

D3.8

Lexical-semantic analytics for NLP - final report

Author(s): Federico Martelli (Sapienza), Marco Maru (Sapienza), Cesare Campagnano (Sapienza), Roberto Navigli (Sapienza), Paola Velardi (Sapienza), Rafael-J Ureña-Ruiz, Francesca Frontini (CNR-ILC), Valeria Quochi (CNR-ILC), Jelena Kallas (EKI), Kristina Koppel (EKI), Margit Langemets (EKI), Jesse de Does (INT), Rob Tempelaars (INT), Carole Tiberius (INT), Rute Costa (NOVA CLUNL), Ana Salgado (NOVA CLUNL), Simon Krek (JSI), Jaka Čibej (JSI), Kaja Dobrovoljc (JSI), Polona Gantar (JSI), Tina Munda (JSI)

Date: 28/07/2022

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

Deliverable Number: D3.8

Dissemination Level: Public

Delivery Date: 28/07/2022

Version: 1.0

Author(s): Federico Martelli (Sapienza), Marco Maru (Sapienza), Cesare Campagnano (Sapienza), Roberto Navigli (Sapienza), Paola Velardi (Sapienza), Rafael-J Ureña-Ruiz, Francesca Frontini (CNR-ILC), Valeria Quochi (CNR-ILC), Jelena Kallas (EKI), Kristina Koppel (EKI), Margit Langemets (EKI), Jesse de Does (INT), Rob Tempelaars (INT), Carole Tiberius (INT), Rute Costa (NOVA CLUNL), Ana Salgado (NOVA CLUNL), Sanni Nimb (DSL), Sussi Olsen (UCPH), Simon Krek (JSI), Jaka Čibej (JSI), Kaja Dobrovoljc (JSI), Polona Gantar (JSI), Tina Munda



Project Acronym: ELEXIS
Project Full Title: European Lexicographic Infrastructure
Grant Agreement No.: 731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
28/07/2022		Federico Martelli (Sapienza), Marco Maru (Sapienza), Cesare Campagnano (Sapienza), Roberto Navigli (Sapienza), Paola Velardi (Sapienza), Rafael-J Ureña-Ruiz, Francesca Frontini (CNR-ILC), Valeria Quochi (CNR-ILC), Jelena Kallas (EKI), Kristina Koppel (EKI), Margit Langemets (EKI), Jesse de Does (INT), Rob Tempelaars (INT), Carole Tiberius (INT), Rute Costa (NOVA CLUNL), Ana Salgado (NOVA CLUNL), Sanni Nimb (DSL), Sussi Olsen (UCPH), Simon Krek (JSI), Jaka Čibej (JSI), Kaja Dobrovoljc (JSI), Polona Gantar (JSI), Tina Munda



Table of Contents

1	Introduction	9
2	Lexical-semantic analytics for NLP	10
2.1	Sense clustering	10
2.1.1	Clusty	10
2.1.2	Results and discussion	11
2.2	Domain labeling of text	12
2.2.1	Our approach	12
2.2.2	Results and discussion	14
2.3	Diachronic distribution of senses	14
2.3.1	Corpus creation and analysis	15
2.3.1.1	Bulgarian	15
2.3.1.2	Dutch	24
2.3.1.3	Slovenian	32
3	The ELEXIS parallel sense-annotated dataset	40
3.1	Composition of the corpus	41
3.2	Language-specific lexical-semantic analysis	42
3.2.1	Bulgarian	44
3.2.2	Danish	44
3.2.3	Dutch	46
3.2.4	English	49
3.2.5	Estonian	51
3.2.6	Hungarian	52
3.2.7	Italian	54
3.2.8	Portuguese	57
3.2.9	Slovenian	58
3.2.10	Spanish	61
4	Conclusion	63
5	References	64

List of Tables

Table 1 - Clusty - in vitro evaluation against our gold standard of 300 words.	11
Table 2 - Clusty - in vivo evaluation.	12
Table 3 - Results in the English domain labeling setting.	14
Table 4 - Results in the low-resource languages setting.	14
Table 5. Statistical information for the Diachronic Corpus of Bulgarian Texts.	16
Table 6. Number of words (word forms and lemmas), sentences in 6 periods of historical dataset.	17
Table 7: The top 20 most frequent senses in all time periods (Bulgarian).	19
Table 8: Top 10 most frequent senses present in all time periods (Slovenian).	33
Table 9: Top 10 most frequent senses in the sense-annotated IMP corpus.	35
Table 10: Top 20 most frequent lemmas with a shift in sense distribution.	36
Table 11 - ELEXIS-WSD datasets with sense annotations.	42
Table 12. Sense-annotated tokens by part-of-speech.	43

List of Figures

Figure 1 - Transformer-based architecture for domain-labeling.	13
Figure 2 - Distributions over time.	22
Figure 3: Concordances of the noun <i>duim</i> annotated with sense bn:00046209n.	25
Figure 4: Duim - sense distribution over time.	26
Figure 5: Dienst - sense distribution over time.	28
Figure 6: Dienst - examples.	29
Figure 7: Zwaar - sense distribution over time.	30
Figure 8: Bezetten - sense distribution over time.	31
Figure 9: Analysis of individual cases.	37



D3.8 Lexical-semantic analytics for NLP - final report

Figure 10: Semantic description of headword alcohol with associated ELEXIS-WSD-sl sentences in Lexonomy dictionary editor.

60



1 Introduction

The present document illustrates the work carried out in task 3.3 (work package 3) focused on lexical-semantic analytics for Natural Language Processing (NLP). This task aims at computing analytics for lexical-semantic information such as words, senses and domains in the available resources, investigating their role in NLP applications.

Specifically, this task concentrates on three research directions, namely i) *sense clustering*, in which grouping senses based on their semantic similarity improves the performance of NLP tasks such as Word Sense Disambiguation (WSD), ii) *domain labeling of text*, in which the lexicographic resources made available by the ELEXIS project for research purposes allow better performances to be achieved, and finally iii) analysing the *diachronic distribution of senses*, for which a software package is made available.

In this deliverable, we illustrate the research activities aimed at achieving the aforementioned goals and put forward suggestions for future works. Importantly, we stress the crucial role played by high-quality lexical-semantic resources when investigating such linguistic aspects and their impact on NLP applications. To this end, as an additional contribution, we address the paucity of manually-annotated data in the lexical-semantic research field and introduce the *ELEXIS parallel sense-annotated dataset*, a novel entirely manually-curated parallel corpus available in 10 European languages and featuring 5 annotation layers.



2 Lexical-semantic analytics for NLP

We now describe the tasks under consideration and provide a detailed report of the current status of the research activities for each task. D3.1 already reports the work on sense clustering, whereas D3.3 illustrates the work on domain labeling of text, which we briefly summarize hereafter. We then move on to the new work on the analysis of the diachronic distribution of senses and the multilingual corpus creation effort.

2.1 Sense clustering

Sense clustering can be defined as the computational task of grouping senses into sets of senses based on their semantic relatedness. The cardinality of each set can vary and depends on the degree of granularity desired. One of the goals of sense clustering is to reduce the fine granularity of sense inventories, thus leading to a significant increase in WSD performances (Navigli, 2009). Interestingly, when asked to determine the most appropriate sense within a fine-grained sense inventory, human annotators show an agreement ranging around 0.8. This suggests that, in some cases, granularity should be reduced and sense clustering could allow us to achieve this goal.

2.1.1 Clusty

We describe Clusty, a knowledge-based approach to sense clustering. This approach consists of the following two steps:

- 1) *Extraction of a lexical vector*, in which we extract content words from definitions, creating bags of words. We use the bag of words to extract the most relevant Wikipedia pages. Then, we compute a lexical vector of the target sense by exploiting the words contained in the target Wikipedia pages. Each vector component is

10



weighted according to its importance for the target senses. To do this, we use the lexical specificity (Lafon 1980) against the Wikipedia corpus.

Metric	Scores
Macro accuracy	0.714
Micro accuracy	0.730

Table 1. Clusty - in vitro evaluation against our gold standard of 300 words

- 2) *Sense clustering* in three steps: i) we calculate the similarity between all vector representations, ii) we sort using the cosine similarity score, and iii) we identify three thresholds to determine which senses should be included in the same cluster.

2.1.2 Results and discussion

We evaluate our approach both in vitro and in vivo. As for the first scenario, we create a gold standard by manually clustering WordNet (Miller 1995) senses of ambiguous words with polysemy degree ranging from 3 to 10 according to WordNet. The clustering is performed by expert linguists with proven experience in this type of task. We report the results in Table 1. As far as the in vivo evaluation is concerned, we evaluate the performance of a neural WSD system (Vial et. al 2019) when adopting different sense inventories, i.e. the WordNet sense inventory, Lexnames¹, WordNet domains (Magnini and Cavaglià, 2000) and ours. As can be seen in Table 2, the model under consideration achieves 88.9 F1, only one point below WordNet domains which is a manual clustering.

A challenging research direction to be addressed in future works could be to cluster complex lexical units such as idioms based on their semantic commonalities (Sag et al. 2002, Navigli and Martelli 2019, Tedeschi et al. 2022). Given the well-known coverage issues affecting lexical knowledge bases especially in languages other than English, one viable approach could

¹ <https://wordnet.princeton.edu/documentation/lexnames5wn>



consist in exploiting the inherent ability of WSD systems to discriminate between senses in monolingual and cross-lingual scenarios, dropping the requirement of fixed sense inventories (Pilehvar et al. 2019, Martelli et al. 2021).

Inventory	F1	PPL	geometric mean
fine-grained	80.5	2.45	14.05
lexnames	87.5	1.86	12.77
WordNet domains	89.9	1.89	13.05
our coarse-grained sense inventory	88.9	1.89	12.97

Table 2. Clusty - In vivo evaluation.

2.2 Domain labeling of text

For the purposes of the ELEXIS project, we can define domain labeling as the text classification task of tagging dictionary definitions with labels indicating domains of knowledge such as biology, computer science or medicine.

2.2.1 Our approach

In the light of recent NLP advances, novel neural architectures were proposed such as the Transformer (Vaswani et al. 2017) which proved to outperform almost all alternative architectures in several NLP tasks. Therefore, we exploit this architecture to create a neural classifier capable of assigning domain labels to dictionary definitions. Specifically, we create a sequence classification model based on m-BERT (Devlin et al. 2018). For each dictionary definition provided as input, our model produces a dense representation using the [CLS] token and finally performs label classification. Our architecture is illustrated in Figure 1. In order to train our model, we used data extracted from BabelNet



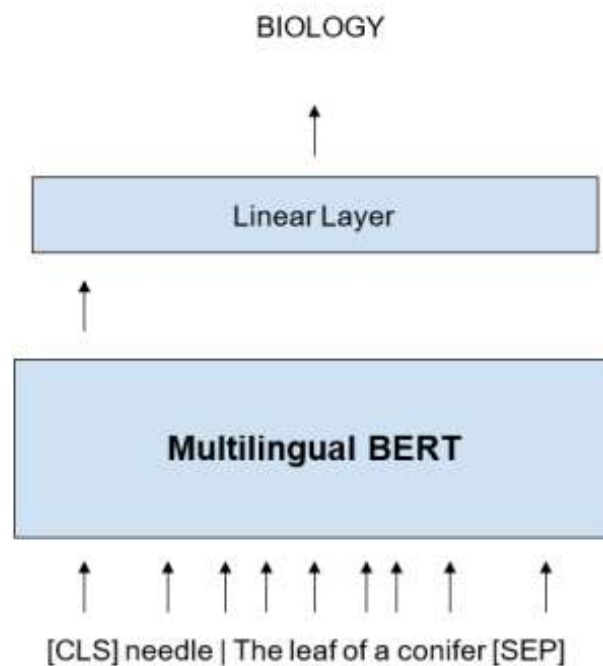


Figure 1 - Transformer-based architecture for domain labeling

(Navigli and Ponzetto 2012) and from the lexicographic resources made available within the ELEXIS project. Specifically, we extract all WordNet nominal synsets and their lemmas in BabelNet and their corresponding domain labels since each of them has been manually tagged with one of those 37 domain labels. As a result, for each considered definition we obtain a triple composed of: i) the domain label, ii) the definition and iii) the main lemma. As for the ELEXIS lexicographic resources, we extract triples from 5 dictionaries in the following languages: Bulgarian, Croatian, Spanish, Serbian and Slovenian. We perform experiments in two different settings, one in which we test on English and the other one on the aforementioned low-resource languages.

