the list created by individuals based on the available resources for the old and new JLPT exam and a word frequency analysis to create a most plausible list for the new 5-level JLPT.⁷

4.2 Development of Learner's GDEX for Japanese

Difficulty-level tailored GDEX for Japanese, or Learner's GDEX for Japanese, consists of a number of configurations, with each configuration adjusted for a particular difficulty level. The differences between the configurations are thus mainly in penalising words and their lemmas that appear in the corpus but are more difficult than the specified level, and awarding bonus for sentences containing a certain percentage of words for a particular level. So the final list of the configurations is as follows:

- Japanese-v1v-jlpt1-5 (the most difficult level, penalising words outside the whole JLPT list, preference for all the words listed in the list from levels 1 to 5)
- Japanese-v1v-jlpt2-5 (penalising words outside the levels 2, 3, 4 and 5, preference for words listed in the JLPT levels 2 to 5)
- Japanese-v1v-jlpt3-5 (penalising words outside the level 3, 4 and 5, preference for words listed in the JLPT level 3 to 5)
- Japanese-v1v-jlpt4-5 (penalising words outside the level 4 and 5, preference for words listed in the JLPT level 4 and 5)
- Japanese-v1v-jlpt5 (the easiest level, penalising words outside the level 5, preference for words listed in the JLPT level 5).

For the purposes of configuration classifier related to word lists, we merged items from various levels, depending on the level for which the configuration was devised. Since

⁶ So far, the following list is prepared for the beginning of a sentence: また|それ|これ|あれ|で|しかし|と|つまり|ただ|なぜ|で|する|って|しかも|こう|そう|その|この|あの|とりあえず|さあ, and the following list for the end of a sentence: に|は|な|

⁷ For example, refer to <http://www.tanos.co.uk/jlpt/> and www.jlptstudy.net/.

some words at the beginner's levels (4 and 5) are written in the syllabic script hiragana instead of with kanji characters (kanji characters are preferred options in lemma annotation), certain mismatch of the items was possible due to differences in the used script. To avoid these mismatches, we manually rechecked the word lists of those two levels and prepared the items in an appropriate script. In addition, some items in the word lists appear as multiword units, which is not in compliance with the narrow annotation used for corpora, so some readjustment was done to cover the items from the list properly. For example, the word お父さん *otousan* ('father' - polite form/used when talking about fathers of other persons, and for addressing own father), consists of three elements in the corpus annotation:

- 御 *o* (prefix for politeness – annotated as lemma in the corpus using the preferred kanji character script, but usually written with the syllabic script hiragana お *o*, and as such appears in the JLPT list).
- 父 *ou* ('father' - written in kanji characters) and
- さん *san* (polite suffix for persons - written in the syllabic hiragana).

Although the JLPT list provides this word as one unit, it is not tagged as such in the corpus, so we considered that each of the three elements should appear separately in the list of items used in the configuration file. Similar approach was used for other multi-element units. The advantage of using a list of lemmas in the configuration was to cover script variations of words that were annotated in the corpus under one lemma (for example, the combination of lemmas 御+父+さん covers more script variations, e.g. 御父さん, お父さん, おとうさん). The other advantage was to cover various grammatical forms of a word that were annotated under one lemma and are not specifically listed in the JLPT list although they appear in different forms in the real data – this is mainly valid for verbs and adjectives as they have inflection in Japanese.

After the lists were prepared, the penalization and preference parameters were adjusted in the configurations according to the difficulty levels, and to avoid vocabulary outside of a particular learning level as much as possible. Some fine-tuning of the settings was needed, including of other classifiers, for example threshold for less frequent lemmas and sentence length. This provided us with the first versions of Learner's GDEX for Japanese (one configurations per level; for level 5, we devised three different configurations), which we then evaluated.

4.3. Evaluation of Learner's GDEX for Japanese

For the evaluation of the Learner's GDEX for Japanese configurations, we used one word per difficulty level:⁸ 黒い *kuroi* ('black' [adjective, JLPT5]), 研究 *kenkyuu* ('research' [noun, JLPT4]), 扱う *atsukau* ('treat' [verb, JLPT3]), 思い込む *omoikomu* ('to be convinced/under impression' [verb, JLPT2]), 資産 *shisan* ('assets' [noun, JLPT1]). Then, we searched for collocations of the words

using the Word Sketch functionality and the option Tickbox Lexicography. We chose collocates of the same or lower level than the keyword and compared sentences offered by the configurations. In addition, we used the Reading Tutor's Vocabulary functionality⁹ to examine difficulty levels of the offered candidate examples.

Table 2 shows the evaluation results for items in the configuration files for various levels before the final fine-tuning was done. The findings confirmed that the corpus sentences offered become more difficult with each difficulty level, and that overall they are less demanding for learners than the sentence candidates provided by GDEX for Japanese. It was noticed that overall difficulty depends on the collocation relation as well; for example, candidate examples with N+N collocations seem to be more complex and demanding for learners than other combinations (c.f. *kenkyuu* 'research'). However, the analysis confirmed that we need stricter restrictions for words from the levels of difficulty higher than the level of target configuration, which was then tested for level 5, devising the improved configurations 5a and 5b.

Keyword	Collocate	L5	L3-5	L1-5
黒い(5) <i>kuroi</i> 'black'	物(5) <i>mono</i> 'thing'	**	***	***
	色(5) <i>iro</i> 'color'	**	***	***
	髪(5) <i>kami</i> 'hair'	**	***	**
	部分(3) <i>bubun</i> 'part'	***	***	***
研究(4) <i>kenkyuu</i> 'research'	計画(4) <i>keikaku</i> 'plan'	****	****	****
	センター(3) <i>sentaa</i> 'center'	*****	*****	*****
	会(1) <i>kai</i> 'meeting'	****	****	****
	開発(1) <i>kaihatsu</i> 'development'	*****	*****	*****
扱う(3) <i>atsukau</i> 'to treat'	物(5) <i>mono</i> 'thing'	***	***	***
	問題(5) <i>mondai</i> 'problem'	****	****	****
	商品(3) <i>shouhin</i> 'product'	***	****	****
	情報(3) <i>jouhou</i> 'information'	****	****	****
資産(1) <i>shisan</i> 'properties'	いる(5) <i>iru</i> 'to be'	***	***	***
	持つ(5) <i>motsu</i> 'to bring'	****	****	****
	する(5) <i>suru</i> 'to do'	***	***	****
	増やす(2) <i>fuyasu</i> 'to increase'	****	****	****

Table 2: Difficulty level of examples per collocation offered by configurations L5, L3-5 and L1-5 as measured by Reading Tutor: * Easy, ** A bit easy, *** Normal, **** A bit difficult, ***** Difficult.

Figure 4 shows comparison of the configurations 5a and 5b for JLPT level 5 for the keyword 黒い *kuroi* 'black' [adjective, level 5] and its collocate 服 *fuku* 'clothes' [noun, level 5]. Sentences offered by configuration 5b are

⁸ We used the sample list (see Table 1). Since none of the sampled words belonged to level 2 and 1, we searched for related words (consisting of the same element) in the JLPT list and chose 資産

instead of 資, and 思い込む instead of 駆け込む.

⁹ Reading Tutor's home page: http://language.tiu.ac.jp/index_e.html

improved and easier for learners, if compared to sentences offered by configuration 5a, since the sentences with more difficult words and those words not present in JLPT that appear with configuration 5a are not found among the first six sentences in configuration 5b. This is supported by the scores of the Reading Tutor tool as candidate examples offered by the configuration 5b are evaluated as Very easy (*), while the examples in offered by 5a, are evaluated as Normal (***) and contain more difficult words.

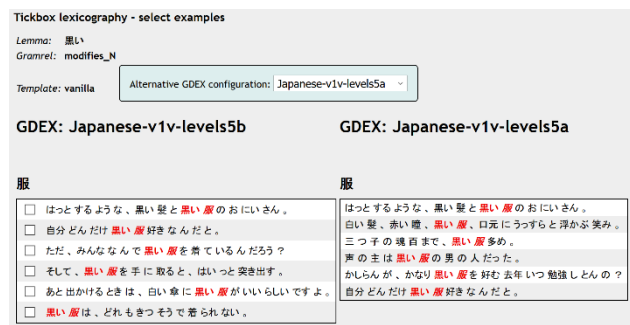


Figure 4: Comparing difficulty-level tailored configurations for JLPT level 5 (configurations 5a and 5b).

Finally, there is still some room for improvement. Besides further evaluation and fine-tuning of the configurations, there is still some work to be done in relation to the adjustment of difficulty levels of kanji characters, as well as difficulty levels of various grammatical patterns. Also, the morphological analysis of tokens and lemmas in the corpus sometimes does not fully correspond to word forms and the script used in the JLPT list – thus, some additional work is needed to further clean the data. In future research, we plan to use additional resources, such as new JLPT textbooks, to improve the five level JLPT list. Also, the use of some other resources that measure language difficulty for foreign learners will be considered, e.g. Instructional Vocabulary List with around 18 thousand words and 6 different levels (Sunakawa et al., 2012).

5. Conclusion

This paper presents the design and preliminary evaluation results of two GDEX configurations for Japanese. The first configuration, GDEX for Japanese, is aimed at the needs of corpus-based lexicography. It was created by adjusting to Japanese various criteria and parameters used by configurations for other languages as well as default language independent configuration in the Sketch Engine tool. The second configuration or group of configurations, Learner’s GDEX for Japanese, are difficulty-level tailored GDEX configurations that use the JLPT word lists to provide more learner-friendly set of example candidates and can thus be useful for Japanese language learners and makers of learner’s dictionaries.

Some of the most important achievements of the GDEX configuration designed specifically for Japanese are as follows. Firstly, configurations offer example candidates that have full sentence form (instead of fragments, or two

or more sentences), which above all contributes to their suitability and informativeness. Rewarding second collocates, together with penalizing less frequent words and non-Japanese characters, help in identifying examples that are more typical of a collocation. Clarity of example candidates is achieved by penalizing unnecessary or potentially distractive elements such as proper nouns, symbols, brackets etc. Both configurations make use of some of the advantages of the newer morphological annotation tool set (MeCab and UniDic), such as covering various script variants under one lemma and more fine-grained set of part of speech categories with precisely annotated various types of symbols, punctuation marks, character strings, and numbers.

Learner’s GDEX for Japanese configurations, which implemented data on JLPT difficulty levels, include the method of rewarding group of words that belong to a particular difficulty level, and penalizing words outside that level. The analysis of the results showed that a very high percentage of retrieved examples were grammatically well formed and acceptable. Overall, the first versions of the difficulty-level tailored Learner’s GDEX for Japanese offer more learner-friendly examples than GDEX for Japanese. The difficult of offered examples increases with the difficult of levels targeted by the different configurations. Nonetheless, more fine-tuning of configurations is needed in terms of rewarding words that belong to a particular difficulty level or penalising words outside that level, in order to identify the examples that are suitable for the level in question.

6. Bibliographical References

- Atkins, B.T.S, Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Gantar, P., Kosem, I., Krek, S. (in print). Discovering automated lexicography: the case of the Slovene Lexical Database. *International Journal of Lexicography*. Advance Access: <http://ijl.oxfordjournals.org/content/early/2016/03/27/ijl.ecw014.abstract>
- Hmeljak Sangawa, K., Erjavec, T., Kawamura, Y. (2009). Automated collection of Japanese word usage examples from a parallel and a monolingual corpus. In S. Granger & M. Paquot (Eds.), *eLexicography in the 21st century: new challenges, new applications: Proceedings of eLex 2009*. Louvain: Presses Universitaires de Louvain, pp. 137--147.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the Thirteenth EURALEX International Congress*. Barcelona, Spain: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 425--432.
- Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I. & Tiberius, C. (2010). A Quantitative Evaluation of Word Sketches. In *Proceedings of the XIV Euralex International Congress*. Leeuwarden: Fryske Academy, 7pp.

http://nlp.fi.muni.cz/publications/kilgarriff_xkovar3_et_al/kilgarriff_xkovar3_et.al.pdf

- Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. *Proc. EURALEX*, Lorient, France.
- Kosem, I., Gantar, P., Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (Eds.), *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17–19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 32--48.
- Kosem, I., Husak, M., McCarthy, D. (2011). GDEX for Slovene. In I. Kosem & K. Kosem (Eds.), *Proceedings of eLex 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 151--158.
- Kosem, I. (2015). Interrogating a Corpus. *The Oxford Handbook of Lexicography*, pp.76--93.
- Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., Den, Y. (2013). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*. Netherlands: Springer.
- Pomikálek, J., Suchomel, V. (2012). Efficient web crawling for large text corpora. Kilgarriff, A. & Sharoff, S. (eds.) *Proceedings of the Seventh Web as Corpus Workshop (WAC7)*. Lyon, 2012.
- Rundell, M., Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds.), *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Amsterdam: John Benjamins, pp. 257--281.
- Sunakawa, Y., Lee, J.-H., Takahara, M. (2012). The Construction of a Database to Support the Compilation of Japanese Learners' Dictionaries, *Acta Linguistica Asiatica*, 2(2), pp. 97--115.
- Srdanović, I., Erjavec, T., Kilgarriff, A. (2008). A web corpus and word-sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing)*, 15(2), pp. 137--159.
- Srdanović, I., Ida, N., Shigemori Bučar, C., Kilgarriff, A. & Kovar, V. (2011). Japanese word sketches: advantages and problems. *Acta Linguistica Asiatica*, 1(2), 63-82.
- Srdanović, I., Suchomel, V., Ogiso, T., Kilgarriff, A. (2013). Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen. In *Proceeding of the 3rd Japanese corpus linguistics workshop*, Department of Corpus Studies/Center for Corpus Development, NINJAL, pp. 229--238.

Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case

Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, Guadalupe Aguado-de-Cea

Ontology Engineering Group
Universidad Politécnica de Madrid
{jbosque,jgracia,emontiel,lupe}@fi.upm.es

Abstract

Lexicographic resources can highly benefit from Semantic Web technologies, specifically, linked data technologies, since such resources cannot only become easy to access and query, but also easy to share and link to resources that contain complementary information, contributing to the creation of a huge graph of interlinked lexical and linguistic resources. In this paper, we present the methodology we have followed for the transformation of a lexicographic resource, namely, the Spanish dataset of K Dictionaries's Global series, from its proprietary XML format to RDF, according to the *lemon-ontolex* model, a *de-facto* standard for representing lexical information in the Web of Data. We describe in detail the original resource, the design decisions taken for the transformation process, the model chosen for the representation of the dataset, as well as the extensions made to the model to accommodate specific modelling needs of the original source. The core of the representation model is described in detail in order to illustrate the issues encountered and how they have been solved in this first prototype, which could serve to lay the foundations for future transformations.

Keywords: Linked Data, e-lexicography, lemon-ontolex, multilingual dictionary.

1. Introduction

Recently, the field of lexicography has experienced a remarkable evolution marked by the adoption of language technologies to assist content creators in their job and to make dictionary data more easily accessible to experts and final users (Fuertes-Olivera and Bergenholtz, 2011; Moulin and Nyhan, 2014). Integration of lexicographic systems into bigger knowledge systems and the ability to exploit them in collaborative contexts is also essential for modern lexicography.

In this context, we argue that linked data technologies (Bizer et al., 2009) constitute a major opportunity for representing, sharing, interlinking, and accessing lexicographic information at a Web scale by following Semantic Web standards. In short, linked data refers to a set of best practices for exposing, sharing and connecting data on the Web. By following these practices, data is shared in a way that can be read automatically by computers. Linked data relies on the Resource Description Format (RDF) (Manola and Miller, 2004) as main mechanism to describe the data.

In fact, we are nowadays witnessing a growing trend in publishing not only lexicographic data but any type of linguistic data and language resources (lexicons, corpora, dictionaries, etc.) as linked data on the Web. As a result, the so-called linguistic linked open data (LLOD) cloud¹ is emerging. This cloud is constituted by the RDF version of such language resources, whose data are linked to one another. We are using here (and all throughout this paper) the term “linking” in the “linked data” sense, that is, a resource on the Web (e.g., a person, an image, a lexical entry, a definition, ..., which is uniquely identified at a Web scale) can be linked to another resource in such a way that useful information can be stated about the former. The set of all links constitutes a graph that can be traversed and queried in a straightforward and standardised manner. This goes well

beyond the notion of cross-reference, i.e., the exact location in a dictionary where a lexical entry can be found.

In this work, we report on our experience in modelling the linked data version of a very rich lexicographical resource, namely, the Spanish dictionary core of the K Dictionaries (KD) multilingual Global series dataset. We present the methodology we have followed for the transformation of the resource from its proprietary XML format to RDF, according to the *lemon-ontolex* model², a *de-facto* standard for representing lexical and linguistic information on the Web of Data. We focus here on the modelling part, leaving aside the issues related to the generation and publication of the RDF data.

The rest of this paper is organised as follows. In Section 2, we describe the methodology followed in the definition of the model. In Section 3 the analysis of the source data is described, while the model used to represent such data in RDF is presented in Section 4. Then, in Section 5, the application of the selected model to the particular characteristics of the KD data is discussed. Some related work is introduced in Section 6, and, finally, the paper is concluded in Section 7.

2. Methodology

In this section we summarise the methodology that we have followed to define the RDF version of the dictionary data from its XML counterpart, as well as the main design decisions we took prior to starting the process. As basis for the modelling we relied on existing guidelines for multilingual linked data generation such as the ones described by Vila-Suero et al. (2014) as well as the guidelines for linked data generation of language resources developed in the context of the W3C Best Practises for Multilingual Linked Open

¹<http://linguistic-lod.org/llod-cloud>

²https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

Data (BPMLOD) community group³.

2.1 Tasks

The applied process can be divided into these steps:

1. *Analysis of the data* (Section 3). The original KD data were analysed in detail, based on its document type definition (DTD) and a representative sample of XML entries from the Spanish dataset. Datasets of two other relevant languages (English and German) were explored as well.
2. *Model definition* (Section 4). The last version of the W3C *lemon-ontolex* model was considered to cover the representation needs of the KD data. Other linguistic models (e.g. LexInfo) were also taken into account. Ad-hoc solutions or extensions to the *lemon-ontolex* ontology have been proposed in cases in which the representation needs were not met. A suitable strategy was devised to construct permanent Uniform Resource Identifiers (URIs) that identify the elements in the model.
3. *Conceptual mappings definition*. A set of mappings was defined in a platform independent format. These mappings represent the links between the entities in the original model (XML-based) and their counterpart in RDF.

According to the above referred guidelines (Vila-Suero et al., 2014), these tasks should be followed by two additional steps, namely, the RDF generation and RDF publication processes, which are not covered in this paper (focused on the modelling process). In such steps a set of conversion scripts should be generated and run to generate the RDF of the dictionary data. Then, the generated RDF should be indexed and published in a linked data server in order to allow navigation and querying mechanisms based on Semantic Web standards.

2.2 Design Decisions

We highlight here three important design decisions that were taken prior to starting the modelling activities and which influenced the whole process:

- The unique identifiers of the dictionary elements (the URIs) were built with *reusability* and *linking* in mind. That is, on the one hand, adding the necessary elements to the URI in order to prevent the collision of identifiers at a Web scale, and, on the other hand, ensuring the reuse of URIs of already defined lexical entries across dictionaries. The URI naming strategy goes back to the guidelines defined in (Gracia et al., 2016). As an example, the URI of the entry *cintura* ‘waist’ would read `:lexiconES/cintura-n`. The part-of-speech ending prevents collisions of entries that share written representation but not category (e.g. the verb and the noun *book*). By defining the URIs in this fashion, we do not depend on knowing the entry identifiers in order to link resources.

- We tried to *prevent any loss of information* in the conversion process. That is, all the modelling ingredients that were interpreted in the same way both in the KD data and in *lemon-ontolex* were mapped into *lemon-ontolex* entities (or LexInfo entities, if needed). However, in cases in which no clear equivalences were found, we opted for creating tailored entities in a KD namespace (e.g. the `kd:InflectionBlock` groups together different inflected forms). Information had to be preserved in order to guarantee the reconstruction of the original XML entries from their RDF representation. For this reason, some elements created under the KD namespace do not encode linguistic information but data to allow for this reconstruction.
- *Issues* detected in the source data. For instance, as a result of a non-strict interpretation of the KD guidelines on the editors’ side, contradictions or misuses were reported, but not corrected, during the transformation process. In that way, the responsibility of correcting possible issues and introducing improvements in the data is left to the KD editorial process. Thus, by correcting the data in origin according to the reported issues and improvement suggestions, future iterations of the conversion process will lead to RDF data of higher quality.

3. Analysis of the data

The representation model proposal was defined for the Spanish dataset of KD’s Global series, although the modelling and conceptual mappings were extended to some tags, values, and structures found in the English and German datasets, too. The original data in XML were provided by KD at the start of the project, along with the style guidelines for editors, the DTD, and additional documentation on specific tags used in the files. The Spanish dataset includes translations of headwords to Brazilian Portuguese, Norwegian, Japanese and Dutch.

An entry in these datasets encodes grammatical and phonetic information about the headword at the “Headword Container” level, although geographical usage restrictions and pragmatic aspects may be listed here as well. A snippet of the headword container of the word *abadía* ‘abbey’ is given in Example 1.

Example 1: XML of the entry *abadía* ‘abbey’

```
<Entry HomNum="" hw="abadía" identifier="EN00000019" pos="noun">
  <DictionaryEntry identifier="DE00000020" version="1">
    <HeadwordCtn>
      <Headword>abadia</Headword>
      <Pronunciation>aβa'ðia</Pronunciation>
      <PartOfSpeech value="noun" />
      <GrammaticalGender value="feminine" />
    </HeadwordCtn>
    [...]
  </DictionaryEntry>
```

³<https://www.w3.org/community/bpmlod/>

</Entry>

The senses of the headword follow this first container and each of them encapsulates its definition, translations equivalents of the headword, usage examples and translations, along with antonyms, synonyms, selectional and usage restrictions, etc., if available. Example 2 shows a synonym for *abadía* in Spanish, its definition, *monasterio con territorio propio* ‘monastery with own territory’ and the translation container with the Dutch translation for *abadía*, *abdiĳ*. Morphological, syntactico-semantic and pragmatic information about the translation may be provided as well (see *GrammaticalGender* in the example).

Example 2: XML with the translations of the headword *abadía* ‘abbey’

```
<SenseBlock>
  <SenseGrp identifier="SE00000039"
    version="1">
    <Synonym>convento</Synonym>
    <Definition>monasterio con
      territorio propio</Definition>
    <TranslationCluster identifier="
      TC00000097" text="monasterio
      con territorio propio" type="
      def">
      <Locale lang="nl">
        <TranslationBlock>
          <TranslationCtn>
            <Translation>abdiĳ</
              Translation>
            <GrammaticalGender value
              ="feminine" />
          </TranslationCtn>
        </TranslationBlock>
      </Locale>
      [...]
    </TranslationCluster>
    [...]
  </SenseGrp>
  [...]
</SenseBlock>
```

In addition to headword translations, senses comprise usage examples and their translations to other languages. Example 3 shows the translation into Dutch of a usage example of the first sense of *abadía*, *la Abadía de Westminster* ‘Westminster Abbey’.

Example 3: XML of the entry *abadía* ‘abbey’ with a usage example and its translation into Dutch

```
<ExampleCtn type="sid" version="1">
  <Example>la Abadía de Westminster</
    Example>
  <TranslationCluster identifier="
    TC00000098" text="la Abadía de
    Westminster" type="exmp">
  <Locale lang="nl">
    <TranslationBlock>
      <TranslationCtn>
```

```
<Translation>de Abdiĳkerk
  van Westminster</
  Translation>
</TranslationCtn>
</TranslationBlock>
</Locale>
[...]
</TranslationCluster>
</ExampleCtn>
```

Lastly, some senses include a *CompositionalPhrase* element that encodes idioms, collocations and frequent combinations. These groupings in turn have their own definition and set of senses with the corresponding translations, usage examples, and translated examples. For instance, one of the senses of the entry *agente* ‘agent’ provides the *CompositionalPhrase* *agente comercial* ‘commercial agent’.

In a first step, an exploratory analysis of the original data was carried out, together with the study of their associated guidelines. During the process a number of modelling doubts arose which were related to the use of certain tags (e.g. synonyms, cross-references) and to annotation differences across dictionaries. Some of these discrepancies might be due to the fact that different languages require different structures and elements to describe them (e.g. we only found the tag *Case* in the German dictionary), whereas others, such as, for instance, viewing a transitive verb and its reflexive use as two senses of the same entry or as two dictionary entries, are structural differences which call for further analysis to discern whether they are language-motivated or not. In addition, some errors were detected on the editors’ side (e.g. the incorrect use of a tag or its free value attribute), but this is frequent in complex dictionaries with several versions, as has been already pointed out (Declerck et al., 2015).

The task of defining a model proposal for the representation of these datasets in RDF was a challenging one due to two aspects: firstly, usage examples are not addressed in *lemon-ontolex* and neither are translations among them. This brought up the question of how the meaning of an example is to be captured on the first place and at which level a translation relation must be established, given that examples are not conceived as lexical entries; secondly, the design decisions (2.2) taken prior to the model definition phase involve the preservation of all information available in the XML, which is intrinsically related to the creation of new elements that must be both *lemon-ontolex* and *LexInfo* compliant as well as applicable to the three datasets.

4. Representation Model

In this section we briefly present the main features of the *lemon-ontolex* model and focus on one of the modules that constitute it, namely, the *vartrans* module, which serves to represent translations, amongst other linguistic descriptions. Then, we briefly discuss some cases in which extensions to the *lemon-ontolex* model were called for in order to illustrate the difficulties in the transformation process.

4.1 The *lemon-ontolex* model

The *lemon-ontolex* model is the resulting work of the W3C Ontology Lexica Community Group since 2011 to build a rich model that serves as interface between an ontology and the natural language descriptions that lexicalise the knowledge represented and structured in the ontology. It is largely based on the *lemon* model (McCrae et al., 2012), which, in turn, brings together the design principles of several previous models such as LexInfo (Buitelaar et al., 2009), LIR - Linguistic Information Repository (Montiel-Ponsoda et al., 2011), LMF - Lexical Markup Framework (Francopoulo et al., 2006) or SKOS - Simple Knowledge Organization System (Miles and Bechhofer, 2009).

lemon-ontolex has been implemented according to the RDF vocabulary, one of the recommended open Web standards for publishing linked data. It is a modular model that in its current version consists of a core set of classes and several modules that can be used depending on the type of linguistic descriptions that need to be represented. The core set of classes will be described in the following, as well as the *vartrans* module, which records lexico-semantic relations across entries in the same or different languages (translations, for instance). The other modules are the *decomp* module, for the decomposition of compound words or multiword expressions; the *synsem*, or Syntax and Semantics module, to establish correspondences between the syntactic structure of a certain linguistic description and its semantic realisation in the ontology; and the *lime* module, to account for the metadata related to the ontology-lexicon interface.

The model builds on the principle of *semantics by reference* (Buitelaar, 2010), which means that the semantics of linguistic descriptions is captured in the ontology by means of the classes, properties and individuals that represent a certain conceptualisation/concept. And, as said in the Final Model Specification of the model, "in some cases, the lexicon itself can add named concepts which are not made explicit in the ontology". However, *lemon-ontolex* can also be used to describe and represent linguistic resources that do not necessarily have a conceptualisation behind them, by creating the entities that represent that knowledge in an ad-hoc fashion as `skos:Concepts`, for instance. Finally, another principle in the design of the model is conciseness, in the sense that the model aims to offer the scaffolding to which deeper levels of linguistic descriptions are to be added by means of links to external ontologies that capture that knowledge (e.g. LexInfo).

Figure 1 depicts the main classes and properties of the *lemon-ontolex* model. The main class of the core of the model is the class `LexicalEntry`. A lexical entry can be a word, a multiword expression or an affix with a single part-of-speech, morphological pattern, etymology and set of senses. A lexical entry needs to be associated with at least one form, which represents one grammatical realisation of a lexical entry. Lexical entries can be linked to ontology entities in two ways: directly by the `denotes` property, or by means of an intermediate element called `LexicalSense`, which is intended to capture the particular sense of a word when referring to an ontology entity. The latter element allows us to attach additional properties (usage) to a pair consisting of a lexical entry and an on-

tology entity that describe under which conditions (context, register, domain, etc.) the lexical entry can be regarded as having the ontological entity as meaning. We can also represent the fact that a certain lexical entry evokes a mental concept or unit of thought that can be lexicalised by a given collection of senses. In that case, we would use the `LexicalConcept` class, which is a subclass of `skos:Concept`.

As for the *vartrans* module⁴, it has been developed to record lexico-semantic relations across entries in the same or different languages: those among senses and those among lexical entries and/or forms. Lexico-semantic relations among senses have a semantic nature and include synonymy, antonymy, hyperonymy-hyponymy, and terminological relations (dialectal, register, chronological, discursive, and dimensional variation) among senses in the same language, and translation relations among senses in different languages. In contrast, relations among lexical entries and/or forms concern the surface form of a term and encode morphological and orthographical variation, among other aspects.

4.2 Extension of *lemon-ontolex* for the K Dictionaries case

Some elements and structures in the original XML do not have a counterpart in *lemon-ontolex* or LexInfo allowing for their representation in RDF, for instance, usage examples⁵, translation relations among examples, identifiers of entries and dictionary entries⁶, aspects related to the display of the data in the interface, groupings of inflected forms, number of homographs, and other distinctive elements of the KD data (e.g. the tag `SenseQualifier`). This might be due to the fact that *lemon-ontolex* was not developed to model a dictionary with all its lexicographical annotations but to lexicalise an ontology. Furthermore, these dictionaries have been compiled with the human as final user too as well as for their use by NLP applications. Besides the above-mentioned elements, the KD vocabulary proposed as extension includes classes, individuals and properties that could not be directly mapped to LexInfo, primarily for two reasons: (1) mismatches between the DTD values of a tag and LexInfo classes (e.g. predefined values of a tag in the DTD are individuals from a class in LexInfo that is not compatible with the tag in KD), and (2) different level of granularity in the predefined values in the DTD and the individuals in LexInfo (e.g. `kd:possessive` vs `lexinfo:possessiveAdjective`, `lexinfo:possessiveDeterminer`, etc.). Given that in these cases a one-to-one mapping from KD into LexInfo was not viable, new elements had to be created under the KD namespace. Example of these KD tailored elements are `kd:SenseQualifier`, `kd:GeographicalUsage`, `kd:neuter-masculineGender`, and `kd:TranslationExampleCluster`, among others.

⁴See *lemon/ontolex* Final Model Specification.

⁵In contrast to *lemon*, which included the class `lemon:UsageExample`.

⁶In the KD data, an element with tag `<Entry>` may comprise one or several `<DictionaryEntries>`, see Example 1.

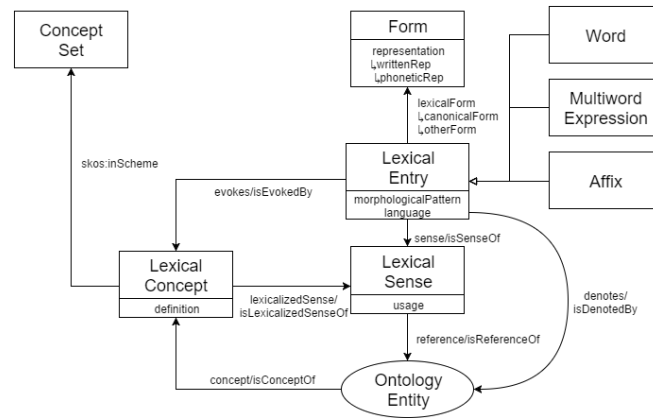


Figure 1: The *lemon-ontolex* core

5. Applying the model to the KD data

The aim of the project was to develop a prototype for the representation of the KD Spanish dataset that was general enough to be applicable to the English and German ones while preventing any loss of data in the conversion. If the latter two dictionaries were to be migrated, however, extensions to the model would be required. This prototype relies primarily on *lemon-ontolex* and has been defined in an iterative process involving project team meetings and feedback gathering. For those lexical descriptions in the original KD model that had not a direct correspondence in the *lemon-ontolex* model, ad-hoc KD entities were created (see section 4.2). In the following we present the core of the representation model, which addresses monolingual entries, and its extension to multilingual entries. Conceptual mappings from KD tags into *lemon-ontolex* and LexInfo elements which do not belong to the core or its multilingual extension, such as those annotations related to the geographical usage and selectional restrictions, compositional phrases, register, inflection groups, etc., were covered in the project but are not detailed in this paper.

Let us go back to the entry *abadía* introduced in section 3 and its sense ‘monastery with own territory’. Figure 2 depicts the representation suggested to capture the monolingual information associated with it, including definitions and synonyms. The model includes other properties and elements necessary to recreate the original XML from the RDF resulting from the conversion, such as entry identifiers, display elements, etc., but these are not shown in the diagram for simplification purposes.

Dictionary entries in the KD data are modelled as `ontolex:LexicalEntries` and each of their senses is represented with an `ontolex:LexicalSense` and a `skos:Concept`. The IRI of the entry *abadía* (`:lexiconES/abadía-n`) allows to easily reuse the resource every time another entry (from the same or different dictionary) provides information about the word *abadía* without the need to know the entry and dictionary entry identifiers. Following *lemon-ontolex*, the pronunciation and the written representation are provided at the level of the `ontolex:Form`. Additional pronunciations or transcriptions in different alphabets, if this were the case, would come at this level as well. The

part-of-speech and gender in nouns are attached to the `ontolex:LexicalEntry`, since they do not vary depending on different forms. KD Senses are conceived as `skos:Concept(s)` and are part of a conceptual layer that is language-independent and aims at representing the actual meaning denoted by the headword. Following this reasoning, definitions are linked to the `skos:Concept` at hand. The `ontolex:LexicalSense` class reifies the relation between a word and its meaning. Synonymy and antonymy are therefore understood as relations between lexical senses (sense relations), i.e. synonymy is seen as an equivalence relation between a relation x from a word y and a concept z and a relation h from a word i and the same concept z . Note that we do not have access to the information about the entry *convento* ‘convent’, which is given as synonym in the XML. This leads us to create an artificial lexical sense for *convento* from the lexical sense of *abadía* in which it occurs as a synonym.

Translations, usage examples and translated examples are treated similarly. Figure 3 shows the same sense for *abadía* mentioned before with its translations to Dutch. Just as synonymy and antonymy, translations are modelled as relations among lexical senses, being one sense in the source language (Spanish) and the other one in the target language (Dutch). The lexical sense of the target language and its corresponding entry were created artificially, since no pointers to the entries in other dictionaries are provided in the XML. This approach allows for the automatic growth of the Dutch lexicon, which would be extended later on if the KD’s Dutch dataset were converted. In addition, each sense is complemented with usage examples, which are represented in the model by the class `kd:UsageExample`. The IRI of the usage example of *abadía* shown here includes the translation cluster identifier given in the XML. The translation of a usage example is considered as a usage example itself and is therefore modelled as `kd:UsageExample` and linked to the lexical sense of the entry in Dutch (see *de Abdijkerk van Westminster*). This element resembles a lexical sense in that it may take part in a translation relation, reified here in the class `kd:TranslationExampleCluster`. In fact, new translations of the Spanish examples could be added in future versions of the dictionary and linked to this