2.2.2 Results and discussion

In our experiments, a crucial goal is to demonstrate that higher performances can be achieved thanks to the lexicographic resources provided within the ELEXIS project.

model	#train	#dev	#test	macro-F1	weighted-F1
EN (base)	83,892	2,377	4,890	65.8	75.8
EN+ELEXIS data	102,004	2,377	4,890	67.3	76.2

Table 3 - Results in the English domain labeling setting

model	#train	#dev	#test	macro-F1	weighted-F1
LR (base)	109,677	8,454	17,752	55.0	70.9
LR+ELEXIS data	127,789	8,454	17,752	56.8	71.6

Table 4 - Results in the low-resource languages setting

As illustrated in Table 3 and 4, we show how the integration of the data provided by ELEXIS allows us to attain better performance with 67.3 macro-F1 compared to 65.8 in the English scenario, and 56.8 against 55.0 in the low-resource languages setting.

2.3 Diachronic distribution of senses

This task aims at investigating the distribution of senses over time. In fact, in NLP the most frequent sense has been used as a solid baseline. However, in the light of the evolution of language, the most frequent sense of a given lexeme might change over time. In this task, we explore such phenomenon in multiple languages and conduct a linguistic analysis. Our findings are of vital importance when dealing with texts dating back to past centuries. To achieve the aforementioned goals, we create 4 different corpora where each sentence S is tagged with the indication of time in which S was written. Subsequently, we automatically



disambiguate those corpora using a state-of-the-art multilingual WSD system, called AMuSE (Orlando et al. 2021). Finally, we create an algorithm which aggregates the data and allows us to analyse relevant changes in terms of most frequent sense over time. In the following sections, we detail the creation of each corpus and report our language-specific analysis.

2.3.1 Corpus creation and analysis

In this section we describe the creation of corpora for analysing the diachronic distribution of senses.

2.3.1.1 Bulgarian

The Diachronic Corpus of Bulgarian Texts is part of the Bulgarian National Corpus. Initially 54 texts were extracted from the Bulgarian National Corpus, amounting to 754 814 words of running text covering periods up to 1950 (for the periods 1951-1990 and 1990-2021 no available texts were found in the Bulgarian National Corpus with open access).

Additionally, to cover the periods of 1951-1990 and 1990-2021 texts were collected automatically from the several online sources and 11 new texts were added amounting to 341 926 words. All text units are supplied with extensive metadata following the metadata conventions of the Bulgarian National Corpus. The selected time periods are: 1851-1880; 1881-1910; 1911-1930; 1931-1950; 1951-1990; 1991-2021 (Table 5).

The domains included in the dataset are: fiction, news, science. The selection of domains was based on observations on the coverage of the domains across time periods. Administrative and other types of texts are rare in the earlier periods and are thus not included in the Diachronic Corpus (Table 5).

Period	Number of texts	Number of words	Number of authors	Coverage of domains



D3.8 Lexical-semantic analytics for NLP - final report

1850-1880	5	154 886	4	Fiction, News, Science
1881-1910	10	252 426	7	Fiction, News, Science
1911-1930	24	180 241	10	Fiction, News, Science
1931-1950	15	167 261	5	Fiction, News, Science
1951-1990	5	195 500	5	Fiction
1991-2021	6	146 426	6	Fiction
Total	65	1 096 740	37	Fiction, News, Science

Table 5. Statistical information for the Diachronic Corpus of Bulgarian Texts

The Diachronic Corpus is distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence and can be obtained at: <http://dcl.bas.bg/bulnc/en/dostap/izteglyane/> Statistical analysis shows relatively similar distribution of words, word-to-lemma ration, etc. across time periods. Similar is also the complexity of sentence structure in terms of number of classes in the sentence (average of 2.0 to 2.4 clauses per sentence). Significant divergence is observed in the length of the sentence - 21.2 in the earliest period (1850-1880) down to 14.7 and 14.3 in the latest ones (1951-1990 and 1991-2021) (Table 6). Further analysis is needed to confirm whether this is due to text selection.



D3.8 Lexical-semantic analytics for NLP - final report

	#words per sent	#clauses per sent	#unique wordforms per 1000 words	#unique lemmas per 1000 words
1850-1880	21.18	2.39	161.57	92.97
1881-1910	17.01	2.37	174.28	100.13
1911-1930	19.34	2.40	176.24	98.84
1931-1950	19.38	2.30	185.84	106.09
1951-1990	14.74	2.26	200.29	109.08
1991-2021	14.29	2.04	198.37	111.03

Table 6. Number of words (word forms and lemmas), sentences in 6 periods of historical dataset

Observations on the top 100 most frequent word senses from each time period shows: (a) A total of 222 word senses appear in the top 100 of at least one time period. (b) 44 senses appear in all time periods and further 13 appear in 5 out of the 6 periods - these cover frequent words of general meaning such as godina (year), myasto (place), pat (road), mislya (believe), balgarski (Bulgarian), rabota (work), imam (have), zhivot (life), etc.

We have separately analysed the filtered senses in several directions:

(1) We have counted all occurrences of each BabelNet sense (counting together all lemmas representing the sense) and analysed the most frequent senses across different time periods. In particular, we are interested in cases of discrepancy in frequency which possibly mark changes in the usage of these concepts in language.

The top 20 most frequent senses that appear in all time periods are present in the following Table:



D3.8 Lexical-semantic analytics for NLP - final report

BabelNet Sense	Definition	Bulgarian words	Total frequency
bn:00114234r	Without any others being included or involved	изключително_AD V; само_ADV;	1298
bn:00114626r	And nothing more	само_ADV;	1248
bn:00067975n	An open way (generally public) for travel or transportation	път_NOUN;	863
bn:00088733v	Cause to move; cause to be in a certain position or condition	има_VERB;	851
bn:00088818v	Organize or be responsible for	има_VERB;	845
bn:00116381r	In the historical present; at this point in the narration of a series of past events	сега_ADV;	800
bn:00069619n	An educational institution	училище_NOUN;	727
bn:00116382r	In these times	сега_ADV;	682
bn:00083294v	Undergo a change or development	стана_VERB;	647
bn:00018345n	A young person of either sex	младеж_NOUN; дете_NOUN;	634
bn:00080721n	Any artifact consisting of a road or path affording passage from one place to another	път_NOUN;	587
bn:00085337v	Come to pass	стана_VERB;	498
bn:00115530r	(often used as a combining form) in a good or proper or satisfactory manner or to a high standard ('good' is a nonstandard dialectal variant for 'well')	добре_ADV;	462
bn:00007630n	A very young child (birth to 1 year) who has not yet begun to walk or talk	дете_NOUN;	440
bn:00103673a	Morally admirable	добър_ADJ;	424
bn:00103675a	Of moral excellence	добър_ADJ;	402
bn:00117603r	Thoroughly or completely; fully; often used as a combining form, "well-satisfied customers"	добре_ADV;	391



D3.8 Lexical-semantic analytics for NLP - final report

bn:00028934n	The solid part of the earth's surface	земя_NOUN; суша_NOUN;	378
bn:00084388v	Utter a sudden loud cry	извикам_VERB;	369
bn:00062019n	The sound made by the vibration of vocal folds modified by the resonance of the vocal tract	глас_NOUN;	324

Table 7: The top 20 most frequent senses in all time periods (Bulgarian)

The most frequent senses belong to the commonly used lexicon, and this is evidence that words that nominate unchanging objects and abstractions also do not change.

Some examples of word senses with different distributions across time periods are shown below. Proper conclusions can be drawn only for words with relatively high frequency (in general) to avoid phenomena occurring by chance.

Example 1. Words with general meaning occurring in all periods but with lower frequency in certain periods and relatively equal distribution in the rest of the periods

bn:00036632n A group of people with a common ideology who try together to achieve certain general goals движение_NOUN (movement) Total freq. 203

1850-1880.tsv	43	21.2 %
1881-1910.tsv	35	17.2 %
1911-1930.tsv	57	28.1 %
1931-1950.tsv	59	29.1 %
1951-1990.tsv	7	3.4 %
1991-2021.tsv	2	1.0 %

Example 2. Words occurring in all periods but with lower frequency in certain periods and high frequency in a particular period



D3.8 Lexical-semantic analytics for NLP - final report

bn:00018819n One of the groups of Christians who have their own beliefs and forms of

worship църква_NOUN (church) Total freq. 186

1850-1880.tsv 108 58.1 %

1881-1910.tsv 24 12.9 %

1911-1930.tsv 18 9.7 %

1931-1950.tsv 30 16.1 %

1951-1990.tsv 3 1.6 %

1991-2021.tsv 3 1.6 %

Example 3. Words occurring in earlier periods but not in later ones

bn:00060072n A civil or military authority in Turkey or Egypt паша_NOUN Total freq.

92

1850-1880.tsv 19 20.7 %

1881-1910.tsv 24 26.1 %

1911-1930.tsv 22 23.9 %

1931-1950.tsv 27 29.3 %

1951-1990.tsv 0 3 %

1991-2021.tsv 0 3 %

Example 4. Words occurring only in recent periods

bn:00013723n The act of constructing something строителство_NOUN Total freq. 8

1951-1990.tsv 6 75.0 %

1991-2021.tsv 2 25.0 %

Example 5. Words occurring only in earliest and most recent periods and not in the middle periods, most likely with a newly developed meaning in recent times

bn:00047693n The act of issuing printed materials издаване_NOUN Total freq. 7

1850-1880.tsv 4 57.1 %



 D3.8 Lexical-semantic analytics for NLP - final report

1881-1910.tsv 2	28.6 %
1991-2021.tsv 1	14.3 %

(2) For each BabelNet synset (sense) we considered the occurrences of all lemmas of that sense. We are interested in noticeable differences in the usage of different lemmas which can show speakers' preferences for certain lexical units in general or in different time periods.

Example 6. Synonyms that are both equally used across all time periods

bn:00021644nA state at a particular time состояние_NOUN; положение_NOUN;

Example 7. Synsets for which a certain synonym dominates the usage across all time periods.

bn:00028934nThe solid part of the earth's surface земля_NOUN; суша_NOUN;

The word 'земля' is preferred (375 occurrences) over 'суша' (3 occurrences), all appearing in different time periods.

bn:00005846nA distinctive odor that is pleasant мирис_NOUN; миризма_NOUN

The word 'миризма' is preferred (12 occurrences) over 'мирис' (5 occurrences), all appearing in different time periods.