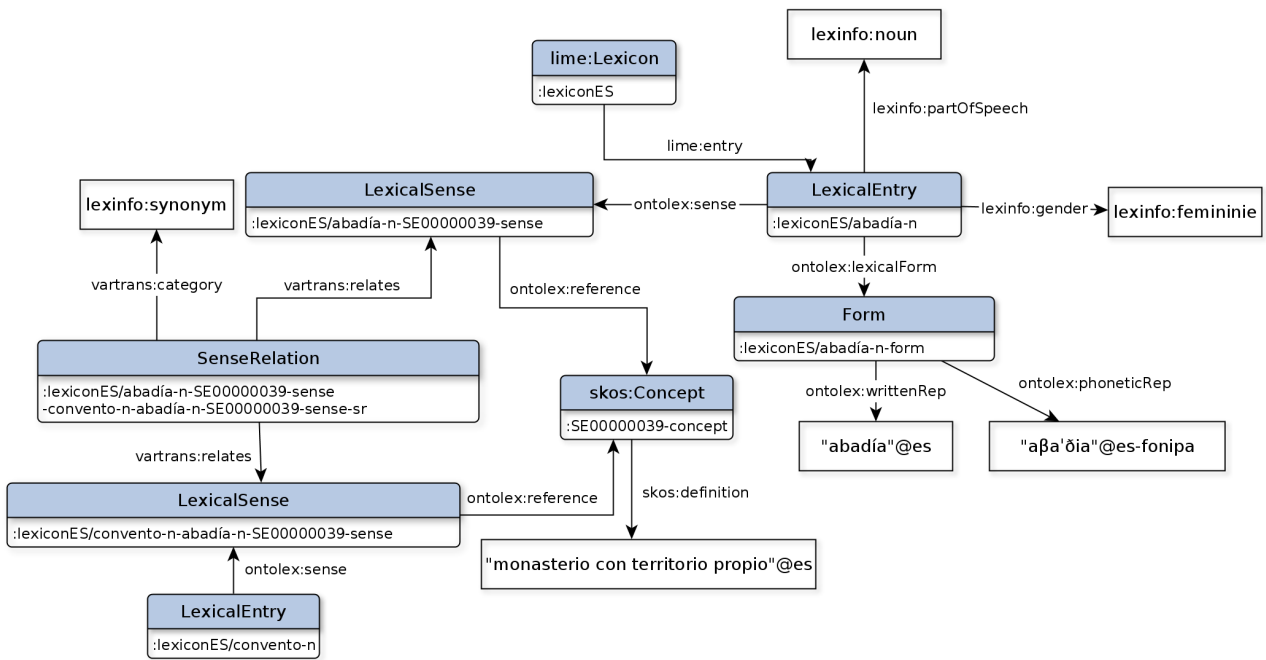


Figure 2: Core representation of the ontalex-based model for KD

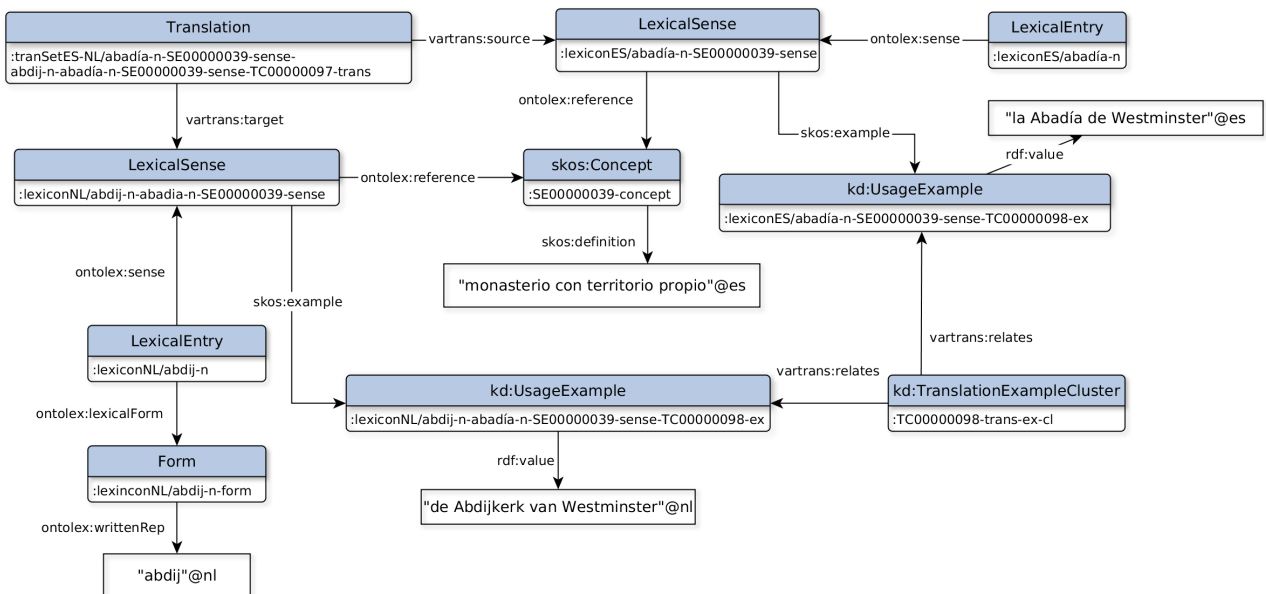


Figure 3: Representation of a KD multilingual entry

class.

6. Related Work

The idea of linking dictionaries or transforming them for their future linking is not new. For instance, Pearsall (2013) envisions a lexical hub with semantically annotated content that gradually grows with the integration of complementary lexical and lexicographic resources from the same domain. Our work is part of a recent trend that is exploring the use of linked data technologies in order to make such scenario real. For instance, Declerck et al. (2015) have proposed a proof-of-concept for encoding etymological and dialectal lexicographic data on the basis of *lemon-ontalex*. The au-

thors' ultimate aim is to generate a multilingual net of dictionaries grounded in concepts defined in external monolingual resources and which are shared across them. A cloud of interlinked lexical and lexicographic content like this one would foster the cooperation between lexicographers and collaborative researchers in general. A major difference between our work and this approach is that they do not propose the migration of the whole source representation scheme but only the subset of the lexical information that can be explicitly linked and merged.

The generation of RDF-based versions of bilingual dictionaries has also been tackled. In particular, the *lemon*-based version of the Apertium family of bilingual dictionaries has

been recently created (Gracia et al., 2016), linked to other resources such as BabelNet⁷ (Navigli and Ponzetto, 2010) and published on the Web of Data. In this work, the authors highlight the potential of a graph of interlinked bilingual dictionaries to support translation techniques, especially in the case of languages currently under-represented in the Linguistic LLOD cloud.⁸ Although we have largely relied on this previous experience, the complexity of the KD data meant that we had to define conversion strategies for many more lexical features than mere translations.

Special mention deserves the work carried out for the conversion of the German monolingual KD dataset with the *lemon* model (Klimek and Brümmer, 2015). Our approach shares some aspects with the authors', namely, the linking to LexInfo and the creation of a KD native vocabulary to represent those concepts for which *lemon-ontolex* or LexInfo had no counterpart. The dataset studied by the authors is however monolingual, and *lemon-ontolex* and LexInfo provide now means to encode information (new classes, individuals and properties) which were not available when they developed their work. Further, some core design decisions differ in both efforts, such as the rules for constructing URLs and the idea of preserving all the source information to allow the backwards conversion that we have followed (see Section 2.2).

Furthermore, some work has been done towards the alignment of medieval Latin dictionaries with the software Semantic MediaWiki, which allows for the integration of Latin terms with maps, time-lines and graphics in a collaborative friendly environment to researchers with less background in computer science (Bon and Nowak, 2013).

Finally, in addition to the above referred works which, in one way or another, rely on Semantic Web techniques, there are also works pursuing the automatic linking of lexicographical information but understanding the notion of *linking* in different terms than the Semantic Web community does. Just to mention an example, Renders et al. (2015) have accomplished the automatic linking of etymological dictionaries in terms of references between articles and not as part of the linked data paradigm. The authors propose a framework to link Gallo-Romance historical lexicographic resources that do not necessarily share lexical units on the basis of references to a common ancestry defined in the scientific reference work *Französisches Etymologisches Wörterbuch*.

7. Conclusions

In this work we have proposed a representation model to convert the data from the Spanish dataset of the KD multilingual Global series into RDF relying primarily on *lemon-ontolex*. The design decisions taken prior to the start of this effort are defined to guarantee both the reusability of the resulting lexical entries as well as the preservation of all the information encoded in the original XML files, which would allow for their reconstruction from the RDF output. This last point as well as some mismatches between the KD model and LexInfo led us to the creation of a KD vocabulary to serve as an extension to *lemon-ontolex* and LexInfo.

The core elements of the proposed model and its extension to account for translations, usage examples, and translated examples, which reuse elements from the *vartrans* module and include new ones inspired by them, were discussed. Future steps towards the development of a lexical hub of interlinked resources in collaboration with KD would involve the conversion of further KD datasets, thus extending the possibilities of exploiting and combining the multilingual lexical information contained across different lexica.

8. Acknowledgements

This work is supported by the Linked Data Lexicography for High-End Language Technology Application project (LDL4HELTA) of Semantic Web Company and K Dictionaries, by the Spanish Ministry of Economy and Competitiveness through the project 4V (TIN2013-46238-C4-2-R), the Excellence Network ReTeLe (TIN2015-68955-REDT), and the Juan de la Cierva program and by the Spanish Ministry of Education, Culture and Sports through the Formación del Profesorado Universitario (FPU) program.

9. References

- Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22.
- Bon, B. and Nowak, K. (2013). Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki. In *Electronic Lexicography in the 21st century: Thinking outside the paper: Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, pages 407–420.
- Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. (2009). Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference (ESWC09)*, pages 111–125.
- Buitelaar, P., (2010). *Ontology and the Lexicon*, chapter Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions, page 212–223. Cambridge University Press.
- Declerck, T., Wandl-Vogt, E., and Mörth, K. (2015). Towards a pan european lexicography by means of linked (open) data. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, pages 342–355, Herstmonceux Castle, United Kingdom, August.
- Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*.
- Fuertes-Olivera, P. A. and Bergenholtz, H., editors. (2011). *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. Continuum, London, New York.
- Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. (2016). The Apertium Bilingual Dictionaries on the Web of Data. *Semantic Web Journal [submitted for peer review]*.
- Klimek, B. and Brümmer, M. (2015). Enhancing lexicography with semantic language databases. *Kernerman Dictionary News*, 23:5–10.

⁷babelnet.org

⁸linguistic-lod.org/llod-cloud

- Manola, F. and Miller, E. (2004). RDF primer. Technical report, W3C Recommendation, February.
- McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., et al. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.
- Miles, A. and Bechhofer, S. (2009). SKOS-Simple Knowledge Organization System Reference.
- Montiel-Ponsoda, E., de Cea, G. A., Gómez-Pérez, A., and Peters, W. (2011). Enriching ontologies with multilingual information. *Natural Language Engineering*, 17(3):283–309.
- Moulin, C. and Nyhan, J. (2014). The Dynamics of Digital Publications. *Exploring the Paradigm Shift*, pages 47–61.
- Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.
- Pearsall, J. (2013). The future of dictionaries. *Kernerman Dictionary News*, 21:2–5.
- Renders, P., Baiwir, E., and Dethier, G. (2015). Automatically Linking Dictionaries of Gallo-Romance Languages Using Etymological Information. In *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 452–460. Ljubljana/Brighton.
- Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J., and Aguado-de Cea, G., (2014). *Publishing Linked Data: The Multilingual Dimension*, pages 101–118. Springer Berlin Heidelberg, August.

EcoLexicon: New Features and Challenges

Pamela Faber, Pilar León-Araúz, Arianne Reimerink

Department of Translation and Interpreting, Universidad de Granada

Buensusceso 11, 18071 Granada (Spain)

E-mail: pfaber@ugr.es, pleon@ugr.es, arianne@ugr.es

Abstract

EcoLexicon is a terminological knowledge base (TKB) on the environment with terms in six languages: English, French, German, Modern Greek, Russian, and Spanish. It is the practical application of Frame-based Terminology, which uses a modified version of Fillmore's frames coupled with premises from Cognitive Linguistics to configure specialized domains on the basis of definitional templates and create situated representations for specialized knowledge concepts. The specification of the conceptual structure of (sub)events and the description of the lexical units are the result of a top-down and bottom-up approach that extracts information from a wide range of resources. This includes the use of corpora, the factorization of definitions from specialized resources and the extraction of conceptual relations with knowledge patterns. Similarly to a specialized visual thesaurus, EcoLexicon provides entries in the form of semantic networks that specify relations between environmental concepts. All entries are linked to a corresponding (sub)event and conceptual category. In other words, the structure of the conceptual, graphical, and linguistic information relative to entries is based on an underlying conceptual frame. Graphical information includes photos, images, and videos, whereas linguistic information not only specifies the grammatical category of each term, but also phraseological, and contextual information. The TKB also provides access to the specialized corpus created for its development and a search engine to query it. One of the challenges for EcoLexicon in the near future is its inclusion in the Linguistic Linked Open Data Cloud.

Keywords: Terminology, knowledge representation, terminological knowledge base

1. Introduction

EcoLexicon (ecolexicon.ugr.es) is a multilingual visual thesaurus of environmental science (Faber, León-Araúz, and Reimerink 2014). It is the practical application of Frame-based Terminology (FBT; Faber et al. 2011; Faber 2012, 2015), a theory of specialized knowledge representation that uses certain aspects of Frame Semantics (Fillmore 1985; Fillmore and Atkins 1992) to structure specialized domains and create non-language-specific representations. FBT focuses on: (i) conceptual organization; (ii) the multidimensional nature of specialized knowledge units; and (iii) the extraction of semantic and syntactic information through the use of multilingual corpora. EcoLexicon is an internally coherent information system, which is organized according to conceptual and linguistic premises at the macro- as well as the micro-structural level.

From a visual perspective, each concept appears in a network that links it to all related concepts. The semantic networks in EcoLexicon are based on an underlying domain event, which generates templates for the most prototypical states and events that characterize the specialized field of the Environment as well as the entities that participate in these states and events. This type of visualization was selected because a semantic network is an effective representation method for capturing and encapsulating large amounts of semantic information in an intelligent environment (Peters and Shrobe 2003). The representations generated for each concept are obtained from the information extracted from static knowledge sources such as a multilingual corpus of texts and other environmental resources.

EcoLexicon currently has 3,599 concepts and 20,106 terms in Spanish, English, German, French, Modern Greek, and Russian, though terms in more languages are currently

being added. This terminological resource is conceived for language and domain experts as well as for the general public. It targets users such as translators, technical writers, and environmental experts who need to understand specialized environmental concepts with a view to writing and/or translating specialized and semi-specialized texts.

2. Frame-based Terminology

Frame-based Terminology (FBT) is the theoretical approach used to create EcoLexicon. Based on cognitive semantics (Geeraerts 2010) and situated cognition (Barsalou 2008), specialized environmental knowledge is stored and structured in the form of propositions and knowledge frames, which are organized in an ontological structure.

FBT is a cognitively-oriented terminology theory that operates on the premise that, in scientific and technical communication, specialized knowledge units activate domain-specific semantic frames that are in consonance with the users' background knowledge. The specification of such frames is based on the following set of micro-theories: (i) a semantic micro-theory; (ii) a syntactic micro-theory; and (iii) a pragmatic micro-theory. Each micro-theory is related to the information in term entries, the relations between specialized knowledge units, and the concepts that they designate (Faber 2015).

More concretely, the semantic micro-theory involves an internal and external representation. The internal representation is reflected in a definition template used to structure the meaning components and semantic relations in the description of each specialized knowledge unit (see Section 5). The external representation is a domain-specific ontology whose top-level concepts are OBJECT, EVENT, ATTRIBUTE, and RELATION. The ontology is based on the conceptual representations of physical objects and

processes (e.g. ALLUVIAL FAN, GROUYNE, EROSION, WEATHERING, etc.). This set of concepts acts as a scaffold, and their natural language descriptions provide the semantic foundation for data querying, integration, and inferencing (Samwald et al. 2010).

The syntactic micro-theory is event-based and takes the form of predicate-argument structures. The nature of an event depends on the predicates that activate the relationships between entities. According to FBT, terms and their relations to other terms have a syntax, as depicted in graph-based micro-grammars, which not only show how hierarchical and non-hierarchical relations are expressed in different languages, but can also tag corpus texts for information retrieval (León and Faber 2012).

Finally, the pragmatic micro-theory is a theory of contexts, which can be linguistic or extralinguistic. Linguistic contexts are generally regarded as spans of +5 items before and after term occurrence. They are crucial in the design stage of a terminological knowledge base (TKB) for a wide variety of reasons, which include: (i) term disambiguation; (ii) definition formulation; (iii) linguistic usage; (iv) conceptual modeling; and (v) term extraction. Such contextual information is important because it shows how terms are activated and used in specialized texts in the form of collocations and collocational patterns.

In contrast, extralinguistic contexts are pointers to cultural knowledge, perceptions, and beliefs since many specialized knowledge units possess an important cultural dimension. Cultural situatedness has an impact on semantic networks since certain conceptual categories are linked to the habitat of the speakers of a language and

derive their meaning from the characteristics of a given geographic area or region and, for example, the weather phenomena that typically occur there

Based on these theoretical premises, EcoLexicon has evolved and has made significant advances since it was first created a decade ago. Section 3 explains the interface of the application, the knowledge provided to users, and the various interaction options. Section 4 describes the contextualization of knowledge to avoid information overload. Section 5 explains how natural language definitions are created according to FBT premises. Section 6 shows the search possibilities of the EcoLexicon corpus. Section 7 addresses one of the future challenges of the resource, its inclusion in the Linguistic Linked Open Data Cloud, and Section 8 draws some final conclusions.

3. User interface

Users interact with EcoLexicon through a visual interface with different modules that provide conceptual, linguistic, and graphical information. Instead of viewing all information simultaneously, they can browse through the windows and select the data that is most relevant for their needs.

Figure 1 shows the entry in EcoLexicon for FAN. When users open the application, three zones appear. The top horizontal bar gives users access to the term/concept search engine. The vertical bar on the left of the screen provides information regarding the search concept, namely its definition, term designations, associated resources, general conceptual role, and phraseology.

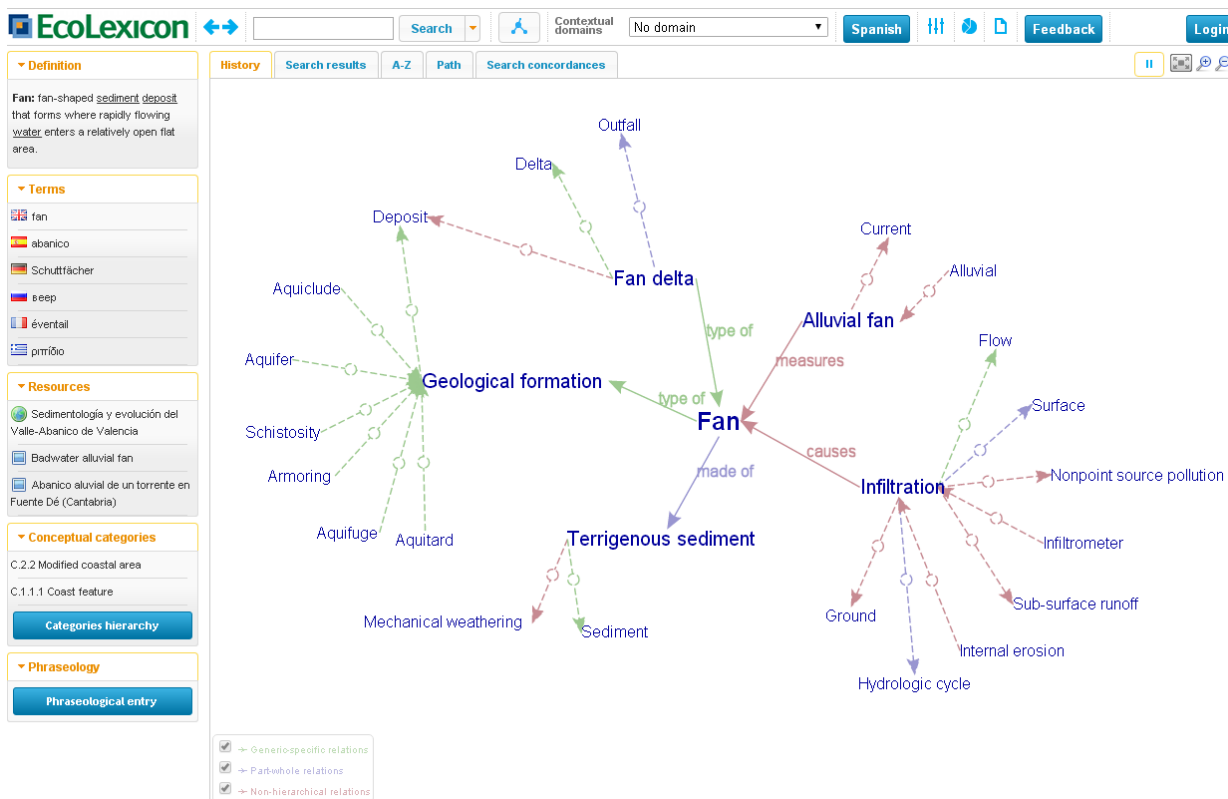


Figure 1: EcoLexicon user interface

The topmost box shows the definition of the concept. Each definition makes category membership explicit, reflects a concept's relations with other concepts, and specifies essential attributes and features (see Section 5). Accordingly, the definition is the linguistic codification of the relational structure shown in the concept map. The words in each definition also have hyperlinks to their corresponding concept in the knowledge base.