Example 8. Cases where a synonym is used only in earlier time periods.

bn:00003242nThe feeling that accompanies something extremely surprising

удивление_NOUN; изумление_NOUN;

The word 'изумление' is much less frequent and appears only in earlier texts before 1910 while 'удивление' is used in later periods as well (except in the most recent one).

Example 9. Cases where a synonym is used only in later time periods.



D3.8 Lexical-semantic analytics for NLP - final report

bn:00001304nThe act of delivering a formal spoken communication to an audience

реч_NOUN; изказване_NOUN; слово_NOUN;

As the diagram below shows, the word ‘изказване’ only appears in recent times and takes over the other words in the synset which are prevailing in earlier periods.

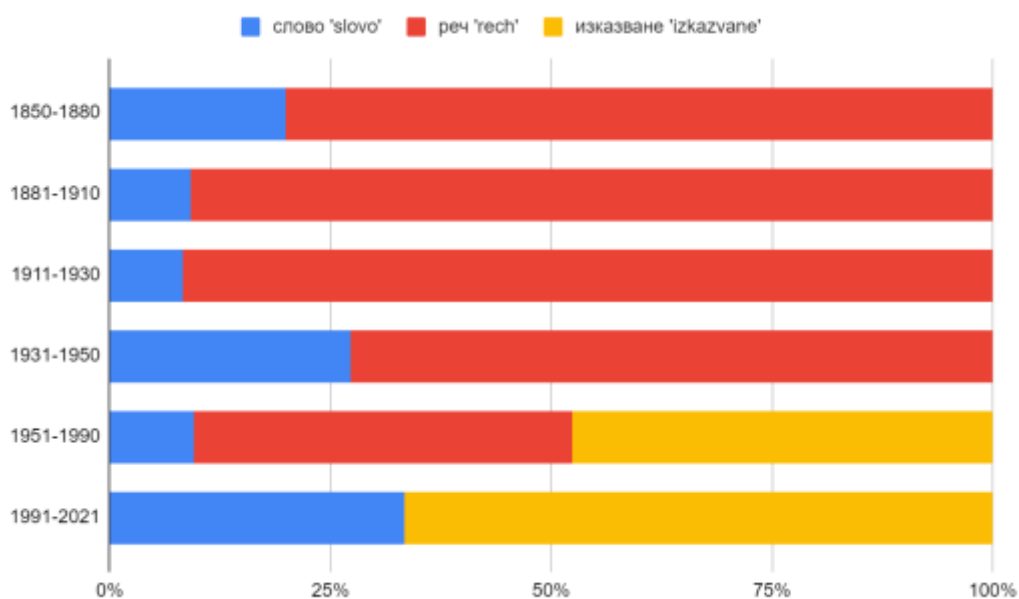


Figure 2 - Distributions over time

(3) For each word (main form) we can further analyse the different senses it appears with. This can be useful to analyse the development of polysemy and occurrence of semantic neologisms.

Example 10. Metaphorical use of a word in certain periods.

bn:00074459nA violent commotion or disturbance буря_NOUN (storm) (1850-1880 and 1881-1910) Total freq. 6

bn:00074459nA violent commotion or disturbance буря_NOUN (storm) (all time periods) Total freq. 58



 D3.8 Lexical-semantic analytics for NLP - final report

The example shows the metaphorical use of the word ‘буря’ (storm) in texts of the earlier periods related to describing riots and national suffering before the liberation of Bulgaria from the Ottomans.

Example 11. Use of homonyms.

(a) bn:00062658n A long-handled hand tool with sharp widely spaced prongs for lifting and pitching hay вила_NOUN (pitchfork)

bn:00080000n Country house in ancient Rome consisting of residential quarters and farm buildings around a courtyard вила_NOUN (villa)

The word ‘вила’ (pitchfork) is found only in texts from the period 1911-1930, while ‘вила’ (villa) occurs only from 1931 onwards.

(b) bn:00036632n A group of people with a common ideology who try together to achieve certain general goals движение_NOUN (movement)

bn:00056030n A natural event that involves a change in the position or location of something движение_NOUN (movement)

Both appear in all time periods.

(c) bn:00023310n An area wholly or partly surrounded by walls or buildings двор_NOUN (yard)

bn:00024541n The enclosed land around a house or other building двор_NOUN (yard)

bn:00023306n The family and retinue of a sovereign or prince двор_NOUN (court)

The first two appear in all time periods while the last one is used only in texts up to 1950.

Example 12. Cases where a word develops new senses.

bn:00034949n An object firmly fixed in place (especially in a household) инсталация_NOUN (installation)

bn:00046934n The act of installing something (as equipment) инсталация_NOUN (installation)



The first sense of the word appears after 1930 and the second one only after 1990.

2.3.1.2 Dutch

Description of the corpus. The corpus used for Dutch has been specially compiled for this task from quotations from the *Woordenboek der Nederlandsche Taal* (WNT, Dictionary of the Dutch language), one of the largest historical dictionaries in the world. The WNT describes the meaning and history of hundreds of thousands of words in written Dutch from 1500 to 1976. It contains about 95,000 main headwords and has about 1,700,000 quotations.

We sampled the quotations in such a way that a reasonable distribution over the 16th-20th centuries was obtained. This resulted in a corpus of 247755 quotations and 5349552 tokens (the complete quotation corpus has about 16 million tokens).

The corpus was sent to Sapienza, where word sense disambiguation according to the BabelNet sense inventory was performed. In order to analyze the results, we looked for cases where the Most Frequent Sense (MFS) undergoes a significant change, and this is supported by a minimum of 20 quotations per sense. To see the actual concordances of the words, the annotated corpus was loaded into the Corpus Query System BlackLab². An example of the concordances of the noun *duim* ‘thumb’ annotated with BabelNet sense *bn:00046209n* : *A unit of length equal to one twelfth of a foot* is given below.

² <https://inl.github.io/BlackLab/>



D3.8 Lexical-semantic analytics for NLP - final report

...wierd , by beurten wierd aangewonnen , en dagelyks eenige	duimen	toenam , maar niet schielyk en op eene reis ,...
...maeken , — Oordeele een weer-glas , wiens Pypje vyftien	duimen	lang is , en eene holte , als de nevenstaende...
...en daer aan een zeer dun uitgeblazen bolletje van een	duim	en twee linien Rhylandse maat diameters is , van bequaem...
...steekt ... een takje van kragtig een-jaerig hout , vyf	duimen	lang , omtrent drie vingeren diep in de aerde ...
...besloten blyven , Dat de vorst zelden dieper als agtien	duimen	in de aerde dringt ; zelfs hebbe het ys in...
...eerlang aan vinken van een half voet lang en twee	duim	breet gekapt en gesneeden zynde , stryken enz. ; Vermits...
...walvischoog) , was in zyn Diameter zeer na 2	duim	, Dies wil hy (de walvisch) liever in...
...den rug , en onder aan den buik omtrent zes	duim	dik , Die ' t spel niet en kan ,...
...diameteraal wyd is , en in de lengte 3 1/4	duim	, van dezelfde ruimte als de uitbreiding C. G (...
...eindelyk de kwikzilver gezogen tot op de hoogte van zevenentwintig	duim	, De kwik wederom uit gelopen zynde , heeft de...
...dekken , Deze kopjens hebben een voetje , van een	duim	hoog , In ' t openen der Kassen , of...
...eerlang aan vinken van een half voet lang en twee	duim	breet gekapt en gesneeden zynde , stryken het met hunne...
...waar na ' t genoemd werd) vyf of zes	duim	lang , en omtrent twee vingeren breed ... Dit...
...huis-uurwerk is , De bladeren vallen omtrent vier of vyf	duim	lang , ... zeer schoon groen , en glad van...
...vyf voeten lang ; dog agter aan vier a vyf	duimen	breed , en als een zaag vol tanden is ,...
... , aan de andre kant 26 in getal , 2	duim	lang zynde , verto onen , Dat hy ... door...
...ou de quelque étoffe avec les ciseaux , Van vier	duimen	in ' t vierkant . Di quattro dita in quadro...
...De vuist die nu den scepter zwaait En op wiens	duim	de werelt draeit is ... Wel haest bekneelt in yzre...
...9 , 10 à 11 voeten lang , en 8	duim	kant , — Tegen dit groot gebrek (het niet-vloeien...
...Te zamen spannen , moet geen schoone vrou verschynen ,	Duim	is een zoon geboren . Laet hierop ' t zorgen...

 Figure 3: Concordances of the noun *duim* annotated with sense bn:00046209n

This allowed us to inspect the individual occurrences of the words annotated with a particular sense. For the manual evaluation, the results for a small sample of nouns, verbs and adjectives were studied in detail, focussing on two aspects: 1) are the words correctly disambiguated, i.e. were they annotated with the correct sense, and 2) how are the senses distributed over time. We will discuss four of those here, i.e. *duim* (noun), *dienst* (noun), *zwaar* (adjective) and *bezetten* (verb).

Analysis of sample. For each of the sample words, we present the data in tabular form.

A short explanation:

The *Salient Most Frequent Sense changes* are pairs of periods (centuries) (c_1, c_2) for which:

- The most frequent sense in c_1 and c_2 is different
- The relative frequency change of both most frequent senses is at least 0.2
- The change is supported by at least 20 occurrences in the corpus

The semantic profile per period is presented in a table, where

- the entries are of the form *absolute frequency*_(rank)



D3.8 Lexical-semantic analytics for NLP - final report

- the entries corresponding to a most frequent sense are in bold
- the entries relevant to a salient MFS change are underlined

We also provide a bar chart visualising the evolution of the semantic profile. We start with the results for the noun *duim*. In the corpus, two senses occur, i.e. *A unit of length equal to one twelfth of a foot* (in brown) and *The thick short innermost digit of the forelimb* (in red).

o Duim, NOUN

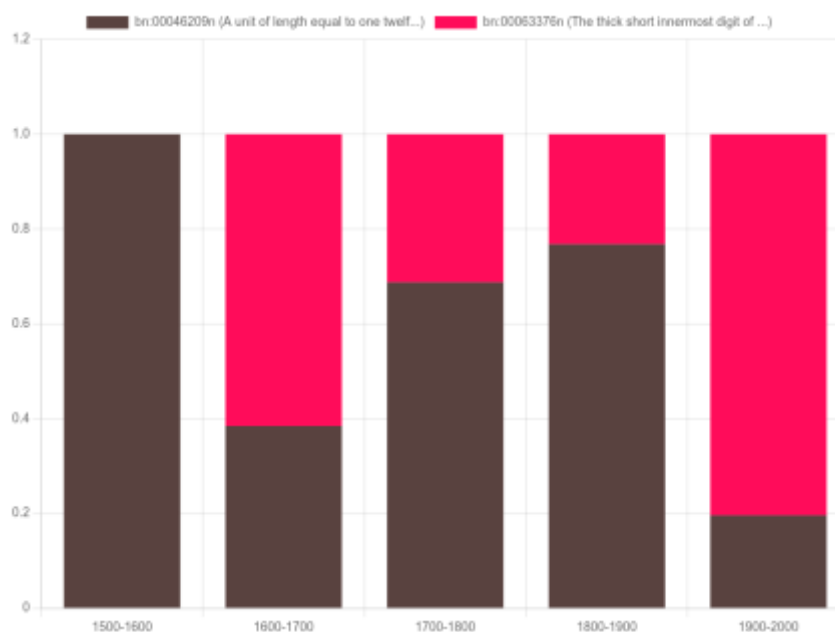


Figure 4: Duim - sense distribution over time

Definitions:

bn:00046209n : A unit of length equal to one twelfth of a foot

bn:00063376n : The thick short innermost digit of the forelimb

Salient MFS changes:

c_1 =Century 1	c_2 =Century 2	mfs_1 =MFS in c_1	$relfreq(c_1, mfs_1)$	$relfreq(c_2, mfs_1)$	mfs_2 =MFS in c_2	$relfreq(c_1, mfs_2)$	$relfreq(c_2, mfs_2)$
<u>1700-1800</u>	<u>1900-2000</u>	bn:00046209n	0.69	0.20	bn:00063376n	0.31	0.80
<u>1800-1900</u>	<u>1900-2000</u>	bn:00046209n	0.77	0.20	bn:00063376n	0.23	0.80

Frequency data:

period	bn:00046209n	bn:00063376n
1500-1600	1 ⁽¹⁾	0
1600-1700	10 ⁽²⁾	16 ⁽¹⁾
1700-1800	33 ⁽¹⁾	<u>15</u> ⁽²⁾
1800-1900	86 ⁽¹⁾	<u>26</u> ⁽²⁾



 D3.8 Lexical-semantic analytics for NLP - final report

 1900-2000 10 (2) 41 (1)

Judging from the above results, the sense *A unit of length equal to one twelfth of a foot* is more frequent in the material from 1700-1800 and 1800-1900, whereas in the material from 1900-2000 the other sense (*The thick short innermost digit of the forelimb*) occurs most frequently.