The box directly below shows the terms designating the search concept in various languages. The list is organized according to language and term type (main entry term, synonym, variant, acronym, etc.). At the left of each term is the flag of the country where the language is spoken. A click on the term provides further linguistic information regarding language, term type, gender, part of speech, and concordances.

The third box provides resources (images, documents, URLs, audiovisual material, etc.) associated with each concept/term. The fourth box shows the very general conceptual role that the concept normally has within the Environmental Event (EE). The EE is a basic template in which any environmental process is conceived of as initiated by an agent, affecting a patient (environmental entity), and producing a result, often in a geographical area. Each concept is associated with one or more conceptual categories, which are shown as a list. Also included is a *Category Hierarchy* icon, which shows the concepts in a hierarchical format in which nodes can expand or retract.

The Phraseology box is currently under construction and shows a list of verbs most commonly used with the term within different phraseological patterns. So far, this option is only available for a small number of terms, such as *hurricane* (Figure 2).

The screenshot shows a 'Phraseology' window with two sections. The first section is for the nuclear meaning 'ACTION', with a meaning dimension 'to_come_against_sth_with_sudden_force' and a phraseological pattern 'NATURAL FORCE comes against PATIENT with sudden force, affecting it negatively.' It lists verbs: hit, batter, strike, blast3. The second section is for the nuclear meaning 'CHANGE', with a meaning dimension 'to_cause_to_change_for_the_worse' and a phraseological pattern 'NATURAL DISASTER causes a PATIENT to change for the worse.' It lists verbs: affect, damage, demolish, destroy, devastate, injure, sweep away, wreck, ravage.

Figure 2: Phraseological information for *hurricane*

The center area has tabs that access the following: (i) the history of concepts/terms visited; (ii) the results of the most recent query; (iii) all the terms alphabetically

arranged; (iv) the shortest path between two concepts; and (v) concordances for a term (see Section 6).

On the center of the screen, the conceptual map is shown as well as the icons that permit users to configure and personalize it for their needs (see Section 4). The standard representation mode shows a multi-level semantic network whose concepts are all linked in some way to the search concept, which is at its center.

When users click on any of the concepts in the map, (for example, FAN DELTA), the network rearranges itself. In this new map, FAN DELTA is at the center along with its set of related concepts (see Figure 3).

By right-clicking on a concept in the map, the user can access the contextual menu (Figure 3). This menu can be used to perform any of the following actions: (i) centering the concept; (ii) fixing a node by dragging it to a certain position; (iii) showing details of the concept (definition, associated terms, resources, etc.) by selection on the sidebar; (iv) generating a URL for direct access to the concept selected; (v) searching Google Images, Google, and Wolfram Alpha; (vi) removing a concept and its related concepts from the map. Any of these actions enhances concept representation by providing a rich quantity of conceptual information, according to the specific needs of each end user.

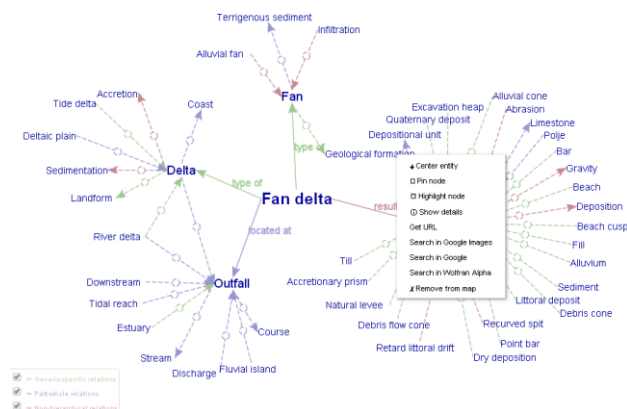


Figure 3: Conceptual map of FAN DELTA and contextual menu

EcoLexicon also includes icons to personalize concept map visualization such as *Zoom map*, *Zoom out map*, and *Fullscreen*. *Stop layout* deactivates the automatic arrangement of concepts in the network, thus allowing users to configure the map by dragging concepts to the desired location.

The *Settings* icon further customizes semantic networks by allowing users to establish the depth of the network, namely, its maximum conceptual level. Similarly, they can also decide whether they wish to visualize the names of all semantic relations since, by default, relation labels only appear when the relation includes the central concept. If this value is activated, all relations will have labels.

4. Information overload and multidimensionality

The scope and multidimensionality of the environmental domain, as well as the great deal of conceptual propositions represented in EcoLexicon, has resulted in an information overload problem. This problem has been solved in different quantitative and qualitative ways: (i) by letting the user filter overloaded networks by relation type, (ii) by offering a recontextualized view of concepts according to subject-field based contextual constraints, and (iii) by providing different access modes to the visualization of concepts' behaviour (network mode, tree mode, and path mode).

In the lower left-hand corner of the conceptual map (Figure 1 and 3) there is a text box that allows users to identify the three categories of conceptual relation in EcoLexicon: (i) hyponymic (*type_of*) relations; (ii) meronymic (*part_of*) relations; (iii) non-hierarchical relations (*has_function*, *located_at*, *causes*, *affects*, *result_of*, etc.). These relations, which are related to Pustejovsky's (1995) *qualia*, belong to a closed inventory that is currently being revised to make them more fine-grained and provide them with greater relational power. The checkboxes at the left of each label can be used to activate or deactivate the visualization of a certain type of relation so that it does not appear on the map. This allows users to filter overloaded networks based on relation types. Recontextualized networks can be visualized by choosing one of the contextual domains from a pull-down menu (upper ribbon in Figure 1).

This is a qualitative way to solve the information overload problem while enhancing the representation of multidimensionality. Recontextualized networks are reshaped according to how the relational behaviour of concepts varies according to perspective. Instead of representing all possible dimensions of a concept, conceptual propositions are activated or constrained based on their salience in different subject fields (León-Araúz et al. 2013).

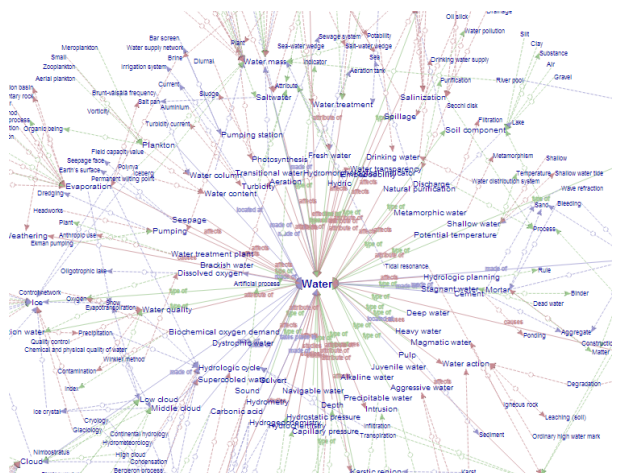


Figure 4: Context-free overloaded network of WATER

In Figure 4, WATER appears in a context-free overloaded network – hardly meaningful to users – while in Figure 5 the same concept is framed in the Civil Engineering domain, whose network is substantially reduced.

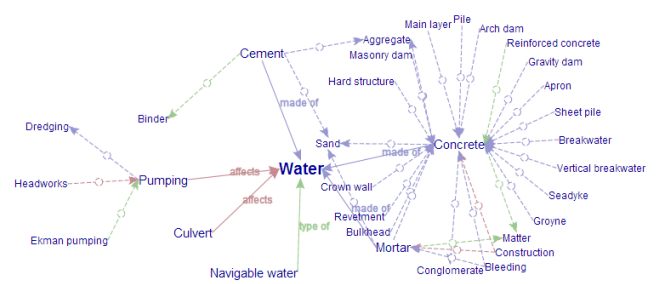


Figure 5: Network of WATER the Civil Engineering domain

Regarding the representation mode, users can also choose between a tree mode and a path mode. The tree mode generates a *type_of* hierarchy for the concept (Figure 6). In contrast, in the path mode users choose two concepts that will be the beginning and end of the path, and the application calculates and draws the shortest distance between them (Figure 7).

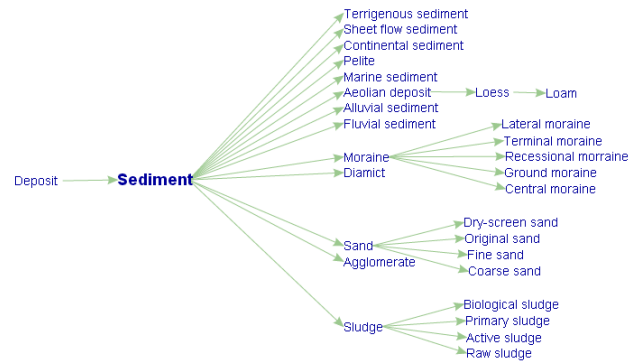


Figure 6: Tree mode of SEDIMENT



Figure 7: Path mode of HURRICANE and SAND

5. Natural language definitions

In EcoLexicon, definitions are based on the most representative conceptual propositions established by the concept in EcoLexicon. Each conceptual proposition is considered to be a feature of the concept and the representativeness of each feature is determined by the category assigned to the concept being defined. Each category has a set of representative conceptual relations that describe it, which is schematically represented in a definitional template (León Araúz, Faber, and Montero Martínez 2012: 153–154).

When applying a template to a concept, it may only inherit the relation with the defined concept in the template or activate a more specific concept than the one in the template. An example would be the template for HARD_COASTAL_DEFENCE_STRUCTURE (Table 1), which is applied to the definition of GROUYNE (Table 2), a member of

this category.

HARD COASTAL DEFENCE STRUCTURE	
<i>type_of</i>	CONSTRUCTION
<i>located_at</i>	SHORELINE
<i>made_of</i>	MATERIAL

Table 1: HARD_COASTAL_DEFENCE_STRUCTURE definitional template (León Araúz et al. 2012: 156)

GROYNE	
Hard coastal defence structure made of concrete, wood, steel and/or rock perpendicular to the shoreline, built to protect a shore area, retard littoral drift, reduce longshore transport and prevent beach erosion.	
<i>type_of</i>	HARD COASTAL DEFENCE STRUCTURE
<i>located_at</i>	PERPENDICULAR TO SHORELINE
<i>made_of</i>	CONCRETE WOOD METAL ROCK
<i>has_function</i>	SHORE PROTECTION LITTORAL DRIFT RETARDATION LONGSHORE TRANSPORT REDUCTION BEACH EROSION PREVENTION

Table 2: Definition of GROUYNE after the application of the HARD_COASTAL_DEFENCE_STRUCTURE definitional template (León Araúz et al. 2012: 156)

As explained in Section 4, the multidimensional nature of the environment can cause information overload because some concepts present a high level of contextual variation. This can be prevented if the information shown is reduced according to the propositions present in specific conceptual domains. These versatile concepts, therefore, behave differently according to the contextual domain chosen. This has consequences for how these concepts are defined. In the same way that a single network becomes overloaded, a single definition cannot encompass all propositions present in the entire environmental domain and is therefore not sufficiently informative (San Martín and León-Araúz 2013).

For that reason, we are working on the creation of ‘flexible definitions’. A flexible definition is a system of definitions for the same concept composed of a general environmental definition along with a set of recontextualized definitions derived from it, which situate the concept in different domains (San Martín 2016). Table 3 is an example of the resulting definitions for the entry SAND.

SAND	
Environment as a whole	Mineral material consisting mainly of particles of quartz ranging in size of 0.05-2 mm.
Geology	Sediment consisting mainly of particles of quartz ranging in size of 0.05-2 mm that is part of the soil and can be found in great quantities in beaches, river beds, the seabed,

	and deserts.
Soil Sciences	Unconsolidated inorganic soil component consisting mainly of particles of quartz ranging in size of 0.05-2 mm that are the result of weathering and erosion. It renders soils light, acidic, and permeable.
Civil Engineering	Natural construction aggregate consisting mainly of particles of quartz ranging in size of 0.05-2 mm that is mixed with cement, lime and other materials to produce concrete and mortar.

Table 3: Extract of the flexible definition of SAND

6. The EcoLexicon corpus

In EcoLexicon, a specialized corpus was specifically compiled in order to extract linguistic and conceptual knowledge. Then, it was classified and tagged in order to provide our users with a direct and flexible way of accessing the corpus, which is available in the *Search concordances* tab (Figure 1).

Currently, the corpus has more than 50 million words and each of its texts has been tagged according to a set of XML-based metadata (Figure 8). These tags contain information about the language of the text, the author, date of publication, target reader, contextual domain, keywords, etc. Some of them are based on the Dublin Core Schema (<dc>) and some others have been included based on our own needs (<eco>).

```
<?xml version="1.0" ?>
- <metadata xmlns:xsi="http://www.w3.org/2001/XMLSchema-instanc
  xmlns:eco="http://manila.ugr.es/tags/0.1">
- <header>
  <dc:title>Coastal Engineering Manual Part 1 Chapter 2 History of (
  <dc:creator>US Army Corps of Engineers</dc:creator>
  <eco:respon>adm</eco:respon>
  <dc:date>2002-04-30</dc:date>
  <eco:country>us</eco:country>
  <eco:domain>3.2.3</eco:domain>
  <dc:subject>coastal engineering</dc:subject>
  <dc:subject>history</dc:subject>
  <dc:subject>evolution</dc:subject>
  <dc:subject>military</dc:subject>
  <dc:subject>civil engineering</dc:subject>
  <eco:user>s</eco:user>
  <eco:text>book</eco:text>
  <dc:language>en</dc:language>
  <eco:variant>am</eco:variant>
  <eco:note />
</header>
<body>History of Coastal Engineering I-3-i Chapter 3 EM 1110-2-11
```

Figure 8: Corpus metadata

This allows constraining corpus queries based on pragmatic factors, such as contextual domains or target reader. In this way, users can compare the use of the same term in different contexts. For instance, Figure 9 shows the concordances of *sediment* in Environmental engineering texts, while Figure 10 shows the concordances of the same term in an Oceanography context. In the same way, in Figures 11 and 12 the query for *sand* is constrained according to expert and lay settings respectively.

Furthermore, in the future, the corpus will be expanded and annotated with a POS tagger in order to enable richer queries.

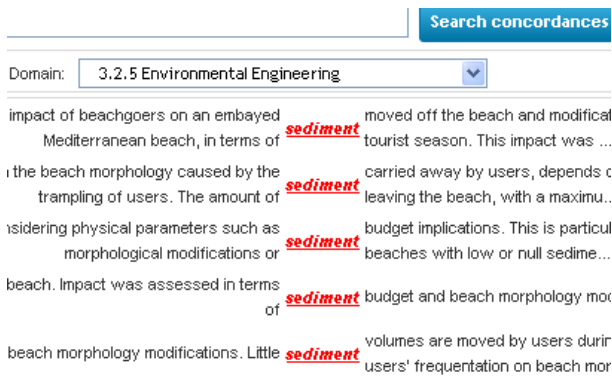


Figure 9: Concordances of *sediment* in Environmental Engineering

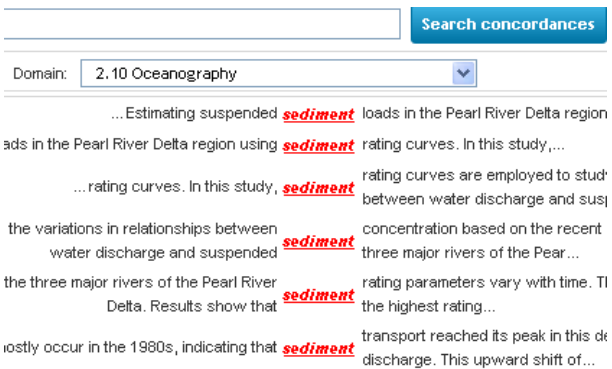


Figure 10: Concordances of *sediment* in Oceanography

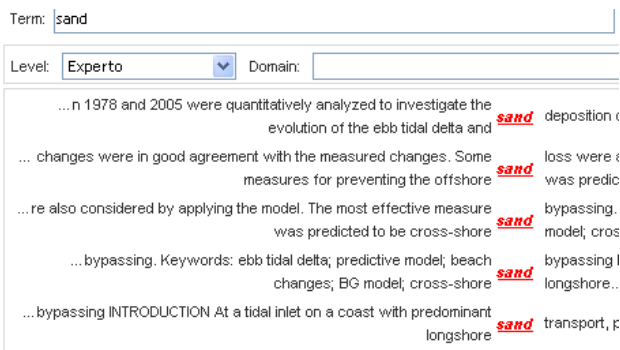


Figure 11: Concordances of *sand* in expert-to-expert texts

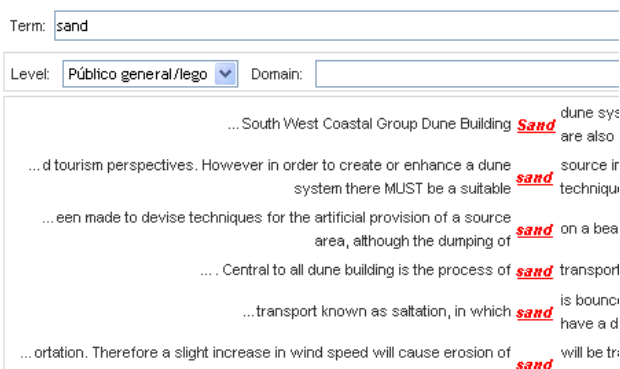


Figure 12: Concordances of *sand* in expert-to-lay texts

7. EcoLexicon-LD

Apart from annotating the corpus, expanding the phraseological module, and creating flexible definitions for all versatile concepts, one of the major challenges in EcoLexicon is to integrate the resource in the Linguistic Linked Open Data Cloud (León-Araúz et al. 2011a, 2011b).

Linked Data is an important initiative for creating a shared information space by publishing and connecting structured resources in the Semantic Web (Bizer et al. 2008). However, the specification of semantic relationships between data sources is still a stumbling block.

First of all, the TKB was converted to an RDF ontology in order to link it to other resources and provide the ways in which other resources can be linked to EcoLexicon. Thus, in the near future EcoLexicon will be available in three ways, as depicted in Figure 13: (i) the web application, as it is currently presented; (ii) another web application where EcoLexicon-LD can be browsed by humans; and (iii) a SPARQL endpoint.