Looking at the concordances for *duim* in the different periods, we observe that the two senses were mostly correctly disambiguated. For instance, if we look at the concordances for *duim* in the period 1700-1800 in the sense *unit of length*, we see words such as *lang* 'long', *dik* 'thick', *breed* 'width' in the context and a quantifier or numeral is often preceding *duim*, whereas the words in the context of the concordances that are annotated with the sense *The thick short innermost digit of the forelimb* are *vinegar* 'finger', *hand* 'hand', and *voet* 'foot'.

This was a promising start. Unfortunately, the results were not as good for the noun *dienst*. For this noun, the following two senses are involved in a sense change, i.e. *A service conducted in a house of worship* (pink) and *Work that you are obliged to perform for moral or legal reasons* (purple)

dienst, NOUN



D3.8 Lexical-semantic analytics for NLP - final report

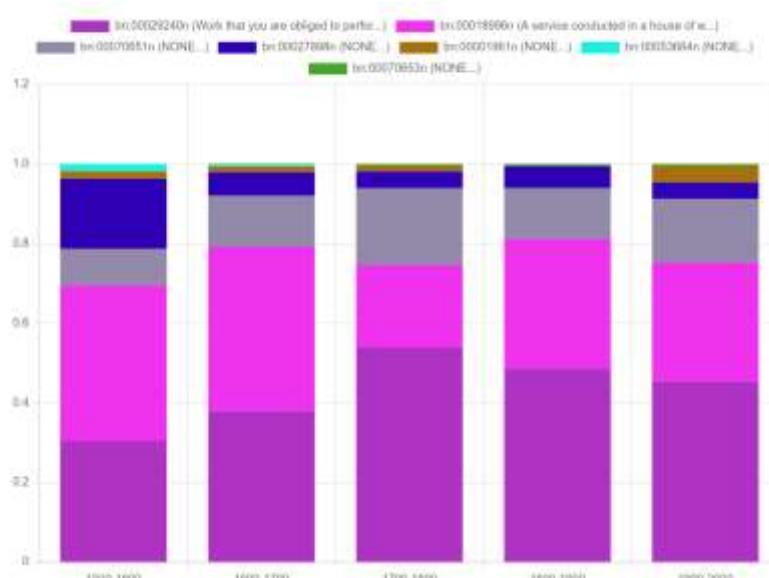


Figure 5: Dienst - sense distribution over time

Definitions:

bn:00029240n : Work that you are obliged to perform for moral or legal reasons
 bn:00018996n : A service conducted in a house of worship
 bn:00070651n : _ (only the senses involved in a MFS change where defined in the Sapienza set)
 bn:00027898n : _
 bn:00001961n : _
 bn:00053664n : _
 bn:00070653n : _

Salient MFS changes:

c_1 =Century 1	c_2 =Century 2	mfs_1 =MFS in c_1	$relfreq(c_1, mfs_1)$	$relfreq(c_2, mfs_1)$	mfs_2 =MFS in c_2	$relfreq(c_1, mfs_2)$	$relfreq(c_2, mfs_2)$
1600-1700	1700-1800	bn:00018996n	0.41	0.21	bn:00029240n	0.38	0.54

Frequency data:

period	bn:00029240n	bn:00018996n	bn:00070651n	bn:00027898n	bn:00001961n	bn:00053664n	bn:00070653n
1500-1600	33 (2)	42 (1)	10 (4)	19 (3)	2 (6)	2 (5)	0
1600-1700	182 (2)	198 (1)	62 (3)	28 (4)	7 (5)	2 (6)	1 (7)
1700-1800	115 (1)	44 (2)	41 (3)	9 (4)	3 (5)	0	1 (6)
1800-1900	88 (1)	59 (2)	23 (3)	10 (4)	0	0	1 (5)
1900-2000	88 (1)	58 (2)	31 (3)	8 (5)	8 (4)	0	1 (6)

Based on these results, the most frequent sense in the period 1600-1700 seems to be bn:00018996n: *A service conducted in a house of worship*, whereas in the period 1700-1800



D3.8 Lexical-semantic analytics for NLP - final report

the sense bn:00018996n: *Work that you are obliged to perform for moral or legal reasons* seems to be more frequent in the material.

However, when we have a close look at the concordances for 1600-1700, we see that there are actually very few instances where the sense *A service conducted in a house of worship* clearly applies. The first instance that was correctly annotated was found after 26 lines of concordance output.

... syne vele ende menichfuldige gedane **diensten** inde kercke alhyer Int predicken ...

Also the 44 instances in the 1700-1800 period which are annotated with this sense do not seem to be unproblematic.

Before	Hit	After
... condoleren , maar ook syn	dienst	heeft doen aanbieden tot het ...
... trouw , en zo veel	dienst	vergelden ? Wil hoopen en ...
... meer quaats , in zynen	dienst	, dan datze hunne eige ...
... trouw , en zo veel	dienst	vergelden ? C . Met ...
... de bazuinen , — Geen	dienst	van tweemaal zeven jaar ; ...
... en die met haar effective	dienst	doen en tot het schip ...

Figure 6: Dienst - examples

Due to the errors in the annotation, it is difficult to draw further conclusions.

For the adjective *zwaar*, the sense annotation seemed more reliable again, especially for those instances which were annotated with the sense *Of comparatively great physical weight or density* (green).

zwaar, ADJ



D3.8 Lexical-semantic analytics for NLP - final report

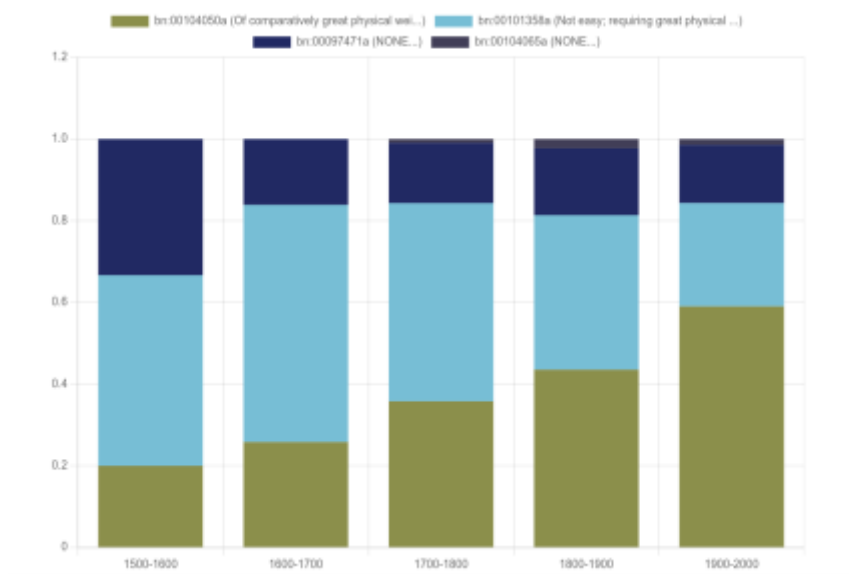


Figure 7: Zwaar - sense distribution over time

Definitions:

bn:00104050a : Of comparatively great physical weight or density

bn:00101358a : Not easy; requiring great physical or mental effort to accomplish or comprehend or endure

bn:00097471a : _

bn:00104065a : _

Salient MFS changes:

c_1 =Century 1	c_2 =Century 2	mfs_1 =MFS in c_1	$relfreq(c_1, mfs_1)$	$relfreq(c_2, mfs_1)$	mfs_2 =MFS in c_2	$relfreq(c_1, mfs_2)$	$relfreq(c_2, mfs_2)$
<u>1700-1800</u>	<u>1900-2000</u>	bn:00101358a	0.49	0.25	bn:00104050a	0.36	0.59
<u>1600-1700</u>	<u>1900-2000</u>	bn:00101358a	0.58	0.25	bn:00104050a	0.26	0.59
<u>1600-1700</u>	<u>1800-1900</u>	bn:00101358a	0.58	0.38	bn:00104050a	0.26	0.44

Frequency data:

period	bn:00104050a	bn:00101358a	bn:00097471a	bn:00104065a
1500-1600	3 ⁽³⁾	7 ⁽¹⁾	5 ⁽²⁾	0
1600-1700	<u>24</u> ⁽²⁾	54 ⁽¹⁾	15 ⁽³⁾	0
1700-1800	<u>73</u> ⁽²⁾	99 ⁽¹⁾	30 ⁽³⁾	2 ⁽⁴⁾
1800-1900	175 ⁽¹⁾	<u>152</u> ⁽²⁾	66 ⁽³⁾	9 ⁽⁴⁾
1900-2000	245 ⁽¹⁾	<u>105</u> ⁽²⁾	59 ⁽³⁾	6 ⁽⁴⁾

According to the results, the adjective *zwaar* is mostly annotated with the sense *Not easy; requiring great physical or mental effort to accomplish or comprehend or endure* in the material from 1600-1700 and 1700-1800, although the other sense is gradually becoming more frequent. In 1800-1900 the most frequent sense becomes *Of comparatively great*



D3.8 Lexical-semantic analytics for NLP - final report

physical weight or density, although the other sense is still quite frequent too. In the material from 1900-2000, the sense *Of comparatively great physical weight or density* is clearly the most frequent and the frequency of the other sense drops significantly.

o bezetten, VERB

With the verb *bezetten*, we encounter another limitation, which has to do with the rather limited and imperfectly balanced corpus we are using. The sense *Decorate or cover lavishly (as with gems)* (purple) seems overrepresented in the material in the period from 1700-1800. Furthermore, the disambiguation of the two senses does not as good as expected although the two senses are clearly different.

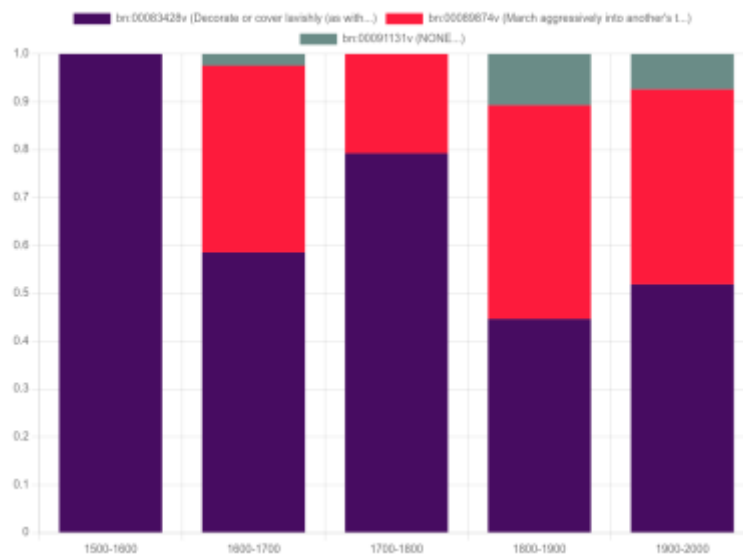


Figure 8: Bezetten - sense distribution over time

Definitions:

bn:00083428v : Decorate or cover lavishly (as with gems)

bn:00089874v : March aggressively into another's territory by military force for the purposes of conquest and occupation

bn:00091131v : _

Salient MFS changes:

c_1 =Century 1	c_2 =Century 2	mfs_1 =MFS in c_1	$relfreq(c_1, mfs_1)$	$relfreq(c_2, mfs_1)$	mfs_2 =MFS in c_2	$relfreq(c_1, mfs_2)$	$relfreq(c_2, mfs_2)$
1700-1800	1800-1900	bn:00083428v	0.79	0.45	bn:00089874v	0.21	0.45

Frequency data:

period	bn:00083428v	bn:00089874v	bn:00091131v
1500-1600	1 (1)	0	0



D3.8 Lexical-semantic analytics for NLP - final report

1600-1700	24 ⁽¹⁾	16 ⁽²⁾	1 ⁽³⁾
1700-1800	42 ⁽¹⁾	<u>11</u> ⁽²⁾	0
1800-1900	<u>25</u> ⁽²⁾	25 ⁽¹⁾	6 ⁽³⁾
1900-2000	14 ⁽¹⁾	11 ⁽²⁾	2 ⁽³⁾

Discussion

From the above analysis we can draw a few conclusions:

- Processing historical Dutch is still difficult, both for the WSD system and the more basic tasks; PoS or lemmatization errors also lead to wrong conclusions on semantic change, e.g. the lemmatization of “zyt” to “zitten” (to sit), which should have been “zijn” (to be).
- Especially in cases where there is a coarse sense distinction (e.g. *duim*), valid conclusions may be obtained from the material, even if the WSD is probably not extremely reliable in these cases.

For future research it would be interesting to:

- Compare the BabelNet and WNT sense inventories
 For a complete and thorough evaluation of sense distribution over time, the BabelNet sense-ids would ideally have been added to the TEI of the original input corpus, which contained WNT-based sense distinctions for those words corresponding to the headword of the article from which the quotations were taken.
- Improve the performance of the basic NLP tools (tagging, lemmatization).

2.3.1.3 Slovenian

The IMP digital library contains historical Slovene books and other publications, together 658 texts with over 45,000 pages from the period 1584-1919. These texts were annotated to be used as a language corpus with 17,723,566 tokens. In the corpus each word is marked-up with its modernised form, lemma, and morphosyntactic description (fine grained PoS tag). The annotations are automatic, so they contain a fair amount of errors.



D3.8 Lexical-semantic analytics for NLP - final report

The corpus is available in the CLARIN.SI repository in source TEI P5 XML and in the simpler and smaller vertical format, used by various concordancers, e.g. CWB and Sketch Engine:

- CLARIN.SI repository: <http://hdl.handle.net/11356/1031>
- CLARIN.SI NoSketch Engine: IMP ([Concordance](#))
- CLARIN.SI Kontext: IMP ([Concordance](#))

For the purposes of word sense annotation the IMP corpus data was split into four periods with the following data:

Period	Number of texts	Number of words
17th century	1	7,433
18th century	1	206,500
19th century	18	3,881,603
20th century	395	9,833,815
TOTAL	415	13,929,351

Table 8: Top 10 most frequent senses present in all time periods (Slovenian)

The distribution of senses in the corpus was first analyzed automatically by aggregating the number of different senses and counting their frequencies. Table 8 shows the top 10 most frequent senses present in all time periods - as expected, these refer to very general concepts like human beings (*človek*) or common actions or states (*biti* - to be, *imeti* - to have, *reči* - to say) as well as expressions of time (*potem* - after) and degree (*tako* - so (much)). These senses are shown to be relatively stable as they remain present and frequent throughout the four centuries included in the corpus.



D3.8 Lexical-semantic analytics for NLP - final report

BabelNet Synset Id	Definition	Example lemma	Frequency
bn:00046516n	A human being	človek ~ NOUN	102,212
bn:00083185v	Happen, occur, take place	biti ~ VERB	52,561
bn:00083181v	Have the quality of being; (copula, used with an adjective or a predicate noun)	imeti ~ VERB	39,700
bn:00093287v	Express in words	reči ~ VERB, povedati ~ VERB, praviti ~ VERB	28,073
bn:00117043r	To a very great extent or degree	tako ~ ADV	25,656
bn:00095597v	Use one's feet to advance; advance by steps	iti ~ VERB, hoditi ~ VERB, oditi ~ VERB	23,863
bn:00089240v	Have or possess, either in a concrete or an abstract sense	imeti ~ VERB	23,626
bn:00114626r	And nothing more	ali ~ ADV	23,029
bn:00114165r	Happening at a time subsequent to a reference time	potem ~ ADV	20,735
bn:00082788v	Reach a destination; arrive by movement or progress	prići ~ VERB	20,504

Table 9: Top 10 most frequent senses in the sense-annotated IMP corpus.

Table 9 shows the top 10 most frequent lemmas that have shown a shift in their distribution of senses in different centuries. The list contains both very general words (e.g. imeti - to have, prići - to come, morati - must, misliti - to think) where the changes in sense distribution are



D3.8 Lexical-semantic analytics for NLP - final report

less obvious (or likely coincidental due to automatic annotation), but also a number of homonyms, e.g. *svet* (world/council), where the sense "A body serving in an administrative capacity" becomes more frequent through time (compared to "All of the living human inhabitants of the earth"). An interesting example is also *kraj* (place), a word with an archaic sense meaning "end" - although the archaic sense is not present in BabelNet, the word has nonetheless been shown to have undergone a semantic shift.

Lemma	Frequency
imeti_VERB	56,902
pritti_VERB	34,407
morati_VERB	26,926
gospod_NOUN	20,004
misliti_VERB	18,248
star_ADJ	17,320
gledati_VERB	15,139
pot_NOUN	13,895
povedati_VERB	13,030
stati_VERB	12,961
res_ADV	11,710
svet_NOUN	11,555
storiti_VERB	9,402
kraj_NOUN	9,331
dolg_ADJ	9,257
najti_VERB	8,327
vesel_ADJ	8,325

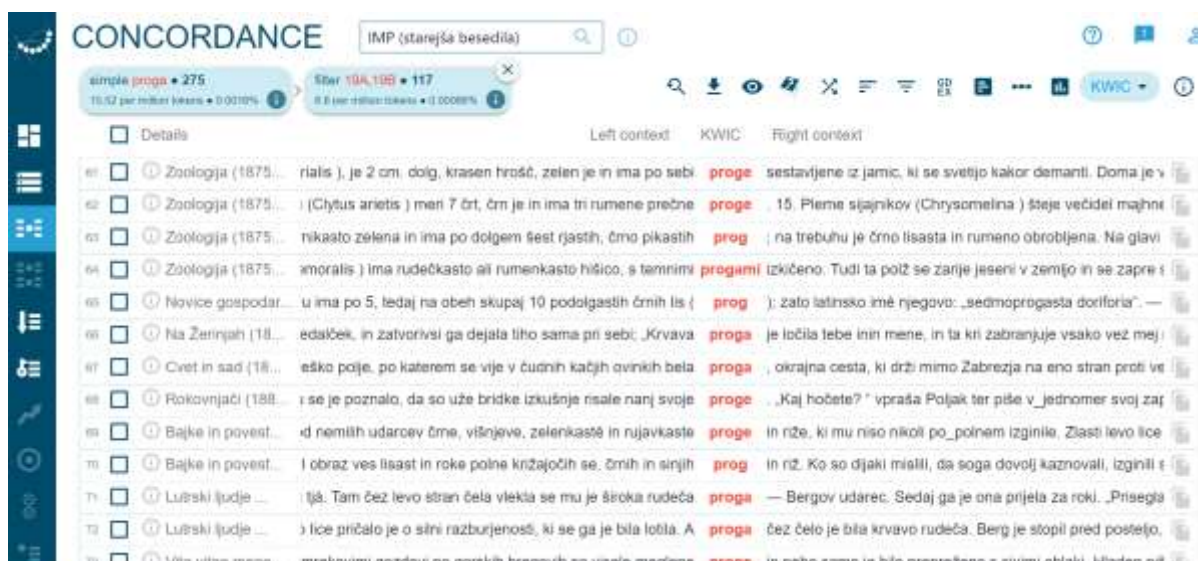


D3.8 Lexical-semantic analytics for NLP - final report

pravi_ADJ	7,662
zemlja_NOUN	7,522
hitro_ADV	6,516

Table 10: Top 20 most frequent lemmas with a shift in sense distribution.

In order to analyze the results manually and in more detail, we used the extracted data where the Most Frequent Sense (MFS) undergoes a significant change, and this is supported by a minimum of 20 quotations per sense (see below). To analyse the actual concordances of the words, the NoSketchEngine installation of IMP corpus was used, which allows text period filtering.



The screenshot shows the 'CONCORDANCE' tool interface. At the top, there is a search bar containing 'IMP (starejša besedila)'. Below the search bar, there are statistics for the search results: 'example proga • 275' and 'Star 104,189 • 117'. The main area displays a list of concordance entries, each with a checkbox, a document title, a snippet of text, and the word 'proga' highlighted in red. The entries are organized into columns: 'Left context', 'KWIC', and 'Right context'. The list includes various documents such as 'Zoologija (1875...', 'Novice gospodar...', 'Na Zerinjah (18...', 'Cvet in sad (18...', 'Rokovnjači (188...', 'Bajke in povest...', 'Luški ljudje ...', and 'Vila vitas mese...'. The word 'proga' is consistently highlighted in red across all entries.

Figure 9: Analysis of individual cases

Analysis of individual cases

We now analyse individual cases of the diachronic distribution of senses over time in Slovenian.



PALEC

sl_1800.tsv bn:00046209n 126 bn:00063376n 101
 sl_1900.tsv bn:00063376n 42 bn:00046209n 7

BabelNet synsets:

- bn:00046209n: A unit of length equal to one twelfth of a foot
- bn:00063376n: The thick short innermost digit of the forelimb

In texts from the 19th century the forelimb digit sense (bn:00063376n) is prevalent, but in the 20th century texts the unit sense (bn:00046209n) indicating the length of cca 2.5 cm became the more frequent one.