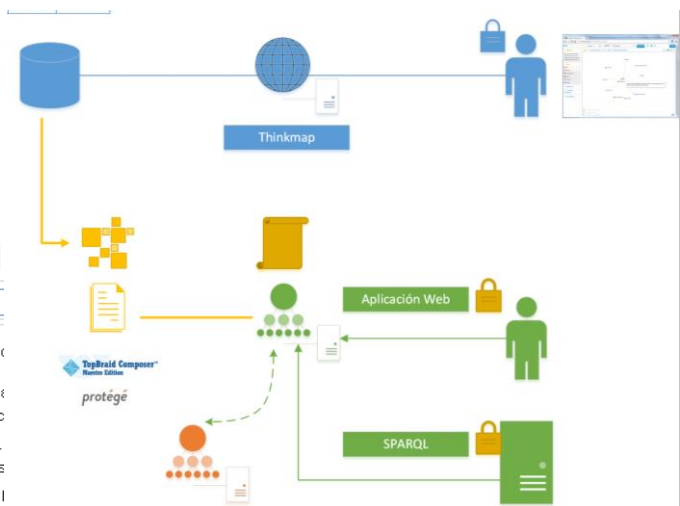


Figure 13: Access to EcoLexicon-LD

After that, a linking algorithm was designed in order to automatize the mappings between DBpedia and EcoLexicon (Figure 14).

Instead of mapping one-to-one manual correspondences between the entities contained in each of the resources, the matching algorithm performs sense disambiguation by exploiting the semantics of each data set. The data categories that are used from EcoLexicon are those related to linguistic variants, multilingual choices and semantic relations, which are mapped against the properties in DBpedia containing text.

Therefore, the first step in the data linking process is the comparison of the string of all English variants in EcoLexicon with the *rdfs:label* property of DBpedia. Since these strings may match various entries in DBpedia and lead to erroneous mappings, disambiguation is then performed by comparing other multilingual equivalents.

1. Get all ECOLEXICON concepts $C = \{c_1, \dots, c_i, \dots, c_n\}$
2. For each c_i in C
 - 2.1. Search in DBPEDIA resources $D = \{d_1, \dots, d_j, \dots, d_m\}$ such that $c_i.rdfs:label == d_j.rdfs:label$ (exact match @en)
 - 2.2. if $|D| == 0$
 - # No match, end procedure
 - 2.3. if $|D| == 1$
 - # Match
 - $R = \{d_1\}$
 - 2.4. if $|D| > 1$
 - # Disambiguation required
 - 2.4.1. Search in ECOLEXICON $T^{c_i} = \{t_1, \dots, t_k, \dots, t_p\}$ such that t_k is a term of c_i (any language)
 - 2.4.2. For each d_j in D
 - 2.4.2.1. Search in DBPEDIA $L^{d_j} = \{l_1^{d_j}, \dots, l_i^{d_j}, \dots, l_q^{d_j}\}$ such that $l_i^{d_j} == d_j.owl:sameAs$ (any language)
 - 2.4.3. Select $D^{max} = \{d_j\}$ such that $\max(|T_{c_i} \text{ intersection } L_{d_j}|)$
 - 2.4.4. if $|D^{max}| == 1$
 - # Match
 - $R = \{d_j\}$
 - 2.4.5. if $|D^{max}| > 0$
 - # Disambiguation required
 - 2.4.5.1. $T_{c_i} = T_{c_i} \cup T_{c_i}^*$ such that c_i^* is associated to c_i in ECOLEXICON and lemmatized
 - 2.4.5.2. For each d_j in D^{max}
 - 2.4.5.2.1. $X^{d_j} = \{x_1, \dots, x_s, \dots, x_t\}$ such that $(x_s == d_j'.rdfs:comment \ || \ x_s == d_j'.dbpedia-owl:abstract)$ and lemmatized
 - 2.4.5.3. Select $D^{max_text} = \{d_j\}$ such that $\max(|T_i \text{ intersection } X^{d_j}|)$
 - 2.4.5.4. $R = D^{max_text}$

Figure 14: Linking algorithm

Nevertheless, in those cases in which polysemy also occurs at a cross-linguistic level – or no multilingual choices are available – semantic information comes into play. If any term belonging to the same contextual domain of the search concept appears in any of the text-related DBpedia properties (i.e. *rdfs:comment*; *dbpedia-owl:abstract*, etc.), then concepts are considered equivalents (Figure 15).

The image shows three DBpedia entries for 'Accretion'. Each entry has a title, a list of terms, subjects, and a comment. Red circles highlight specific terms in the comments: 'ice' in the atmosphere entry, 'sediment', 'beach', and 'weather' in the coastal management entry.

Figure 15: DBpedia dataset for ACCRETION

The final step will be to provide access to EcoLexicon-LD, where any registered user will be able to validate and evaluate the reliability of each link (Figure 16).

The image shows the 'fan' concept page in EcoLexicon-LD. It lists several linked resources with their DBpedia URIs and version numbers (e.g., 69.5, 70.8, 66.0, 66.0). Below the links, there are definitions in English and Spanish. There are also sections for 'terms' and 'relations'.

Figure 16: EcoLexicon-LD validation form

This will allow for the development of a validation protocol, from which new conclusions could be drawn for the future linking of new resources and the improvement of the algorithm.

8. Conclusion

In the past decade, EcoLexicon has evolved and made significant advances in the representation of environmental knowledge. As well as the specialized domain the TKB represents, it must grow and adapt to new scientific advances. Apart from adding new conceptual knowledge and improving the already existing modules, e.g. adding phraseological information to all entries of the TKB, we have been able to broaden our scope by giving access to contextualized networks, a specialized corpus on the environment, and to other web-related options such as Google images and Wolfram Alpha. The next challenge is to improve the reusability of all this coherently organized knowledge. One way we envision to this end is linking EcoLexicon to other knowledge bases in the Linguistic Linked Open Data Cloud.

9. Acknowledgements

This research was carried out as part of project FF2014-52740-P, *Cognitive and Neurological Bases for Terminology-enhanced Translation (CONTENT)*, funded by the Spanish Ministry of Economy and Competitiveness.

10. References

- Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology* 59, pp- 617–645.
- Bizer, C., Heath, T. and Berners-Lee, T. (2008). Linked Data: Principles and State of the Art. *World Wide Web Internet And Web Information Systems*.
- Faber, P. (2011). The dynamics of specialized knowledge representation: simulational reconstruction or the perception-action interface. *Terminology* 17(1), pp. 9–29.

- Faber, P. (Ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/New York: de Gruyter.
- Faber, P. (2015). Frames as a framework for terminology. In H. Kockaert and F. Steurs (Eds.), *Handbook of Terminology*, Amsterdam/Philadelphia: John Benjamins, pp.14-33.
- Faber, P., León-Araúz, P. and Reimerink, A. (2014) Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era. Educational Linguistics*, 19:267-301. Springer
- Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica* 6: 222-254.
- Fillmore, C. J., and Atkins. B. T. (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbors. In A. Lehrer and E. Feder Kittay (ed.) *Frames, Fields and Contrasts*, 102:75-102. Hillsdale, New Jersey: Lawrence Erlbaum Assoc.
- Geeaerts, D. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press.
- León-Araúz, P., Reimerink, A. and García Aragón, A. (2013) Dynamism and context in specialized knowledge. *Terminology*, 19(1):31-61. John Benjamins Publishing Company. doi:10.1075/term.19.1.02leo.
- León-Araúz P. and Faber, P. (2012). Causality in the specialized domain of the environment. In *Proceedings of the Workshop Semantic Relations-II. Enhancing Resources and Applications (LREC'12)*. Istanbul: ELRA, pp. 10-17.
- León-Araúz, P., Faber, P. and Magaña Redondo, P.J. (2011a). Linking Domain-Specific Knowledge to Encyclopedic Knowledge: an Initial Approach to Linked Data. In *2nd Workshop on the Multilingual Semantic Web*. 68-73. Bonn.
- León Araúz, P., Faber, P. and Montero Martínez, S. (2012). Specialized language semantics. In P. Faber (ed.) *A Cognitive Linguistics View of Terminology and Specialized Language*, 95–175. Berlin, Boston: De Gruyter Mouton.
- León-Araúz, P., Magaña Redondo, P.J. and Faber, P. (2011b). Integrating Environment into the Linked Data Cloud. In *Proceedings of the 25th International Conference Environmental Informatics. EnviroInfo Ispra 2011*, edited by Pillman, W., Schade, S. & Smits, P., pages 370-379. Shaker Verlag.
- Peters, S. and Shrobe, H. (2003). Using semantic networks for knowledge representation in an intelligent environment. In *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications*. Washington D. C., IEEE Computer Society, pp. 323–337.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.
- Samwald, M.; Chen, H.; Ruttenberg, A.; Lim, E.; Marengo, L.; Miller, P.; Shepherd, G.; and Cheung, K. H. (2010). Semantic SenseLab: implementing the vision of the Semantic Web in neuroscience. *Artificial Intelligence in Medicine* 48, pp. 21–28.
- San Martín, A. and León Araúz, P. (2013) Flexible Terminological Definitions and Conceptual Frames. In *Proceedings of the International Workshop on Definitions in Ontologies (DO 2013)*, Seppälä, S. & Ruttenberg, A. (eds.). Montreal: Concordia University.
- San Martín, A. (2016) *La representación de la variación contextual mediante definiciones terminológicas flexibles*. PhD Thesis. University of Granada.

Determining the Characteristic Vocabulary for a Specialized Dictionary using Word2vec and a Directed Crawler

Gregory Grefenstette
Inria Saclay/TAO, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
gregory.grefenstette@inria.fr

Lawrence Muchemi
Inria Saclay/TAO, , Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
lawrence.githiari@inria.fr

ABSTRACT

Specialized dictionaries are used to understand concepts in specific domains, especially where those concepts are not part of the general vocabulary, or having meanings that differ from ordinary languages. The first step in creating a specialized dictionary involves detecting the characteristic vocabulary of the domain in question. Classical methods for detecting this vocabulary involve gathering a domain corpus, calculating statistics on the terms found there, and then comparing these statistics to a background or general language corpus. Terms which are found significantly more often in the specialized corpus than in the background corpus are candidates for the characteristic vocabulary of the domain. Here we present two tools, a directed crawler, and a distributional semantics package, that can be used together, circumventing the need of a background corpus. Both tools are available on the web.

1. Introduction

Specialized dictionaries (Caruso, 2011) and domain-specific taxonomies are useful for describing the specific way a language is used in a domain, and for general applications such as domain-specific annotation or classification. To create a specialized dictionary, it is first necessary to determine the characteristic vocabulary to be included. These are words that are either specific to the domain, or common words that have specialized usages within the domain. Recent advances using machine learning in natural language processing have led to the development of distributional semantic tools, such as *word2vec*, which use unsupervised training over a large corpus of text to embed words in an N -dimensioned vector space (Goldberg and Levy, 2014). These vectors have the desirable property that words that are substitutable, or found in similar contexts, have vectors that are close together in this vector space, and using a distance function, such as cosine distance, reveals words which are semantically similar or related to a given word, or words. To discover the characteristic vocabulary of a domain, it is interesting to see what words are semantically related within that domain. Since the semantic relationships are learned from an underlying corpus, it seems evident that the corpus should be drawn from texts concerning the domain. As a general solution, we have created a directed crawler to build a corpus for any given domain. From this corpus, we can extract the characteristic vocabulary for the domain, and build more complex lexical structures such as taxonomies.

Here, in this article, we present the various pieces that can be assembled to create specialized vocabularies and domain-specific taxonomies. In the next section, we describe how this

crawler works. This is followed by a description of one distributional semantics tool, *word2vec*. Then we show how these two tools can be used together to extract the basis of a specialized vocabulary for a domain.

2. Building a Directed Crawler

A directed crawler is a web crawler for gathering text corresponding to a certain subject. A web crawler is a program that continuously fetches web pages, starting from a list of seed URLs¹. Each web page fetched contributes new URLs which are added to the list of the remaining URLs to be crawled. A directed crawler (Chakrabarti et al. 1999) only adds new URLs to this list if the fetched web page passes some filter, such as being written in a given language, or containing certain key words.

In our directed crawler, we begin our crawl using a list of seed URLs from the Open Directory Project² (ODP) whose crowd-sourced classification of web pages has been used in many lexical semantic projects (e.g., Osiński and Weiss, 2004; Lee et al., 2013; Ševa et al., 2015). To gather the seed list, we send a query concerning the topic of interest, e.g., Fibromyalgia³, and extract the first 40 URLs returned by the query⁴. These URLs stored in a *ToCrawl* list.

The crawler iterates over this *ToCrawl* list, taking the first URL from the list, fetching the corresponding web page with the Unix *lynx* package⁵, and then removing the URL from *ToCrawl*. We do not fetch the same page twice during the crawl, nor more than 100 pages from the same website.

The textual content of the fetched web page is extracted (by the program *delynx.awk*, see release). The page is roughly divided into sentences (*sentencize.awk*), and sentences with at least three English words in a row are retained (*quickEnglish.awk*). Finally, in order to perform the filtering part of the directed crawl, only those pages which contain one or more patterns found in the *Patterns* file are retained. In our released code, the *Patterns* contains upper and lowercase versions of the topic name (e.g. *Fibromyalgia*, *fibromyalgia*). Retained pages are

¹ URL stands for *Universal Resource Locator*. URLs most

² <http://dmoz.org>. There are almost 4 million URLs indexed in the ODP catalog, tagged with over 1 million categories. It can be used under the Creative Commons Attribution 3.0 Unported licence

³ <https://www.dmoz.org/search?q=Fibromyalgia>

⁴ Code found at <https://www.lri.fr/~ggrefens/GLOBALEX/>

⁵ [https://en.wikipedia.org/wiki/Lynx_\(web_browser\)](https://en.wikipedia.org/wiki/Lynx_(web_browser))

copied into a *GoodText* directory, and the new URLs found in the retained page (by the *delynx.awk* program) are appended to the *ToCrawl* list. Every time one hundred pages are crawled, the *ToCrawl* list is randomly mixed. The crawl ends when a predefined number of retained pages (e.g., 1000) are found. Collecting 1000 pages for a given topic, using the code delivered, takes around 3 hours on the average.

We have crawled text for 158 autoimmune illnesses⁶, and for 266 hobbies⁷, in view of creating taxonomies of terms for each topic (Grefenstette, 2015). Here we will show how to use the distributional semantics tools in *word2vec* to explore these domain-specific corpora.

3. Word2vec

Words that appear in similar contexts are semantically related. This is the Distributional Hypothesis (Harris, 1954; Firth 1957). Implementations of this hypothesis have a long history computational linguistics. To find semantically similar nouns using parsed context, Hindle (1990) compared nouns using their frequencies as arguments of verbs as context for comparison, and Ruge (1991) used the frequencies of other words in noun phrases. Frequency of other syntactic relations were used later (Grefenstette, 1994; Lin, 1998), including frequency of appearance in the same lists (Kilgarriff *et al.*, 2004).

In one of the earliest approaches to embedding words in a reduced, fixed-length semantic space, Latent Semantic Indexing (Deerwester *et al.*, 1990) first represented each word by a vector in which each cell value corresponded to the number of times a word appears in a document in some collection. The number of documents in the corpus defined the length of the initial vector. A matrix compression technique, singular value decomposition, allowed them to replace the original word vectors by much shorter, fixed-length vectors (for example, vectors of 100 dimensions). These shorter vectors, or *embeddings* as they are often called now, can be used to recreate the original larger vector with minimal loss of information. As a secondary effect, words whose embeddings are close together, using a cosine measure, for example, to measure the distance, have been found to be semantically similar, as if the singular value matrix reduction mechanism captures some type of “latent semantics.”

Word2vec (Mikolov *et al.*, 2013) and *GloVe* (Pennington *et al.*, 2014) are two recent tools, among many others (Yin and Schütze, 2015), for creating word embeddings. In *word2vec*, using the *continuous bag of words* setting, word embedding vectors are created by a neural net which tries to guess which word appears in the middle of a context (for example, given the four words preceding and following the word to guess). Using another setting *skip-grams*, the neural net tries to predict the words that appear around a given word. In either case, initial, random word embeddings are gradually altered by the gradient descent mechanism of neural nets, until a stable set is found. Levy and Goldberg (2014) have proved that, with a large number of dimensions in the embedding vectors, and enough iterations, *word2vec* approximates Pointwise Mutual

Information (Church and Hanks, 1989; Tunery and Pantel, 2010). *Word2vec* produces “better” results, since it implements other hyperparameters such as generating negative contextual examples, which push unrelated vectors farther apart, and sampling among the positive examples, ignoring some cases, which helps to generalize the vectors since they are not limited to exact contexts (Levy *et al.*, 2015).

Word2vec is memory-efficient and easy-to-use. The code is downloadable from <https://code.google.com/p/word2vec/> and it includes scripts for running a number of large scale examples, out of the box. For example, a *word2vec* script called *demo-word.sh* will download the first 17 million words of Wikipedia and create short embedded vectors for the 71,000 words appearing 5 times or more, in under fifteen minutes on a laptop computer.

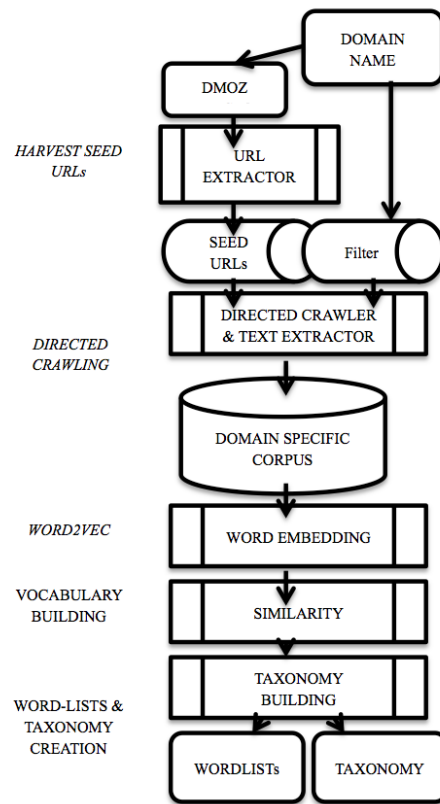


Figure 1. The structure of our approach, involving a directed crawler to gather text in a given domain, and the use of distributional semantics tool to create the characteristic vocabulary and domain taxonomy.

4. Combining a directed crawl and word2vec

Once a domain specific corpus has been crawled (section 2), *word2vec* can be applied to create fixed size word vectors. The input corpus can be transformed by removing all alphanumeric characters, and transposing uppercase characters to lowercase. This is the case of demo programs delivered in the *word2vec* packages, where, in addition ,all numbers are spelled out as digits (e.g., 19 is written as “one nine”) before the word

⁶ <http://www.aarda.org/research-report/> Crawling the 158 topics took about 2 weeks using one computer.