PROGA

sl_1800.tsv bn:00008471n 51 bn:00066021n 30bn:00049900n 17 bn:00051297n
 16
 sl_1900.tsv bn:00066021n 59 bn:00008471n 45bn:00051297n 40 bn:00049900n
 14

BabelNet synsets:

- bn:00008471n: A narrow marking of a different color or texture from the background
- bn:00066021n: A line of track providing a runway for wheels

In the 19th century texts a narrow marking of different color sense (bn:00008471n) is prevalent and in the 20th century the railway track (bn:00066021n) becomes the most frequent sense, which is expected since railways started to be built from mid-century in the area of present-day Slovenia.

PEKEL

sl_1700.tsv bn:00043604n 5 bn:00043603n 5
 sl_1800.tsv bn:00043603n 174 bn:00043604n 84 bn:00034629n 8
 sl_1900.tsv bn:00043603n 123 bn:00043604n 63 bn:00034629n 7

BabelNet synsets:

- bn:00043604n: (Christianity) the abode of Satan and the forces of evil; where sinners suffer eternal punishment
- bn:00043603n: Any place of pain and turmoil;



D3.8 Lexical-semantic analytics for NLP - final report

In the 18th century texts the Christian sense of hell (bn:00043604n) is prevalent, but in the 19th and 20th centuries the metaphorical sense of any place of pain (bn:00043603n) becomes the most frequent one.

PAŠA

sl_1800.tsv	bn:00041539n 433	bn:00060072n 292
sl_1900.tsv	bn:00060072n 542	bn:00041539n 103

BabelNet synsets:

- bn:00041539n: The act of grazing
- bn:00060072n: A civil or military authority in Turkey or Egypt

The two concepts are homonymous - one with the agricultural meaning of “grazing” (sheep) and the other denoting a high ranking official in Turkey or Egypt. In the 19th century texts the agricultural sense (bn:00041539n) is prevalent, but in the 20th century the “Turkish” concept (bn:00060072n:) becomes the most frequent one.

SLINAST

sl_1800.tsv	bn:00110753a 12	bn:00101247a 1
sl_1900.tsv	bn:00101247a 3	bn:00110753a 2

BabelNet synsets:

- bn:00110753a: Covered with or resembling slime
- bn:00101247a: Morally reprehensible

Although the numbers in this case are low, there is a pronounced prevalence of the literal sense “covered with slime” (bn:00110753a) in the 19th century texts, while the metaphorical sense (bn:00101247a) of moral reprehensibility is slightly more frequent in the 20th century texts.

OBVEZATI

sl_1800.tsv	bn:00083072v 36	bn:00082219v 31	bn:00083496v 9
sl_1900.tsv	bn:00082219v 38	bn:00083072v 26	bn:00083496v 5

BabelNet synsets:

- bn:00083072v: Dress by covering or binding
- bn:00082219v: Provide a service or favor for someone

An example of a verb: the concept of binding (with a bandage - bn:00083072v) is more frequent in the 19th century, and the contexts where the concept of “providing a service” (bn:00082219v) is used are more frequent in the 20th century texts.



3 The ELEXIS parallel sense-annotated dataset

A major limitation affecting many research fields including lexicography and NLP is the lack of high-quality manually-curated data which is labour-intensive and costly to produce. Fortunately, recent advances in NLP have shown their effectiveness for the creation and analysis of lexical-semantic resources both within and across languages. However, we believe that such approaches should be a starting point and that robust lexical-semantic studies should rely on manually-curated data whenever possible, which would also encourage deeper connections between the two research fields.

In order to address the aforementioned issues, we introduce the ELEXIS parallel sense-annotated corpus, a novel entirely manually-curated lexical-semantic resource available in 10 European languages, namely Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene and Spanish. The ELEXIS parallel corpus features 5 annotation layers, i.e. tokenization, sub-tokenization, lemmatization, part-of-speech tagging and word sense disambiguation. As for the semantic annotation layer, we used two different senses inventories: i) BabelNet and ii) an open high-quality dictionary or sense-inventory provided by each institution participating in the annotation process. Although initially we explored ways to leverage cross-lingual label propagation to pre-annotate our corpus as outlined in Procopio et al. 2021, given the specificity of our guidelines, e.g. we skipped named entities and addressed the insufficient coverage of the selected sense inventories by manually adding senses and definitions or modifying existing ones, in this case we decided to annotate manually.

Importantly, it will be possible to use the ELEXIS sense-annotated dataset not only to perform lexical-semantic analysis across languages, but also to carry out performance evaluation in



both supervised (Barba et al. 2021) and knowledge-based WSD approaches (Maru et al. 2019). Further details regarding the construction of the corpus can be found in Martelli et al. 2021. In this deliverable, we will focus on the semantic annotation and on a language-specific lexical-semantic analysis in which we will consider both morpho-syntactic and semantic issues encountered during the annotation process which will have to be taken into great consideration when creating lexical-semantic corpora.

3.1 Composition of the corpus

As shown in Table 11, the sizes of the datasets are comparable across languages: all datasets contain 2,024 sentences, with a size ranging from approx. 29,000 tokens to approx. 39,000 tokens, a difference mainly attributed to language-specific characteristics (e.g. use of cases vs. use of prepositions, presence of definite/indefinite articles, use of compound words). The tokens in the datasets were manually annotated as either content-words (and assigned a sense); non-content words (with no assigned sense); or content words with a sense that requires no disambiguation (e.g. named entities). Each dataset contains between 14,000 and 18,000 content words that have been assigned a sense from the selected sense inventories; the senses were mostly assigned from the original sense inventories selected for each language (between 11,000 and 15,000 words), with a smaller percentage annotated with senses from BabelNet (up to 2,600 for Italian and as little as 2 for Dutch). New senses (i.e. those not yet present in the original sense inventories or BabelNet) were also manually added during annotation: up to approx. 5,900 for Danish and as little as 119 for English). The extent of adding new senses was left to the discretion of individual language teams depending on the structure of their selected sense inventories and their goals. The number of unannotated tokens also varies between languages. The unannotated tokens are mostly comprised of named entities; while some language teams added named entities to their sense inventories as separate senses, others decided to process named entities at a later stage.



D3.8 Lexical-semantic analytics for NLP - final report

Lang.	Tokens	Subtokens	All sense-annotated tokens	Sense annotations from original sense inventories	BabelNet annotations	Annotations with newly added senses	Non-content tokens	Unannotated tokens	Annotated multiword expressions
BG	33,978	0	17,621	15,914	4	1,703	15,137	1,186	2,013
DA	32,528	484	18,071	11,385	977	5,709	14,175	164	797
EN	34,228	269	14,415	13,339	957	119	17,452	0	301
ES	37,680	142	16,423	14,441	267	1,715	19,030	0	133
ET	26,304	84	16,072	13,732	1,296	1,044	8,789	1,436	860
HU	29,657	211	13,974	11,439	96	2,439	13,119	2,436	4
IT	39,039	2,576	15,921	11,575	2,637	1,709	20,615	2,594	281
NL	34,923	620	13,656	11,284	2	2,370	18,820	11	93
PT	38,729	2,407	17,471	15,556	852	1,063	20,021	978	849
SL	31,233	0	14,826	14,826	0	0	14,148	2,259	547

Table 11 - ELEXIS-WSD datasets with sense annotations.

Table 11 shows the sense-annotated tokens in the datasets by part-of-speech: as expected, nouns are the most frequent, followed by verbs and adjectives. Adverbs are the least frequent.



D3.8 Lexical-semantic analytics for NLP - final report

Lang.	Nouns	Proper nouns	Verbs	Adjectives	Adverbs
BG	7,868	1,374	3,363	3,207	971
DA	7,442	1,964	4,565	2,662	972
EN	7,449	139	2,933	2,827	1,018
ES	7,974	10	4,303	2,871	1,159
ET	8,082	699	3,963	1,697	1,428
HU	6,859	14	2,488	3,472	1,141
IT	7,950	85	3,097	3,169	1,426
NL	7,179	112	3,036	2,675	552
PT	7,519	1,787	4,059	2,662	1,144
SL	7,430	13	2,569	3,503	1,071

Table 12. Sense-annotated tokens by part-of-speech.

3.2 Language-specific lexical-semantic analysis

In this section, we provide a lexical-semantic analysis focusing mainly on both the lexical and the semantic perspective.



3.2.1 Bulgarian

During the first phase of the annotation we tokenized, lemmatised and part-of-speech tagged the Bulgarian translations. The simple and derived words, including the proper names, contractions, abbreviations and numerical expressions, were automatically annotated with the Bulgarian language processing chain (Koeva et al., 2020). This ensured the correct tokenization, part-of-speech tagging and lemmatization of homonymous verb particles, personal and possessive pronouns, derived numerals and proper names which were not present in the morphological dictionary. The main effort during the manual evaluation and correction was directed towards the re-annotation of multiword named entities as proper names. There are fixed multiword named entities which do not change either in terms of word order or grammar (Yuzhna Amerika ‘South America’) and semi-fixed multiword named entities which also have fixed word order but their constituents in Bulgarian can undergo certain paradigmatic changes within certain grammatical categories (for example, Britanski muzey ‘British museum’ – singular, indefinite, and Britanskiya muzey ‘the British museum’ – singular, definite). Some parts of the multiword names which can be used separately as common nouns had to be marked as proper nouns (for example, all constituents at the organization name Evropeyski socialen fond ‘European Social Fund’, etc.). The lemmas of semi-fixed multiword names in many cases were re-annotated because they differed from the lemmas of the corresponding simple words (for example, the lemmas of the words ruskata ‘Russian’ and pravoslavna ‘orthodox’ from the named entity Ruskata pravoslavna carkva ‘Russian Orthodox Church’ were changed from singular masculine to singular feminine).

3.2.2 Danish



Subtokenisation, lemmatisation, and pos-tagging. Prior to the automatic tokenisation, lemmatisation and pos-tagging, native Danish speakers corrected and adjusted the translation of the English dataset into Danish. In the task of correcting the automatic (sub)tokenisation, lemmatisation and pos-tagging, the most prominent challenge in the Danish dataset was how to deal with compounds, which, as for most Germanic languages, are quite common and relatively dynamically generated, and more importantly: they are written as a single word. Our decision across all 10 languages was that conventionalized compounds found in the dictionary of the language should be kept as such, while compounds not found in the dictionary should be subtokenised, i.e. split into lemmas included in the dictionary, in our case the Danish Dictionary DDO, so as to enable each lemma to be semantically tagged. When splitting compounds with a binding element, e.g. ‘s’ in ‘helbredsansliggende’ (health matters), we decided to keep the binding element during the subtokenisation and POS-tagging phase and to remove it in the lemmatisation phase.

Another problem pertaining to compounds concerned the quite frequent phenomenon where two compounds that share a head are split and one head is left out, as in ‘certificerings- og revisionsmyndighed’ (certification and audit authority). One possibility was to insert the head for both parts ‘certificeringsmyndighed’ og ‘revisionsmyndighed’ in the subtokenisation phase, but we decided that the head in the second part suffices for the disambiguation task and consequently we lemmatised it as ‘certificering’.

The DDO was also used in the cases of participles used as adjectives. Participles with adjective entries in the dictionary were annotated as such, e.g. ‘udstrakt’ (Eng. outstretched, fig: extensive), while those that had only verb entries in the dictionary were annotated as verbs, e.g. ‘samlet’ (Eng. lit: assembled/collected, fig: total).