⁷ https://en.wikipedia.org/wiki/List_of_hobbies

embedding vectors are trained. Once the vectors are built, one can find the closest words to any word using the *distance* program in the package. For example, using word vectors built from a 750,000 word corpus for fibromyalgia, we find the following words closest to *Fibromyalgia*. The closest the cosine distance is to one, the nearer are the words:

Nearest words to Fibromyalgia	Cosine distance
pain	0.573297
symptoms	0.571838
fatigue	0.545525
chronic	0.542895
mysterious	0.517179
fms	0.514373
syndrome	0.514127
cached	0.508570
treatment	0.505819
georgia	0.495497
cfs	0.492857
overview	0.492563
referrals	0.491843
diet	0.487120
condition	0.485280
specialists	0.470644
mcgee	0.467879
comprehensive	0.462546
chronicfatigue	0.462226
fibro	0.459657
constellation	0.459147
perplexing	0.454235
checklist	0.441451
pinpoint	0.441292
webmd	0.441237
controversial	0.440630
conditions	0.438186
fm	0.437467

Fibromyalgia is “a rheumatic condition characterized by muscular or musculoskeletal pain with stiffness and localized tenderness at specific points on the body” and many of the words identified by *word2vec* concern its symptoms (*pain, fatigue, , constellation [of symptoms]*) or synonyms (*fibro, fms, chronic-fatigue, fm*) or its characteristics (*mysterious, chronic, perplexing, controversial*) or its treatment (*treatment, referrals, specialists, webmd, diet*). In order to expand this list, we can

find the closest words to each of the 10 most frequent words of length 6 or more:

acceptance, accompanying, aerobic, ailment, amen, anger, anxiety, approach, approaches, appt, arthritic, arthritis-related, biking, bipolar, bloggers, blogspot, brochure, cached, care, cat, cause, causes, celiac, cfs, characterized, cherokeebillie, chronic, clinically, com, common, comprehensive, concurrent, condition, conditioning, conditions, conducted, considerable, constellation, contributing, cortisol, costochondritis, cycles, degenerative, dementia, depressive, dermatomyositis, discomfort, discusses, disease, diseases, disorder, disorders, disturbance, doc, docs, doctors, documentary, dysthymia, ehlers-danlos, elevated, emedicine, emotions, encephalomyelitis, endocrinologist, everydayhealth, excluded, exercises, exercising, exertion, existing, experiencing, expertise, explanations, extent, fatigue, fetus, fibromyalgia, finance, fiona, fischer, flexibility, flu-like, fms, fmsni, focused, frontiers, funding, georgia, guardian, hallmark, hashimoto, hashimotos, healthcare, health-care, homocysteine, hyperthyroidism, hypothyroidism, hypothyroidmom, ... , situations, someecards, sought, specialist, sponsors, statistics, stretching, studies, study, subjective, substantial, sufferers, surrounding, swimming, symptomatic, symptoms, syndrome, syndromes, temporary, testosterone, therapy, transforming, treatment, treatments, truths, tsh, underactive, undiagnosed, unrefreshing, valuable, variant, walking, warranty, wealth, websites, wellness, widespread, worsen

To demonstrate that it is better to use *word2vec* with a domain specific corpus, rather than a general corpus, consider Tables 1 and 2. In these tables, we compare the closest words found to “pain” and to “examination” in two general corpora, a 10 billion word newspaper corpus, and 17 million word Wikipedia corpus, to 9 domain specific corpora concerning illnesses gathered using the directed crawler of section 2. We see that in the domain specific corpora, the words related to pain are tailored to each illness, whereas the general corpora give words related to pain over a variety of situations. Likewise, for “examination”, we can guess from the closest words, what type of medical examinations are used for each illness, whereas the general corpora confuse the academic and judicial senses of “examination” with any medical senses.

Google News (10 billion words)	First 17 million words Wikipedia	Domain specific corpora (each about 250k words)								
		Hypogammaglobulinemia	Vitiligo	Psoriasis	Vasculitis	Uveitis	Neutropenia	Scleroderma	Lupus	Myositis
discomfort	neuropathic	nausea	fever	swelling	joint	redness	relief	stiffness	joint	tenderness
chronic_pain	nausea	headache	stomach	stiffness	sleeping	tenderness	headache	joint	stiffness	stiffness
excruciating_pain	suffering	vomiting	urination	unbearable	stiffness	stiffness	difficulty	physiotherapy	fatigue	aches
ache	palpitations	itching	knee	itch	fatigue	ache	legs	aches	tenderness	chills
arthritic_pain	headaches	stiffness	vision	stiff	muscle	photophobia	shortness	tiredness	complaints	pains
agony	analgesia	flushing	ulcers	joint	aching	ibuprofen	asthenia	relief	aching	malaise
soreness	discomfort	chills	decreased	abdominal	tingling	painkillers	epistaxis	appetite	pains	fatigue
throbbing_pain	itching	sweats	tooth	weakness	weakness	symptoms	abdominal	shoulder	spasms	redness
dull_ache	convulsions	headaches	teeth	joints	myofascial	blurring	chills	swelling	muscle	cramping
numbness	ailments	weakness	discolored	vision	shoulders	fatigue	fatigue	mood	swelling	anorexia
anxiety	vomiting	dizziness	chest	redness	muscles	spasms	appetite	mobility	fevers	complaint

compartmental_syndrome	insomnia	malaise	feeling	botox	diarrhea	pains	weakness	strength	fever	complain
burning_sensation	anesthesia	dyspnea	redness	intense	relieve	motion	breath	subacromial	shortness	joint
Muscle_spasms	headache	swelling	checker	itching	shortness	sensitivity	edema	exercises	ligaments	aching
aches	fibromyalgia	rashes	thickening	headache	appetite	blurred	malaise	tenderness	weakness	swelling

Table 1 Words closest to the word "pain", using *word2vec* to generate embedded word vectors from different corpora. The first two columns use word vectors from 100 billion words of newspaper text (Google News), and 17 million words of Wikipedia text, the remaining 9 columns correspond to smaller corpora created by directed crawling. The first two corpora give general, wide-ranging type of pain. The domain specific corpora restrict type of pain to the specified illness.

Google News (10 billion words)	First 17 million words Wikipedia	Domain specific corpora (each about 250k words)								
		Hypogammaglobulinemia	Vitiligo	Psoriasis	Vasculitis	Uveitis	Neutropenia	Scleroderma	Lupus	Myositis
examinations	examinations	revealed	wood	suspect	exam	slit-lamp	aspirate	exam	laboratory	reveal
exam	histological	physical	suspect	determine	physical	biomicroscope	findings	tests	exam	distinguish
Examination	baccalaureate	sample	uveitis	diagnosing	piece	reveals	physical	ekg	measurement	careful
evaluation	electromyograph	biopsy	physical	determining	histopathological	physical	aspiration	history	evaluation	confirm
thorough_examination	autopsy	radiograph	exam	examining	radiological	establishing	investigations	perform	absence	exam
exams	study	duodenal	rule	checking	revealed	evaluation	examinations	microscope	microscopic	differentiating
inspection	studies	findings	tests	imaging	work-up	revealed	biopsy	changes	biopsy	electrophysiology
dissection	exam	exam	eye	recognize	removal	accomplished	gross	physical	tests	evaluation
medico_legal_exam	exams	stool	insufficient	physical	conduct	fundus	exam	confirm	physical	specimen
forensic_examination	biopsy	specimen	closed	suspected	specimens	exam	tender	ultrasound	x-ray	radiography
assessment	screening	showed	existence	proper	examine	findings	workup	reveal	microscope	scans
postmortem	procedure	adenopathies	identifying	confirmation	examined	lamp	careful	sensitive	urinalysis	ultrasound
polygraphic_test	tests	mediastinal	perform	dosing	interventional	ophthalmoscopy	specimen	assessed	electrolytes	tomographic
examined	accreditation	examinations	trauma	uncertainty	specimen	biomicroscopy	diagnostically	dimensions	ultrasound	histopathology
microscopic_examination	coursework	perform	qualified	biopsy	confirmation	tessier	smear	definitive	repeated	electromyography

Table 2 Words closest to the word "examination", using *word2vec* to generate embedded word vectors from different corpora. The first two columns use word vectors from 100 billion words of newspaper text (Google News), and 17 million words of Wikipedia text, the remaining 9 columns correspond to smaller corpora created by directed crawling. The first two corpora give criminal, newsworthy types of "examination". The domain specific corpora restrict type of pain to the specified illness. Words sorted by nearness to "examination"

5. Conclusion

In this paper, we explain how we created a directed crawler (code released with publication, see also) that gathers domain-specific text, using open source tools, and also demonstrate how the collected corpus can be exploited by word2vec to discover the basic vocabulary for a given domain.

5.1 Acknowledgments

This work is supported by an Advanced Researcher grant from Inria.

6. References

Caruso, Valeria. "Online specialised dictionaries: a critical survey." In *Electronic lexicography in the 21st century: New Applications for New Users: Proceedings of eLex 2011*, Bled, 10-12 November 2011, pp. 66-75. 2011.

Church, Kenneth and Patrick Hanks. "Word association norms, mutual information, and lexicography." In Proceedings of the 27th ACL, pp. 76-83. 1989

Deerwester, Scott; Susan T. Dumais; George W. Furnas; Thomas K. Landauer; and Richard Harshman. "Indexing by latent semantic analysis". JASIS 41:391-407. 1990

Demartini, Gianluca. "Finding Experts Using Wikipedia." FEWS 290: 33-41. 2007. Chakrabarti, Soumen, Martin Van den Berg, and Byron Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." *Computer Networks* 31, no. 11: 1623-1640. 1999.

Firth, John Rupert. *Papers in linguistics*, 1934-1951. Oxford University Press, 1957.

Goldberg, Yoav, and Omer Levy. "Word2vec explained: Deriving Mikolov et al.'s Negative-sampling Word-embedding Method." arXiv preprint arXiv:1402.3722. 2014.

Gregory Grefenstette. *Explorations in automatic thesaurus discovery*. Kluwer International Series in Engineering and Computer Science, 278, 1994

Grefenstette, Gregory. "Personal Semantics." In *Language Production, Cognition, and the Lexicon*, pp. 203-219. Springer International Publishing, 2015.

Harris, Zellig S. "Distributional structure." *Word* 10, no. 2-3: 146-162. 1954.

Hindle, Donald. "Noun classification from predicate-argument structures". In Proceedings of the 28th Annual Meeting of the ACL, 118-125. 1990.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. "The Sketch Engine". In Proceedings of Euralex.

2004

Lee, Jung-Hyun, Jongwoo Ha, Jin-Yong Jung, and Sangkeun Lee. "Semantic contextual advertising based on the open directory project." *ACM Transactions on the Web (TWEB)* 7, no. 4 : 24. 2013.

Levy, Omer and Yoav Goldberg. "Neural Word Embeddings as Implicit Matrix Factorization." *Proceedings of NIPS*. 2014

Levy, Omer, Yoav Goldberg, and Israel Ramat-Gan. "Linguistic regularities in sparse and explicit word representations." In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Baltimore, Maryland, USA, June. Association for Computational Linguistics. 2014

Levy, Omer, Yoav Goldberg, and Ido Dagan. "Improving distributional similarity with lessons learned from word embeddings." *Transactions of the Association for Computational Linguistics* 3: 211-225. 2015.

Lin, Dekang. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*, pp. 768-774, 1998

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In *NIPS*, pp. 3111-9. 2013

Osiński, Stanislaw, and Dawid Weiss. "Conceptual clustering using lingo algorithm: Evaluation on open directory project data." In *Intelligent Information Processing and Web Mining*, pp. 369-377. Springer, Berlin, 2004.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning "GloVe: Global Vectors for Word Representation". In: *Proceedings of EMNLP 2014*. pp. 1532-1543. 2014.

Ruge, Gerda. "Experiments on linguistically based term associations." In *RIA0'91*, pp.528--545, Barcelona. CID, Paris. 1991

Ševa, Jurica, Markus Schatten, and Petra Grd. "Open Directory Project based universal taxonomy for Personalization of Online (Re) sources." *Expert Systems with Applications* 42, no. 17: 6306-6314. 2015.

Turney, Peter D. and Patrick Pantel. "From frequency to meaning: vector space models of semantics". *Journal of Artificial Intelligence Research*, 37(1):141-188. 2010.

Yin, Wenpeng and Hinrich Schütze. "Learning Word Meta-Embeddings by Using Ensembles of Embedding Sets". In: *eprint: 1508.04257*. 2015

Semi-automatic Compilation of the Dictionary of Bulgarian Multiword Expressions

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetlozara Leseva

Department of Computational Linguistics
Institute for Bulgarian Language, Bulgarian Academy of Sciences
{svetla,iva,maria,zarka}@dcl.bas.bg

Abstract

The paper presents the Dictionary of Bulgarian Multiword Expressions. We outline the main features of Bulgarian MWEs, their description and classification based on morphosyntactic, structural and semantic criteria. Further, we discuss the organisation of the Dictionary and the components of the description of the MWEs, as well as the links to other lexicographic and general language resources. Finally, we present the semi-automatic procedures for the compilation of the MWE entries. The work on the Dictionary is ongoing.

Keywords: multiword expressions, automatic compilation, lexicographic resources

1. Introduction

Modern linguistic theory and lexicographic practice emphasise on the importance of MWEs. For instance, they have been estimated to represent a substantial portion (41% of the literals) of the Princeton WordNet 1.7 (Sag et al., 2002). Other scholars propose that multiword expressions (MWEs) are quantitatively equivalent to simple words (Jackendoff, 1997) or even suggest that the number of MWEs is much more prevalent than the number of single words (Melčuk, 1998). This makes the systematic description of MWEs an important task.

There are two main approaches to the representation of multiword expressions in dictionaries – they are either included in the lexical entry of one (or more) of their components, e.g. the headword (the general practice), or represented as individual lexical entries (rarely). The practice followed in the Bulgarian explanatory dictionaries is for MWEs to be listed (not consistently) as a part of the dictionary entry of one of the words they are made up of, which need not be the head word. The second approach is adopted in the Bulgarian specialised phraseological dictionaries (Nicheva et al., 1974; Nicheva et al., 1975; Ankova-Nicheva, 1993), among others.

While the traditional dictionaries provide a rich basis for the description of MWEs, the MWEs included in them are not clearly marked for the type of category they belong to according to the adopted classification in the respective dictionary, and especially for idioms, there are no clear-cut criteria for the choice of the canonical form and the word order of their components.

Therefore, even though the existing Bulgarian dictionaries provide a starting point for the creation of a comprehensive dictionary of MWEs, the exhaustive description of MWEs and their grammatically correct forms remains a complex task that requires other methods and classifications. At the same time, the proper treatment of MWEs is important not only in lexicography, but also in many NLP applications, such as information retrieval, question answering, sentiment analysis, automatic summarisation.

The work discussed here is based on a repository of 86,373

nominal and verbal MWEs extracted from various lexical resources. By 'nominal' and 'verbal' we mean MWEs with a nominal or a verbal head, respectively. Based on the repository, we have constructed an Inflectional Dictionary of Bulgarian Nominal and Verbal MWEs comprising 21,782 nominal MWEs and 24,201 verbal MWEs, out of which 2,345 subject-verb constructions and 21,856 proper verbal MWEs. We have developed a set of tools for their semi-automatic inflectional classification. The inflectional description is subsequently subject to manual validation and supplementation with the help of a dedicated tool. The work on the Dictionary is ongoing – both in terms of the processing of new MWE entries from the repository and their inclusion in the Dictionary and in terms of the manual verification of the lexicographic description.

2. Related Work

Recent research into MWEs focuses on verbal and other MWEs and the description of their components and structure (Villavicencio et al., 2004; Gregoire, 2010). Different approaches for generation of MWE forms have been proposed, such as the parametrised Equivalence Class Method of the DUELME database (Francopoulo, 2013); the graph-based morphosyntactic generator of Multiflex (Savary, 2009) which combines simple words morphology and MWE forms generation, and linear string description in the POLENG formalism (Gralinski et al., 2010).

Nonetheless, the challenges in the lexicographic description of MWEs posed by morphologically rich languages have not been completely addressed yet. These include: rich inventories of synthetic and analytical verb forms with a complex word order, flexible word order of verbal MWEs, structural features, such as mandatory components, discontinuous components, etc.

With respect to Bulgarian, a framework for the morphosyntactic description of MWEs has been proposed by Koeva (2006). Subsequently, it was partially incorporated in the classification and description of a large database of Bulgarian MWEs (Stoyanova and Todorova, 2014). Building upon these proposals, the work presented here represents a systematic effort towards a uniform grammatical description of

MWEs in Bulgarian (in compliance with the description of single words) covering various POS, components' word order and variations, types of syntactic structure, and features, with a view to their automatic recognition and annotation.

3. Complex Description of MWEs in Bulgarian

3.1. Components of the description of MWEs

The description of MWEs, such as the ones proposed by Nunberg et al. (1994), Baldwin et al. (2003), among others, deal with the restrictions imposed on the internal structure, syntactic behaviour and semantic properties of MWEs, which affect significantly their linguistic annotation and automatic processing.

In our approach the description of lexical entries contains the following information:

(a) **Lemma.**

The lemma of each MWE is represented by a canonical form of the MWE. There are two possible approaches to defining a MWE lemma (Savary, 2008) – to define an abstract lemma (a sequence of the lemmas of the components), or to use a non-abstract lemma which is the most neutral existing form of the MWE. We employ the latter approach in the description of the entries. This makes the Dictionary of Bulgarian MWEs more human-friendly and applicable not only for NLP tasks.

A set of rules were implemented in order to ensure consistency in lemma representation, so that each MWE is identified with exactly one lemma (and canonical form) and the encoding of duplicate entries is avoided. This is particularly relevant for MWEs which allow different word order, such as most of the verbal MWEs.

(b) **Morphosyntactic properties.**

The morphosyntactic characteristics of the MWEs are inherited from the head of the phrase. They determine the set of forms that may be realised (i.e. if the head is a neuter noun, the maximum number of forms is four, but if it is a masculine noun – the maximum number of forms is six). The morphosyntactic properties of the components may also determine the set of paradigmatic forms if the components inflect independently of the head (i.e. nominal components of a verbal MWE). The description also needs to reflect any idiomatic restrictions on the paradigm of the MWE.

(c) **Structural properties.**

The description includes the phrasal structure of MWEs: the list of components with their specific morphosyntactic properties. The structural description also needs to define the order of the components and any possible variations, the slots for mandatory, yet variable, arguments and the possible modifiers of each component.

(d) **Semantic properties.**

The MWEs are classified into categories with respect to the degree of their idiomaticity, or decomposability of the meaning.