Sense annotation. The sense annotation was carried out by seven lexicographers at DSL and CST. The sense inventory from the Danish WordNet “DanNet” was used. The WordNet is an open source resource which is linked to the sense inventory of the corpus-based Danish Dictionary DDO, representing approx. half of the senses in the dictionary (which contains around 140,000 senses).



D3.8 Lexical-semantic analytics for NLP - final report

The WordNet mainly contains noun, verb and adjective senses, but only to a small degree the senses from adverbs and multiword units in the DDO dictionary. In the case of missing senses needed for the annotation task, some of which are in fact often frequent in the corpus, the lexicographers studied the DDO and added the appropriate sense from the dictionary. In this way, the WP3 annotation task also gives very useful information when it comes to future extensions of the sense inventory of DanNet. Only rarely did we find senses not represented in the DDO, the exception being proper nouns, which were sometimes found in BabelNet.

Since DanNet only includes very few adverbs senses, we initially planned that only adverbs derived from adjectives should be annotated (with the adjective sense), however we ended up extending the sense inventory with many senses from adverb lemmas in the DDO.

100 sentences were annotated by three lexicographers in order to be able to calculate the interannotator agreement.

Due to the rather high number of annotators and the risk of divergences of the more technical part of the task, e.g. the treatment of multiword units, all 2000 senses were afterwards checked again in the last phase to obtain a higher degree of consistency across the data. In this round, also many proper nouns (some of which in the first round had been annotated with BabelNet) were added to the Danish sense inventory.

3.2.3 Dutch

As a first step in the data preparation for Dutch, the translations of the English sentences were manually checked and corrected. The adopted strategy was to stay as close as possible to the English source sentence. Proper names were replaced if there was a clear Dutch equivalent, but in most cases, the spelling as used in English was kept. The spelling of all sentences was changed if needed so that it conforms to the latest Dutch spelling reform (2015).

The next step was the manual correction of the automatic (sub)tokenisation, lemmatisation and POS-tagging. Similarly to Danish, compounds also represented a specific challenge for



Dutch. Initially, a compound was subtokenised if it did not occur in the Van Dale dictionary³ following the general principle that was set across all 10 languages. Later, this criterion was slightly relaxed for Dutch and some other transparent compounds were also subtokenised, as we observed that a substantial number of compounds would not otherwise be found in the sense inventory. As in Danish, the binding element of compounds was kept in the subtokenisation phase, but removed in the lemmatisation one. Overall, 620 compounds were subtokenised in the Dutch dataset, mostly in two parts, but sometimes even in three or four parts (e.g. *laryngotracheobronchopneumonitis laryngo*; *tracheo*; *broncho* and *pneumonitis*). An important subclass of compounds in Dutch is formed by the separable complex verbs. These are combinations of a verb and some other word. Examples are *bekendmaken* ‘to announce’, *plaatsvinden* ‘to take place’. They sometimes behave as one word (*het kan plaatsvinden* ‘it can take place’) and sometimes as two (*wanneer vindt het plaats?* ‘when does it take place?’). Separable complex verbs are a known problem in corpus linguistics in Dutch and they presented another challenge for the annotation task. According to the UD guidelines, which are based on a lexicalist view of syntax, separable verbs should be annotated as separate words if they are written as separate words and the dependency relation should be used to identify them. After long discussions, it was decided to deviate from the UD guidelines and to consistently lemmatise separable complex verbs with the ‘complex’ lemma, regardless of whether the parts were separated or not. The latest version of the Alpino parser⁴ also does this and lemmatises separable complex verbs with the ‘complex’ form, including an underscore to mark that it can occur as one word or as two, e.g. *aan_vallen*. The advantage of lemmatising with the complex verb is that the whole verb will be automatically looked up in the semantic annotation phase. This is important, as the meaning of separable complex verbs is not always compositional. Moreover, in some instances the parts of a separable complex verb are not even existing Dutch lemmas, as in the

³ <https://zoeken.vandale.nl/>

⁴ The Dutch UD corpora consist of data annotation with the Alpino annotation tools and guidelines, but do not yet include this. <https://github.com/rug-compling/alpino>



case of *aanmoedigen* ‘encourage’, which can be split into *aan* and *moedigen*, but where *moedigen* cannot occur on its own.

For the semantic annotation Open Dutch WordNet was used. It is a Dutch lexical semantic database, which was created by removing the proprietary content from Cornetto⁵. A large portion of the Cornetto database originated from the commercial publisher Van Dale⁶ preventing it from being distributed as open source. In order to create Open Dutch WordNet, all the synsets and relations from WordNet 3.0 were used as a basis and existing equivalence relations between Cornetto synsets and WordNet synsets were exploited in order to replace WordNet synonyms by Dutch synonyms. Concepts that were not matched through hypernym relations to the WordNet hierarchy were added, as well as manually created semantic relations from Cornetto. The synonyms, concepts and relations were limited to those on which there were no copyright claims. In addition, the inter-language links in various external resources were used to add synonyms to the resource (Postma et al., 2016).

The sense annotation was carried out in LexTag by two annotators from INT (one lexicographer and one computational linguist having more than 15 years of experience with lexicographic projects at the institute). The Dutch team decided not to split the data set into smaller batches, so that both annotators had access to the whole data set throughout the whole annotation process to ensure consistency.

Since Open Dutch WordNet mainly contains nouns, verbs and adjectives, we focussed on the semantic annotation of those categories. An exception are ordinal numbers (such as *eerste* ‘first’) which are tagged as adjectives in the data following the UD guidelines. These adjectives were disregarded in the semantic annotation for Dutch as including all of them would have meant that a lot of missing senses would have had to be added. Another exception are adjectives and nouns that are part of a separable compound verb, e.g. *gebruik* in *gebruikmaken*. They are not annotated separately. Separable compound verbs are always lemmatised as a whole (regardless of their orthographic occurrence as one or two words) and

⁵ <http://www2.let.vu.nl/oz/ctl/cornetto>

⁶ <https://www.vandale.nl/>



semantically annotated as a whole. For adverbs, the approach originally adopted by the Danish team was followed and only adverbs derived from adjectives have been annotated (with the adjective sense). Proper nouns were not annotated in this stage.

Due to the nature of the sentences (from specific domains such as chemistry and astronomy), there were quite a lot of words which could not be found in Open Dutch WordNet. The missing words and senses were added by the annotators relying in a first instance on available resources at the institute such as *Algemeen Nederlands Woordenboek (ANW)*⁷, and *Woordenboek der Nederlandsche Taal (WNT)*⁸. If a word could not be found, it was defined by the lexicographer in the team. During the annotation process, quite a few MWU were also identified and defined.

In addition to missing definitions, we found that the definitions in the sense inventory were not always sufficient for the task. Either they overlapped, leading to arbitrary choices and inconsistencies or they were too vague, too general or too specific or incomplete i.e. the desired definition was missing.

After annotation was completed, two rounds of automatic consistency checks were run by the ELEXIS WSD technical team, and on the basis of the results from these checks, the data was revised and corrected by the Dutch team.

3.2.4 English

The English subsection of the ELEXIS Parallel Sense-Annotated Dataset is the pivotal dataset from which translation into all other languages has been performed. In the first phase of the annotation, the automatically processed sentences have been manually inspected by two independent pre-trained annotators and corrected with respect to tokenization, lemmatization and part-of-speech classification. The relatively scarcely documented English-specific UD guidelines have been complemented with querying the two largest manually

⁷ <https://anw.ivdnt.org>

⁸ <https://gtb.ivdnt.org/search/>



D3.8 Lexical-semantic analytics for NLP - final report

annotated English UD treebanks – EWT (Silveira et al., 2014) and GUM (Zeldes, 2017), especially for resolving lexicon-based linguistic issues. Among others, these included the tokenisation of compounds (e.g. *cease-fire*), lemmatization of group names (e.g. *Muslims*), classification of determiner-like words (e.g. *its*), and the classification of various types of verb particle (e.g. *speed up*).

In case of discrepancies between the two treebanks, which was often the case with the under-specified lemmatisation layer, specific guidelines were drafted to consolidate the annotation of various phenomena, such as demonyms (e.g. lemma *American* of the form *Americans*), inflected adjectives (e.g. lemma *low* of the form *lower*) and personal pronouns (e.g. lemma *they* of the form *them*). In accordance with the general ELEXIS guidelines and the reference English UD treebanks, the constituents of multi-word named entities were annotated as PROPEN regardless of their original POS class, with function words as an exception (e.g. *United.PROPEN States.PROPEN of.ADP America.PROPEN*).

In the second stage of annotation, the word sense disambiguation on the dataset was performed by four pre-trained annotators who were required to manually assign the correct sense from a given repository to all content words in a sentence according to the guidelines. The 2021 Edition of the Open English WordNet (McCrae et al. 2019) was used as the primary sense repository, however, in case of missing senses, BabelNet (Navigli et al. 2021) was consulted as a secondary repository. If the correct sense was missing from both sense repositories, a new sense definition was created by an expert annotator in the final stage of the annotation campaign.

Rather than raising issues with respect to word sense identification, most issues in the annotation process emerged from the conflicting principles for lemmatization and part-of-speech categorization in the UD annotation scheme on the one hand and the sense repositories on the other, which prevented the LexTag tool to automatically identify the relevant definitions in the repositories. In such cases, the lemma and/or POS category of a word produced in the first stage have been changed to match the lemmatization and categorization in the sense repository. This was especially frequent with content words



occurring in named entities (e.g. changing a *Cup.PROPN* to *cup.NOUN*) and with case- and punctuation-sensitive multi-word units of different kinds (e.g. changing a lemma from *secretary-general* to *Secretary General*). Other issues emerging and requiring additional guidelines included the treatment of function words and phrases, such as *be*, *how* or *going to*, which were treated as non-content words, the treatment of borderline named entities, such as days of the week (e.g. *Tuesday*), which were treated as content words, and the treatment of naturally occurring ambiguity, such as deciding whether the word *footballer* relates to soccer or American football. In such cases, the annotators were asked to choose the most probable of competing interpretations given the surrounding context and common-sense reasoning.

3.2.5 Estonian

Subtokenisation, lemmatisation, and pos-tagging. The manual validation of the tokenisation, lemmatisation and POS-tagging of the Estonian dataset was carried out by four lexicographers at the Institute of the Estonian Language (EKI) and generally followed the Estonian-specific UD annotation guidelines. Estonian uses 16 universal POS categories (all UD categories except PART). Regarding lemmatisation and POS tagging also the EKI Combined Dictionary⁹ was used. The EKI Combined Dictionary is the biggest lexicographic database for modern Estonian compiled in the Institute of the Estonian Language.

In the tokenisation phase manual correction was necessary in the case of non-conventionalised compounds (e.g. *puu- ja juurviljad* (fruits and vegetables)), conventionalised compounds were left as one token. For words with splitting element *Shakespeare'i* (Shakespeare's) the splitting elements were kept during the sub-categorisation, but removed in the lemmatisation phase. The most problematic was POS tagging, since Estonian UD POS tags are very different from other morphological annotators developed for Estonian (e.g.

⁹ <https://sonaveeb.ee/?lang=en>



estNLTK¹⁰) , and also from POS nomenclature used in the EKI Combined Dictionary. UD-specific parts of speech are AUX and DET. Conjunctions are also split into CCONJ and SCONJ. On the other hand, the degrees of comparison of adjectives are analysed as ADJ, while it is common for Estonian to analyse positive, comparative and superlative degrees as separate parts of speech.