Nunberg et al. (1994, 497-498) describe idiomaticity as the main characteristic of idioms and outline the following essential semantic properties: conventionality (the discrepancy between the idiomatic meaning and the literal meaning of the phrase), opacity (the ease with which the motivation for the use of a particular idiom can be recovered), and compositionality (the degree to which the idiom's meaning can be analysed in terms of the contribution of its constituents).

Sag et al. (2002) define decomposability as the degree to which the semantics of a MWE can be ascribed to those of its parts. Based on the decomposability, Baldwin et al. (2003) distinguish between non-decomposable (semantically impenetrable), idiosyncratically decomposable (at least some of the components take semantics unavailable outside the MWE) and simple decomposable MWEs (decomposing into simplex senses and generally displaying high syntactic variability).

The degree of semantic decomposability although not reflected directly in the inflection type of MWEs, influences the morphosyntactic variability and flexibility, and hence affects its paradigm. In general, simple decomposable MWEs tend to exhibit much more variation than non-decomposable MWEs in terms of: (a) the number of forms of the MWE; (b) inclusion of external modifiers of particular components; and/or (c) word order.

At present, the semantic description and classification of vMWEs is still in progress.

3.2. Inflection types

According to Koeva (2006) the dictionary description of MWEs needs to include: (a) categorial information, clustering MWEs into grammatical classes (i.e. nouns, verbs, etc.); (b) paradigmatic information, describing the number and types of components and their significant morphosyntactic categories and grouping the MWEs in grammatical subclasses; and (c) word formation alternations, word order, dependencies between components (i.e. agreement), classifying MWEs into grammatical types.

Based on the above, in the context of the Dictionary of Bulgarian MWEs, we define the inflectional type of MWEs to be a complex description defining their morphosyntactic properties, structure, paradigm and variations.

The main criteria to determine the system of inflection types include:

(a) **Morphosyntactic classes and subclasses.**

MWEs are divided into nominal (nMWEs), verbal (vMWEs), adverbial, etc. The analysis presented here is focused on the two largest groups of MWEs – nominal and verbal. The Dictionary also includes a limited number of closed class MWEs such as prepositions and conjunctions.

The morphosyntactic class determines the members of the grammatical paradigm of a given MWE as a whole, e.g. changes according to person, number, tense, etc. of vMWEs, and according to gender, number and definiteness for nMWEs (unless further restrictions apply). For instance, only part of the non-finite forms of a verb participate in the formation of verb forms (in the perfect tenses and the passive voice) while the remaining part of the participles are used in an adjectival function. Therefore, the former but not the latter may be (unless some restrictions apply) part of the paradigm of a vMWE.

Further, the forms of a MWE's paradigm may be restricted to certain morphosyntactic features of the head or of the dependents. For instance, nominal components (NPs) of some vMWEs may have restrictions with respect to definiteness and/or number, e.g. *ri-tam/V kambanata/N* (kick bell.DEF – 'to kick the bucket').

Another factor that determines the MWE's paradigm are some of the lexicogrammatical and/or morphological characteristics of MWEs, such as the gender (lexicogrammatical feature) and number (morphological feature) of the nominal head of nMWEs which impose agreement constraints on any component of the MWE that agrees in gender and number with nouns when forming a phrase.

(b) **Structural types and subtypes.**

The number and morphosyntactic types of components determine the main structure and variations of the MWE forms. Some phrase components (NP) are defined independently and the inflection types use a combination of these definitions to define the structure and the paradigm of the MWE.

(c) **Mandatory arguments and/or modifiers that are not lexicalised (variable slots).**

Some MWEs take an argument or a modifier which is mandatory, not lexicalised and restricted to a particular semantic class (i.e. person) – any word from the class can take the slot.

(d) **External (though syntactically integrated) modifiers and/or adjuncts.**

While external modifiers are not part of a MWE, they are syntactically integrated in its phrase structure so their identification is important for the proper recognition of the MWEs. As in many cases, especially with light-verb constructions, modifiers may be an open class of words and phrases which depend on the semantics and selectional restrictions of the noun.

(d) **Morphophonemic variations in the paradigm of the MWE.**

In Bulgarian, as a morphologically rich language, it is essential for the definition of the full paradigm of the MWE to cover all the possible morphophonemic variations resulting from various phonological phenomena. These substantially increase the numbers of in-

flexion types. In our approach, we classify the inflection types from the most general features to the more specific, thus morphophonemic variations are the most fine-grained classification feature.

(f) **Possible word order variations.**

Word-order variations have to do with the identification of the components and the boundaries of MWEs and is also addressed in the description of the MWEs by specifying the possible reordering of the MWE's components as compared with the neutral word order in the dictionary entry.

4. Classification of MWEs in Bulgarian

While the approach and classification aim at embracing all kinds of MWEs, the majority of entries the Dictionary currently includes belong to nMWEs and vMWEs, which is why henceforth we focus on these two types of MWEs. Moreover, these are the categories which exhibit the most variation and pose major problems for natural language processing.

4.1. Nominal MWEs

Besides the degree of their semantic decomposability, nMWEs are also divided into proper nouns (named entities, NEs) and common nouns. This distinction is important due to the specific properties of NEs – they often have restrictions in the paradigm, do not allow modifiers, have specific orthographic representation (capital letters), etc.

The system of inflectional types of nMWEs is based on the following features:

- Morphosyntactic properties of the head noun and (in some cases) of a noun, part of the prepositional modifier;
- Syntactic structure – number and type of components;
- Insertion of clitics (possessive and interrogative);
- External modifiers (relatively rare);
- Morphophonemic variations;
- Word order variations (rare).

Nominal MWEs are divided into three main structural categories depending on the modifiers – either taking an adjective modifier, a prepositional phrase modifier, or a combination of the two:

- A N – *byala mechka* 'polar bear';
- A A N – *bruten vatreshen produkt* 'Gross Domestic Product';
- N P N – *More na spokoystviето* 'Sea of Tranquility';
- N P A N – *Ministerstvo na vatreshnite raboti* 'Ministry of Internal Affairs';
- A N P N – *Balgarska akademiya na naukite* 'Bulgarian Academy of Sciences';

- A N P A N – *Konsultativen savet za natsionalna sigurnost* ‘Consultation Council for National Security’;
- Others – there are some less frequent types such as N N (*kashta muzey* ‘museum house’), N P N Conj N (*Ministerstvo na obrazovanieto i naukata* ‘Ministry of Education and Science’), etc.

The above structural types comprise a total of 81 inflectional types built by different combinations of the idiomatic realisation of grammatical categories of nouns and their modifiers. For the basic inflection types relevant are noun classes defined by noun gender (lexicogrammatical), number and definiteness, features such as +human/-human, as well as the lexicogrammatical and grammatical categories of the head and the modifiers.

Typically, nominal MWEs have rigid word order and most of them do not allow variations in the constituents’ linear order or omission of components. Optional elements (modifiers) can be added in some cases. The possibilities are encoded in the inflection type of the MWE (Example 1).

Example 1. Nominal MWE.

Lemma: *kiselo mlyako* ‘yoghurt’
 Type: NP_AN
 Subtype: NN3
 Modifiers: Comp1:NoMod, Comp2:NoMod
 Spaces: Space1:InsCl
 Word order: 1-2

Example 1 shows a nominal MWE with a neuter common noun head. The main structural type is NP_AN (a noun with a single adjective modifier), and the subtype NN3 indicates that the MWE has a full paradigm including the following forms: singular – definite/indefinite; plural definite/indefinite. The description also points out that neither the first nor the second component allows a modifier. The information about the spaces shows that there is only one slot and it can be filled (possibly) with clitics – possessive, interrogative or a combination of both. Word order shifts are not allowed, i.e. the word order is fixed.

We adopt the following general classification of nominal MWEs in Bulgarian with regards to their inflectional description:

- N1. Non-decomposable/semi-decomposable nMWEs with invariable components – predominantly non-descriptive named entities (NEs which do not contain descriptors in their regular meaning) and frozen expressions:
 NID1-AN *Byala/A cherkva/N* (‘(town of) Byala cherkva’);
 NID1-NPP *solta/N na/P zemyata/N* (‘the salt of the earth’).
- N2. Non-decomposable/semi-decomposable nMWEs which have a limited or a full paradigm:
 NID2-AN *bradat/A lishey/N* (bearded lichen ‘Usnea’);
 NID2-NPP *Bryag/N na/P slonovata/A kost/N* (‘(Republic of) Ivory Coast’).
- N3. Decomposable nMWEs which have a limited paradigm – predominantly descriptive NEs:

NID3-AN *Evropeyska/A komisiya/N* (‘European commission’);
 NID3-NAN *ski/N alpiyski/A distsiplini/N* (ski alpine disciplines ‘alpine skiing’);
 NID3-NPP *yaytsa/N na/P ochi/N* (‘fried eggs’).

- N4. Decomposable nMWEs which have a full paradigm:
 NID4-AN *poshtenska/A kutiya/N* (‘postbox’);
 NID4-NN *kashta/N muzey/N* (house museum ‘museum house’);
 NID4-NPP *pasta/N za/P zabi/N* (paste for teeth ‘toothpaste’).

The above categories can optionally have a variable slot and/or allow the insertions of modifiers:

- A. Modifiers (X_mYP is labelling the possible modifier YP to the component X of the MWE):

NID5-P N_mAP *glava/N na/P (golyamo/A) semeystvo/N* (‘head of a (big) family’).

No examples were found with inserted modifier to the noun in an adjective–noun construction. If these exist, they will be extremely rare, since the MWE denotes a single concept, thus its components are more closely semantically connected than any external modifier.

- B. Variable slots

Possessive slot within the NP/PP (X_xYP is labelling the possible variable slot YP):

NID6-N_xPP *balsam/V za/P ranite/N na/P + PER-SON* ((literally) balsam for the open wounds (of someone) ‘a remedy for the problems of someone’). The possessive slot can also be filled by a short possessive pronoun *balsam/V za/P ranite/N mi/PronPoss* (my wounds) or an indefinite pronoun *balsam/V za/P nechii/PronIndefPoss rani/N*.

Out of the 59,369 nominal MWEs in the repository, so far we have processed 21,782 MWEs, mostly named entities (geographical names, events, botanical and zoological species).

4.2. Verbal MWEs

The system of inflectional types of vMWEs is based on the following features:

- Morphosyntactic properties of the head verb and (if relevant) of the nominal components;
- Syntactic structure – number and type of components;
- Morphophonemic variations;
- Variable slots for arguments of the MWE;
- Word order variations;
- Insertion of external elements – clitics (personal and possessive pronouns, and interrogative clitic), adverbs, phrases.

The relevant lexicogrammatical and morphosyntactic features of the verbal head include the following:

- (i) personality – Bulgarian verbs fall into the following categories: personal verbs (ones having a full paradigm for the 1st, 2nd and 3rd person); impersonal verbs (verbs that cannot take a subject); 3rd-personal verbs (ones having forms only for the 3rd person); verbs with plural forms only;
- (ii) transitivity – verbs are transitive or intransitive;
- (iii) aspect – perfective or imperfective (the division counts on the inflectional paradigms of the simple words only) (Koeva, 2004).

These properties are inherited by the head verb, except where the idiomatic meaning places restrictions or requires a particular form, e.g. in subject-verb MWEs verbs are only in the 3rd person – *vezdata mi izgryava* ((literally) my star rises, 'to achieve great success'). The personality and transitivity features of a verbal MWE determine the number of its forms and its potential transformations – e.g., only transitive verbs form passives, or object-involving nominalisations, e.g. *vdigna garda* ('raise one's guard') > *vdignat gard* ('raised guard'). The morphological categories – number, person, tense, polarity, mood, voice, and gender and number for non-finite verb forms – may also be constrained by the idiomatic meaning, e.g. *ne iskam akal nazaem* ('to not want unsolicited advice') is restricted to the negative form of the verb.

Example 2.

Lemma:	<i>ne iskam akal nazaem</i> 'to not need unsolicited advice'	
Type:	VP_NP_AdvP	
Subtype:	V_LIT_neg NP_N_NM0 Adv	
Modifications:	Comp1:NoMod; Comp2:NoMod; Comp3:Mod; Comp4:NoMod	
Variable slots:	0	
Spaces:	Space1:InsCl;	Space2:InsAll;
	Space3:InsAll	
Word order:	1-2-3-4; 3-1-2-4; 4-3-1-2; 3-4-1-2	

The above description shows that the MWE is a verb phrase with a transitive verb head in negative form. The nominal component is a masculine singular noun in the indefinite form and it is followed by an adverb. The information about possible modifications shows that modifiers can be assigned to the third component (the noun, e.g. *ne iskam mnogo akal nazaem* – 'to not want a lot of unsolicited advice'). The MWE has no variable slots. The 'Spaces' field shows which elements can be inserted between the components of the MWE. Several word order patterns are possible, and the options are listed starting with the neutral word order.

The Dictionary of Bulgarian MWEs includes 24,201 verbal MWEs: 2,345 subject-verb constructions and 21,856 verbal MWEs.

The division of vMWEs in terms of decomposability represents a detailed classification whose main categories are aligned with the ones proposed within the PARSEME Shared Task on automatic detection of verbal MWEs¹, in

particular: lexicalised pronominal verbs (IPronV), lexicalised combinations of a verb and a preposition (IPrepV) light verb constructions (LVCs), idioms (IDs) and others (OTHs). The last category comprises vMWEs that do not fall under any of the former classes.

Below we present a core classification of verbal MWEs in Bulgarian reflecting their inflectional description:

V1. Lexicalised pronominal verbs.

This category includes verbs with reflexive and reciprocal particles and dative and accusative pronominal clitics:

smeya/V se (laugh REFL.ACC);
vaobrazyavam/V si (imagine REFL.DAT);
marzi/V me (is_lazy me.ACC 'I am lazy', (literally) 'it feels lazy to me');
hrumne/V mi (occurs me.DAT), 'it occurs to me';
gadi/V mi se (feels_sick me.DAT REFL.ACC 'I feel sick', (literally) 'it feels sick to me').

V2. Lexicalised combinations of verb and preposition – *vyarvam/V v/P* (believe in).

V3. Light verb constructions.

Light verb constructions are vMWEs that share a number of distinctive features. An LVC consists of a semantically bleached verbal head and a lexicalised argument that contains a predicative noun (typically denoting an action or an event), usually a direct object (NP), and more rarely – a PP object or a subject NP (Vincze et al., 2016) – the last option does not apply to Bulgarian.

A characteristic of LVCs is that in most cases they readily take modifiers provided they are semantically compatible with the argument (e.g., *vzemam (trudno/vazhno) reshenie* – to make a (difficult/important) decision).

Below we give examples of LVCs that consist of a verb and a lexicalised argument. LVCs that license variable slots are discussed in the appropriate subsection (B):

- LVC with a lexicalised NP argument (LVC-NP): *cheta/V доклад/N* (read report 'to present/to make a presentation').
- LVC with a lexicalised PP argument (LVC-PP): *izpadam/V v/P panika/N* (fall into panic 'become panic-stricken').

V4. Proper vMWEs with a different degree of idiomaticity (idioms, ID).

This category includes vMWEs that exhibit different degrees of semantic idiomaticity and syntactic flexibility – from invariable components (in such cases the verb is the only part of the MWE that may undergo any changes) to ones whose non-verbal components change according to certain categories and/or allow

konstanz.de/parseme/index.php/events/2-general/142-parseme-shared-task-on-automatic-detection-of-verbal-mwes

¹An event within the PARSEME Network, <http://typo.uni->

external elements to intervene in a regular way in the linear structure of the MWE, and/or allow various systemic word-order variations.

The non-verbal MWE components may either be arguments or adjuncts of the original verb's meaning. They may be lexicalised or may represent a variable argument's slot that needs to be filled in by semantically and syntactically compatible material in the context of use.

- ID1: vMWEs with invariable non-verbal component(s).

This subtype of idioms includes verbal idioms whose non-verbal components do not undergo any morphological changes. The following structural types have been identified in the data. The list is non-exhaustive as new entries may be found in the future that represent other structural types.

– ID1-NP

This structural type includes idioms with a lexicalised NP direct object.

ID1-NP *ritam/V kambanata/N* (kick bell.DEF 'to kick the bucket');

ID1-NP *davam/V zelena/A svetlina/N* (give green light.INDEF 'to give the green light');

– ID1-PP

The idioms subsumed in this structural type lexicalise a PP argument or adjunct of the original verb's meaning, usually with adverbial semantics.

ID1-PP *barkam/V v/P kasata/N* (thrust one's hand into the cash box 'to steal public money');

ID1-PP *stoya/V v/P syanka/N* (stay in the shadows 'remain inconspicuous');

The following structural type includes vMWEs with two lexicalised arguments, either expressed as an NP and a PP, or as two PPs:

– ID1-NPPP *hvashtam/V bika/N za/P rogata/N* ('to take the bull by the horns').

– ID1-PPPP *prochitam/V (neshto) ot/P koritsa/N do/P koritsa/N* ('read (something) from cover to cover').

Other structural types – with an AdvP or an SC (a small clause) – may also be found:

– ID1-NPSC *darzha/V ochite/N si/PRON otvoreni/A* (keep eyes.DET my/your/.POSS open 'to keep one's eyes open');

– ID1-PPSC *kazvam/V na/P chernoto/N byalo/A* ('to call black white');

– ID1-AdvP *izvajdam/V nayave/ADV* (expose in_the_open 'to bring to light').

- ID2: vMWEs with semi-fixed non-verbal components.

This category includes vMWEs whose non-verbal components can change in form without

change in meaning. The cases where the non-verbal components have a full paradigm are very rare. Below are presented several types of semi-fixed non-verbal components found in the Dictionary of Bulgarian MWEs.

– ID2-NP *broya/V zvezdi/N* (count stars.INDEF 'to put on airs'), *broya/V zvezdite/N* (count the stars.DEF 'to put on airs');

– ID2-PP *popadam/V v/P kapan/N* (fall into a trap.INDEF), *popadam/V v/P kapana/N* (fall into the trap.DEF)

In the particular vMWEs the non-verbal component may be used both in an indefinite and a definite form.

– ID2-SC

Another example is presented by vMWEs with an SC. As NP and AP small clauses usually agree in gender and number with the subject or the object, their form changes accordingly (if possible). For instance, in the example below, if the subject has a feminine referent, the form of the adjective will change in the feminine: *be the last.FEM to have the word*, and if the subject is plural, the adjective will take the plural, e.g. *be the last.PL to have the word*:

– ID2-NP_ASC *imam/V dumata/N posleden/A* (be the last to have the word 'to have the last word').

V5. MWEs with a lexicalised subject

Although MWEs consisting of a verb and its lexicalised subject are usually referred to the class of verbal idioms, we consider them to be a separate category because they exhibit specific grammatical features (e.g., agreement between the verb and the subject) and obey certain restrictions on the possible forms and derivations.

- ID-SUBJ: *izliza/V3p mi/DAT ime/N* (appears forme.DAT name 'a name sticks (for/to me)').

V6. MWEs with sentential features

- proverbs – *koyto ne raboti, ne tryabva da yade* (who.DET not work not should to eat, 'He who does not work, neither shall he eat.');

- frozen clausal expressions – *kakto i da e* (as it is 'whatever').

Within each of the above types there are subtypes based on the possible combinations of the following features:

A. Modification of components

Certain vMWEs license modifier slots to one or more of their components. The notation X.mYP marks the possible modifier YP of the component X of the MWE. The possible types illustrated below are non-exhaustive.

- Within the NP: N_mAP *zabarkvam/V (golyama/A) kasha/N* ('to make a (big) mess');
- Within the NP: N_mAdvP *vizhdam/V (mnogo/ADV) zor/N* (see (much) hardship 'to find (s.th.) very tough');
- Within the PP: N_mAP *galtam/V s/P (zhadni/A) ochi/N* (swallow with thirsty eyes 'to take with greedy eyes');
- Within the PP: N_mAdvP *stigam/V do/P (mnogo/ADV) sarca/N* (reach to many hearts 'to reach many hearts');
- Within the AdvP: Adv_AdvP *gledam/V (mnogo/ADV) otvisoko/ADV* (look from_above 'to look down (at s.o.)');
- Within the VP: V_AdvP *vdigam/V letvata/N visoko/ADV* (raise stick.DET high 'to raise the bar high');

B. Variable slots

Along with the lexicalised components which constitute part of the vMWE, a MWE may also license one or more variable modifier or argument slots which need to be filled in order for the vMWE to be semantically interpreted. The notation X_xYP marks the variable slot YP to the component X of the MWE.

- A possessive slot within the NP/PP:
 - ID1-PP_N_xPP(na) *hodya/V po/P nervite/N PPos na/P nyakogo/N* -> *hodya po nechii nervi* (walk on nerves.DET of somebody 'get on s.o.'s nerves');
- An argument slot:
 - * An NP slot:
 - ID1-V_xNP_PP: *prochitam/V nestho mezhdu redovete* ('read **something** between the lines');
 - ID2-V_xNP_PP: *hvashtam/V nyakogo v/P kapan/N* ('catch **someone** in a trap');
 - ID1-V_xNP_AdvP: *gledam/V nyakogo otvisoko* (look down on **someone**);
 - ID1-V_xNP_PP: *darzha/V nyakogo/neshto pod/P kontrol/N* ('keep **somebody/something** under control');
 - ID2-V_xNP_ASC: *izkarvam/V nyakogo chisti/A* (make **someone** clean 'to make s.b. as innocent as a baby unborn').
 - * A PP slot:
 - ID1-V_NP_xPP(na): *podavam/V raka/N na nyakogo* ('lend a hand **to someone**');
 - ID2-V_PP_xPP(v): *vlyubvam se/V do/P ushi/N v nyakogo* ('fall head over heels **with someone**');

C. Possible word order variations

All word order variations are listed as a sequence of components.

ID1-NP: *davam/V zelena/A svetlina/N* (give green

light.INDEF 'to give the green light') can have the following word order variations: 1-2-3 and 2-3-1 (it is not possible to exchange the positions of 'green' and 'light' within the NP).

5. Development of the Dictionary of Bulgarian MWEs

The Dictionary of Bulgarian MWEs described herein uses the data compiled by Stoyanova and Todorova (2014). The MWEs are extracted from different sources – traditional dictionaries, corpora and the Bulgarian wordnet. The total of 86,373 nominal and verbal MWEs were further classified into different grammatical groups². We use these nominal and verbal MWEs as a source repository from which the inflectional description in the Dictionary of Bulgarian MWEs is semi-automatically developed.

The proposed work aims at a uniform and consistent and at the same time flexible representation of different types of MWEs, their main properties and the restrictions they obey.

5.1. Layers of the description

The linguistic description includes several layers of information:

- morphosyntactic categories of the head and components (POS, verb aspect, noun gender, etc.) determining the basic paradigm;
- morphosyntactic constraints imposed by the idiomatic meaning (e.g. in *imam zlatno sartse* 'have a heart of gold' the direct object must be singular indefinite) determining the members of the basic paradigm that are really used;
- structural characteristics – possible mandatory components and variations in their linear order;
- syntactic transformations, such as passivisation;
- subcategorisation – specification of arguments taken by a verbal MWE. Apart from the idiomatic arguments inherited from the original verb's semantics, e.g. *podavam raka (na nyakogo)* 'lend a hand (to someone)' (*na nyakogo* is an argument of *podavam*), there are cases in which an argument is subcategorised by the verbal MWE, e.g. *chupya grab pred nyakogo* 'bend one's back (before so.)'. In the latter case *pred nyakogo* is only licensed by the vMWE and is not an argument of the verb *chupya*.

We define spaces between MWE components as bearing one of the following features (Koeva, 2006):

- no external element can be inserted;
- possibility for insertion from a fixed set of pronominal, interrogative and negative clitics;

²The source is available at <http://dcl.bas.bg/en/parseme-shared-task-phase-1/> and <http://dcl.bas.bg/en/parseme-shared-task-phase-2/>

- (iii) possibility for an optional constituent from a certain class, e.g. prepositional phrase;
- (iv) a relatively free position allowing various insertions.

The positions where insertions are possible are also related to word order combinations since they are likely to represent borders between phrasal components (e.g., NP or PP) within the MWE which can change their order. Possible word order variations are specified as a list of permutations of the components, e.g. 1-2-3, 3-1-2, where the numbers indicate the components' order in the MWE's canonical form.

5.2. Generation of the paradigm

We adopt an approach to the definition of inflectional types in which 'complex' types are defined as sequences of 'elementary' types. An elementary type defines the inflection of an idiomatic MWE complement which can be modelled independently, by encoding the following information: the phrasal category (e.g. NP, PP) and internal syntactic structure (e.g. N, AN), etc.; the syntactic category of the component head, e.g. N(oun), P(reposition), Adv(erb), A(djective); specification of fixed and variable grammatical categories; and agreement between components.

Each inflectional type is assigned a regular expression which defines the generation of all forms from the MWE canonical form. Example 3 shows the regular expression generating the forms of the type NP_AN_NNS1. The first form (indefinite) coincides with the canonical forms (both components stay unchanged <1> <2>), while in the second form (definite) the definite article (-to for the neuter gender) is added to the first component, the adjective.

Example 3. Regular expression for the type NNS1 (neuter gender, singularia tantum, allowing both definite and indefinite forms)

Iskarsko defile,NP_AN_NNS1

NP_AN_NNS1=<1> <2>/sno + <1>to <2>/snd

6. Compilation Procedure

The present Dictionary is compiled semi-automatically from the original source by applying the following procedure.

Step 1. Preprocessing.

The preprocessing includes morphosyntactic analysis of the MWE at component level: tokenisation, POS tagging, lemmatisation of components, and is performed using the Bulgarian Language Processing Chain (Koeva and Genov, 2011).

Step 2. Selection of the lemma of the MWE.

The source contains MWEs presented through their lemma or (for verbal MWEs) through several different forms considered a lemma in different dictionaries.

The MWE lemma is selected using a set of heuristics out of all possible forms of the components of an MWE as they appear in a corpus of texts. We used the Bulgarian National Corpus (BulNC) (Koeva et al.,

2012) as a large representative corpus of Bulgarian. We apply a set of simple heuristics to identify the less marked form (e.g., singular is less marked than plural, present tense is less marked than other tenses, etc.). In order to ensure consistency, we introduced rules for the order of components for the canonical form – verb first, followed by NP, PP, AdvP. Reflexive or other particles are immediately next to the verb (either preceding or following it). In SUBJ-ID constructions, the subject precedes the verb. Exceptions include constructions with restriction in word order, as well as cases where the verb is in one of its analytical verb forms (e.g., the verb is only in the negative form).

The lemma is then verified by an expert. Problematic cases include MWEs which have a non-idiomatic counterpart, while the MWE lemma has a limited paradigm (e.g., *vdigam galabite* 'to raise the pigeons' – 'to go away quickly, to buzz off' and *vdigam galaba v raka* 'to lift the pigeon in a hand').

Step 3. Identification of any relevant information about the subcategorisation.

In order to identify relevant subcategorisation information, we analyse the context of the verbal MWE examples obtained from the BulNC. We use the information for certain prepositions within the range to 3 tokens (currently) from the MWE. E.g., *biya duzpatana* 'to send away'. The main problem is to distinguish between arguments and external phrases which are not part of the subcategorisation, even if they appear frequently with the MWE, e.g. *povdigam vapros na* 'to pose a question at' (at a meeting, a court hearing, etc.).

Further, the subcategorisation information can be extended by analysing the semantic properties of identified arguments, e.g., *biya duzpatana* + PERSON 'send away someone'.

Step 4. Identification of the inflection type.

The identification of the inflection type is performed semi-automatically by the following procedure:

- (a) structural analysis of the MWE;
- (b) lexicogrammatical and morphosyntactic description;
- (c) setting of morphosyntactic category's values using heuristics;
- (d) selection of specific subtypes based on the morphophonemic changes of individual components of the MWE;
- (e) assignment of properties to spaces between components based on their position and general syntax rules.

By analysing the structure of the MWE we identify its main type – nominal, verbal or sentential, as well as its structural subtype within the main category (e.g., V-NP, V-PP, etc.). The lexicogrammatical and morphosyntactic properties of the components as independent words give information about the possible

paradigm, which is defined by a set of parameters for each component – nouns can change in number, definiteness, verbs can change in person, number, tense, etc. However, the paradigm of the MWE does not usually allow full realization of components' word-forms, in fact most components are likely to appear in a frozen form.

By analysing the identified occurrences of the MWE in the BulNC, its paradigm is identified. Here again, it is problematic to distinguish between the occurrence of the MWE as compared to its free phrase counterpart. The analysis of the occurrences also allows to deduce the possible external insertions of elements (clitics or phrases), as well as variations in word order.

The inflection type is composed by the set of the listed parameters (see Example 2).

Step 5. Manual verification of inflection types.

The manual verification is a time-consuming task but is necessary in order to ensure the high quality of the resource. For the purposes of manual verification the tool `FLeGen` for visualisation and editing of inflection types was implemented using `Java`. It presents the inflection type as a set of features where each feature has a number of possible values. After editing, the combination of features and values is verified to avoid inconsistencies.

7. Linking the Dictionary of Bulgarian MWEs with other lexicographic resources

The repository of MWEs is obtained by extraction of appropriate candidates from the following lexical (and general) resources:

- Specialised phraseological dictionaries – appropriate entries are manually identified, namely verbal MWEs were selected among idiomatic expressions, fixed similes, fixed syntactic constructions, and proverbs;
- the Bulgarian wordnet³ – literals containing more than one word are automatically extracted as candidates and manually filtered;
- The Explanatory Dictionary of Bulgarian (Andreychin et al., 1999) – MWEs are listed with their definition in the lexicographic entry for one or more of their components, candidates are then manually filtered;
- The multi-volume Dictionary of Bulgarian (RBE, 1977–2015)⁴ – the same procedure as the Explanatory Dictionary;
- Wikipedia – Wikipedia is automatically crawled, and titles of Wikipedia articles and categories are extracted as candidates for nominal MWEs, these include mostly named entities;

- Bulgarian National Corpus⁵ – MWE candidates are automatically identified using collocation measures, frequency analysis and syntactic filters.

The information about the original source is kept for each MWE and thus we have an entry in the inflectional dictionary linked with a lexical resource which can provide its definition (if the source is a phraseological dictionary with definitions, or the Explanatory Dictionary of Bulgarian), synonyms, hypernyms, hyponyms, and other semantic relations (if the source is the Bulgarian wordnet), etc. For example: The MWE *grancharsko kolelo* (potter's wheel) is linked with the respective definition from an Explanatory Dictionary:

grancharski [pottery] adj. Pertaining to a potter, to the production of pottery. **grancharsko kolelo [potter's wheel]** – a device with a rotating horizontal disk driven by foot movements upon which clay is molded by hand. (Andreychin et al., 1999).

The MWE *himichen element* is linked with its synonym *himicheski element* 'chemical element' in the Bulgarian wordnet, its definition 'any of the more than 100 known substances (of which 92 occur naturally) that cannot be separated into simpler substances and that singly or in combination constitute all matter', its hypernym *substance* and many hyponyms like *nitrogen, N, atomic number 7; nobelium, No, atomic number 102; oxygen, O, atomic number 8; phosphorus, P, atomic number 15*, etc.

Wordnet is a very useful resource for an MWE to be linked to because synonymous sets in different languages are connected with a relation of equivalence to the corresponding English synset. So far, the number of verbal MWEs in the Bulgarian wordnet is relatively small compared to nominal MWEs. Thus, an expansion in the reverse direction might be considered further – the inclusion of MWEs from the Inflectional Dictionary of Bulgarian MWEs into the Bulgarian wordnet.

There are also other resources and repositories of MWEs which can be used as a valuable source of MWEs and description, such as the Phraseological Dictionary on the Infolex portal⁶.

The new Dictionary portal of the Institute for Bulgarian Language⁷ brings together the set of resources of the Institute and can be particularly useful for the analysis of MWEs. It links the Dictionary of Bulgarian, the Grammatical Dictionary of Bulgarian, and the resources from the Infolex portal.

8. Conclusion

The paper focuses on the extensive description and classification of the Bulgarian MWEs which exhibit a wide range of structural, morphosyntactic, and semantic properties. We aim at encompassing all main types of MWEs within the unified framework of the large Dictionary of Bulgarian MWEs. Moreover, the Dictionary offers a unified approach

³<http://dcl.bas.bg/bulnet/>

⁴<http://ibl.bas.bg/rbe/>

⁵<http://search.dcl.bas.bg/>

⁶<http://ibl.bas.bg/infolex/idioms.php>

⁷http://ibl.bas.bg/dictionary_portal/index.php

to the description of verbal and nominal MWEs, which are the largest and most problematic groups of MWEs. For the purposes of the Dictionary we have developed a consistent system of inflectional types, which also incorporate information about possible modifications within the MWE, alternations, subcategorisation, word order variations.

The paper also outlines the procedures for semi-automatic inflectional description of MWEs which can be modified and adapted to other languages with rich morphology. By linking the entries in the Dictionary with the corresponding entries in other lexical resources, the Dictionary provides the opportunity for more detailed description of MWEs and can be used in the analysis of their complex linguistic properties and behaviour.

Further, some machine learning techniques can be investigated and applied using as training data the manually verified MWEs, which can then improve the quality of the automatic description of the MWE entries and thus, reduce the need of manual intervention.

Our future work is focused on extending the Dictionary with new entries and types of MWEs (adjectival, adverbial, prepositional phrases, etc.), as well as with new layers of linguistic description, e.g. semantic information.

9. Bibliographical References

- Andreychin, L., Georgiev, L., Ilchev, S., Kostov, N., Lekov, I., Stoykov, S., and Todorov, T. (1999). *Balgarski talkoven rechnik. Dopalнено i preraboteno izdanie ot D. Popov*. Nauka i izkustvo, Sofia.
- Ankova-Nicheva, K. (1993). *Nov frazeologihen rechnik na balgarskiya ezik (in Bulgarian)*. Universitetsko izdatelstvo Sv. Kliment Ohridski, Sofia.
- Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. ACL.
- Francopoulo, G. (2013). *Lexical Markup Framework*. John Wiley and Sons.
- Gralinski, F., Savary, A., Czerepowicka, M., and Makowiecki, F. (2010). Computational lexicography of multi-word units. how efficient can it be? In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications. Coling, August 2010*, pages 2–10.
- Gregoire, N. (2010). Duelme: a dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44:23–39.
- Jackendoff, R. (1997). The architecture of the language faculty. *Computational Linguistics*, 24.
- Koeva, S. and Genov, A. (2011). Bulgarian language processing chain. In *Proceeding to The Integration of multilingual resources and tools in Web applications Workshop in conjunction with GSCL 2011*. University of Hamburg.
- Koeva, S., Stoyanova, I., Leseva, S., Dekova, R., Dimitrova, T., and Tarpomanova, E. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0(1):65–110.
- Koeva, S. (2004). Sintaktichno i semantichno opisane na dialektite v balgarskiya ezik. In *Kognitivna gramatika na balgarskiya i frenskiya ezik – opisane i formalizatsiya. Balgarsko ezikoznanie*, volume 4, pages 182–232. Academic Press Prof. Marin Drinov.
- Koeva, S. (2006). Inflection morphology of bulgarian multiword expressions. In *Computer Applications in Slavic Studies*, pages 201–216. Boyan Penev Publishing House.
- Melčuk, I. (1998). Collocations and lexical functions. In P. Cowie, editor, *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Oxford: Clarendon Press.
- Nicheva, K., Spasova-Mihaylova, S., and Cholakova, K. (1974). *Frazeologichen rechnik na balgarskiya ezik. Volume 1*. BAS Publishing House.
- Nicheva, K., Spasova-Mihaylova, S., and Cholakova, K. (1975). *Frazeologichen rechnik na balgarskiya ezik. Volume 2*. BAS Publishing House.
- Nunberg, G., Sag, I., and Wasow, T. (1994). Idioms. In Stephen Everson, editor, *Language*, pages 491–538. Cambridge University Press.
- RBE. (1977–2015). *Rechnik na balgarskiya ezik (in Bulgarian)*, volume 1–15. Academic Press Prof. Marin Drinov.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing '02*, pages 1–15. Springer-Verlag.
- Savary, A. (2008). Computational inflection of multiword units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1(2).
- Savary, A. (2009). Multiflex: A multilingual finite-state tool for multi-word units. In S. Maneth, editor, *CIAA. Lecture Notes in Computer Science*, pages 237–240. Springer.
- Stoyanova, I. and Todorova, M. (2014). Razrabotvane na rechnitsi na sastavnite leksikalni edinitsi v balgarskiya ezik za tselite na kompyutarnata lingvistika. In *Ezikovi resursi i tehnologii za balgarski ezik*, pages 185–202. Academic Press Prof. Marin Drinov.
- Villavicencio, A., Copestake, A., Waldron, B., and Lambeau, F. (2004). The lexical encoding of mwes. In T. Tanaka, et al., editors, *Proceedings of the ACL 2004 workshop on multiword expressions: Integrating processing. Barcelona, Spain*, pages 80–87.
- Vincze, V., Savary, A., Candito, M., and Ramisch, C. (2016). Annotation guidelines for the PARSEME shared task on automatic detection of verbal MultiWord Expressions. Version 5.0. <http://typo.uni-konstanz.de/parseme/images/shared-task/guidelines/PARSEME-ST-annotation-guidelines-v5.pdf>.