Sense annotation. The sense annotation was carried out by three lexicographers at EKI. The sense inventory from the EKI Combined Dictionary was used. As an additional data source (especially in the case of proper nouns) the BabelNet (Navigli and Ponzetto 2012) sense inventory was used. The most problematic was sense annotation for compounds and MWU's. There were many compounds missing in the inventories as Estonian has a very productive word derivation system where you can easily add words together to make new words. Most of the compounds have a transparent meaning hence are not explained in dictionaries. In cases like these, compounds were annotated as 'sense not found'. Joining lemmas as MWU's was done in the semantic annotation phase, hence none of the MWU's had senses in the inventories and lexicographers had to copy and paste them from the original database (from the EKI Combined Dictionary).

3.2.6 Hungarian

In the LexTag sense annotation task we experienced that a relatively large number of lemmas do not have matching senses in the Hungarian sense inventory. (Thus we had to add many new definitions manually.) The reasons are that the available dictionaries are inappropriate for this task (outdated and/or, containing a low number of lemmas). Our source dictionary is The Explanatory Dictionary of the Hungarian Language (A magyar nyelv értelmező szótára), which was compiled back in the 1960s, however, its meaning structures are the richest among the Hungarian explanatory dictionaries. Being outdated in vocabulary, it obviously lacks many modern words such as: sajtóközpont 'press centre', flashmob, show, biodízel 'biodiesel', etc.

¹⁰ https://github.com/estnltk/estnltk/tree/version_1.6



Moreover, the WSD dataset contains many scientific texts with specific terms which are also missing from the basic explanatory dictionaries (e.g. *vöröseltolódás* 'redshift', *progeszteron* 'progesterone', *androidos* 'running Android'). Another, language specific reason: Hungarian as an agglutinative language builds many ad hoc lemmas with suffixes (e.g. *(alacsony) frekvenciá+jú (fononok)* '(low-)frequency (phonons)', *halál+oz+ás+i (ok)* '(cause) of death', *befecskendez+és+es (dízelmotor)* 'injection pump diesel engine'). Creating compounds (by combining two or more words into one new lemma) is also very common in Hungarian (e.g. *gyermekfocicsapat* 'children's football team', *leopárdalfaj* 'subspecies of leopards' and *repülőgép-baleset* 'aircraft accident') are.

In the sense annotation of multiword PROPNS and multiword NOUNS with PROPNS-elements we gained interesting experiences about the possibilities of defining PROPNS. Multiword units can be interpreted as a whole, and Hungarian orthography helps this interpretation in most cases. Multiword NOUNS, like *Schmidt-távcső* ('Schmidt telescope') on the one hand, can be regarded as one lemma, on the other hand, a combination of a PROPNS-lemma and a NOUN-lemma. However, we also have similarly built (hybrid) multiword PROPNS: *Dusit-palota* ('Dusit palace') and *Fekete-tenger* ('Black Sea'). *Dusit* is a proper name in itself (without "meaning" in the traditional sense), while *palota* ('palace') has the same meaning as when it stands in itself. Considering the PROPNS *Fekete-tenger*, neither of its elements: neither *fekete* 'black' nor *tenger* 'sea' are PROPNS, thus can also be defined as an ADJ and a NOUN. We also had examples of orthographically "unconnected" multiword PROPNS with non-PROPNS elements: in *Oslói Egyetem* ('University of Oslo'), *oslói* is an ADJ ('of Oslo'), while in *Guinness Rekordok Könyve* ('Guinness Book of World Records'), only *Guinness* is a „real” proper name, *rekord* ('record') and *könyv* ('book') have their meanings as NOUNS. It seems that multiple layers of sense annotation will be suitable for solving this problem.

3.2.7 Italian

Translation post-editing. Translations were manually checked and corrected according to the



following priorities: 1. provide missing translations, 2. re-translate really bad sentences, 3. correct typos and other grammar or fluency mistakes. No intervention was made in case of non literal translations, if the meaning was deemed to be preserved. Translation was performed by one of the researchers of the Italian team with a background in Linguistics and a degree in English language and literature. Difficulties were encountered especially with sentences on Physics, Mathematics and other specific scientific domains.

Manual validation and correction of the linguistic annotation. The manual validation and correction of the automatic linguistic annotation of the Italian dataset was performed by a researcher of the Italian team, who generally followed the Italian-specific UD annotation guidelines¹¹ as well as the praxis established by the Italian treebanks. The UD praxis was followed also in cases of clashes with project-level indications, with a few exceptions:

a) abbreviations are treated as single words that may contain punctuation (e.g. U.S.A., UE) except when they indicate units of measure, in this case they are annotated as SYM as in the rest of the datasets;

b) foreign words are annotated as X in titles and long expressions (i.e. when they are incidentals), as they will most probably be annotated as NEs in the following steps.

For the rest of the cases, following the UD guidelines means that: at the level of part-of-speech (POS), nominalised base infinitives (e.g. *il mangiare* lit. the eating, ‘foodstuff’) and participles used predicatively (i.e. introducing implicit relative clauses) are annotated as verbs even when the subject is implied; participles used attributively are annotated as adjectives instead. Possessive adjectives are always tagged as determiners, while pre-determiners and quantifiers are tagged as such if no other determiner is present, adjective otherwise (e.g. *mio* DET *padre* NOUN, ‘my father’, but *il* DET *mio* ADJ *gatto* NOUN, ‘my cat’). Regarding verbs, in the Italian treebanks AUX is generally used also for copulas, so that the verb *essere* “to be” is almost always an AUX. At the level of subtokenisation, sub-tokens in Italian UD is required in

¹¹ <https://universaldependencies.org/it/>



the following cases: 1) complex prepositions (i.e. combined/fused with the definite article, e.g. *nella* “in the.fem”, *del* “of the.masc”); and 2) verbal forms with enclitic pronouns (e.g. *dammelo* “give-to me-it”, *mangiandolo* “eating-it”). Given that this is a quite frequent phenomenon in Italia, the automatic annotation was almost always correct and little manual correction was required. Lemmatisation on the contrary presented many errors and required consistent intervention especially in the case of homograph, irregular and infrequent words. Nouns and verbs are lemmatised canonically with their base forms. Articles and pronouns are lemmatised with their base form (i.e. singular masculine); adjectives with the positive, singular, masculine forms, except for irregular comparative and superlative forms which are lemmatised independently with their masculine singular form.

Sense inventory. The Italian sense inventory was derived by merging and integrating two existing openly available Italian computational databases: PAROLE-SIMPLE-CLIPS (PSC, Lenci et al. 2000) and ItalWordNet (IWN, Roventini et al. 2003). Although their coverage is limited compared with traditional dictionaries they are nonetheless high quality as they were fully manually curated by lexicographers and based, at least for sense definition. PSC, developed within two subsequent European projects PAROLE and SIMPLE, is a large lexical database for the Italian language. In the semantic layer, the main basic blocks are semantic units, Usems, which are provided with definitions and examples; semantic units are linked both to an internal type system (the SIMPLE Ontology) and to other Usems through a rich set of semantic relations (Bel et al., 2000). ItalWordNet (IWN) is a lexical semantic network started within the context of the EuroWordNet project and subsequently enlarged and refined within national projects until 2012. It is mapped and linked to the Princeton WordNet – thus also indirectly, to BabelNet. PSC and IWN were partially mapped in past projects, so that a subset of IWN synsets is linked to PSC corresponding Usems. The mapping is encoded in a separate database preserved at the CNR-ILC (Roventini and Ruimy 2008; Roventini, Ruimy, Marinelli et al. 2007). In order to produce the current sense inventory, the two resources were queried for all the target lemmas present in the Italian ELEXIS dataset and a list of corresponding USEms from PSC and IWN synsets were retrieved together with their definitions, examples and original



D3.8 Lexical-semantic analytics for NLP - final report

identifiers, yielding an 87% coverage. Where a mapping between the two resources was available in the mapping DB, a unique sense was produced, and the definitions merged into a single one by simple concatenation. The resulting sense inventory contains 4,424 lemmas for a total number of 11,298 senses. As PSC and IWN mostly encode nouns and verbs, adjectives and adverbs are highly under-represented. Also multi-word expressions are almost completely missing from the sense inventory, as only IWN contains a few of them as synset members.

Sense annotation. The semantic annotation was carried out at CNR-ILC by 4 students of the University of Siena with a background in linguistics and foreign languages, during their curricular internship. The annotators were closely supervised by two researchers involved in the task, with a strong background in linguistics and annotation. The sense inventory derived from PSC and IWN as described above was used in conjunction with BabelNet in order to account for coverage issues. When none of the resources provided an adequate sense, it was created within the LexTag interface. This will have the positive side-effect of extending the original resources. Proper nouns are almost exclusively annotated with BabelNet as they are not of interest for the internal databases. After annotation was completed, on the basis of the automatic consistency checks run centrally by the ELEXIS WSD technical team, the dataset was revised and corrected whenever possible. This revision highlighted some interesting issues which have no easy fix as they originate from the underlying morphosyntactic annotation and UD guidelines. For instance, the Italian UD guidelines recommend annotating ordinal numbers as adjectives and not as numerals; this would require them to be semantically annotated which is impractical and possibly not useful. When a sense in PSC+IWN or BabelNet is present (e.g. primo ‘first’, second ‘secondo’, terzo ‘third’), the token was annotated.

A subset of the sentences (about 300) were annotated by two annotators, in order to be able to calculate the inter-annotator agreement. These were annotated after the annotator passed a training period, but without confrontation. Simple Cohen’s K was calculated on all



tokens of the dataset, which yielded an agreement of 0,73. At manual inspection, as expected, we saw that in many cases disagreement involves instances of MWUs. Also sometimes disagreement is actually at POS level rather than at sense level. This happened in particular with a class of adjectives/quantifiers that can be used as (pre-)determiners and which, according to UD, should be tagged as DET when they introduce a noun in the absence of other determiners (typically articles or demonstratives), but ADJ otherwise. In such cases, one annotator was more conservative and tended not to override the original POS tagging, while the other preferred to annotate the tokens with an adjectival sense.

3.2.8 Portuguese

We started by correcting the translations of English dataset (missing translations; correcting typos; improving translations, etc.).

The manual validation of the tokenisation, lemmatisation and POS tagging of the Portuguese dataset was carried out by the Portuguese team. One of the major challenges in annotating the Portuguese dataset was presented by lemmatisation in a dictionary that did not always abide by the same annotation criteria applied to corpora. We decided to always annotate the lemma as being the canonical form during the lemmatisation process, ignoring some of the lexical items identified that occur as a headword in a dictionary. For instance, the personal pronoun *ela* (she), the Portuguese feminine form of *ele* (he), is registered as an entry in the Portuguese dictionary, and the recorded lemma is the canonical form *ele*. The option we decided on guarantees better data consistency and coherency. In dictionaries, cases of this type often turn out to be cross-referenced to the canonical form, e.g., the definite article *a* [the Portuguese feminine form of ‘the’] is a cross-reference to *o* [the Portuguese masculine form of ‘the’], which strengthens our decision. Another decision we took concerned the forms corresponding to degrees of adjectives and adverbs. Although in the ACL dictionary we find comparative and superlative forms as headwords, e.g., *pior* (worse; worst), we considered the positive form as a lemma according to Universal Dependencies recommendations.



D3.8 Lexical-semantic analytics for NLP - final report

Generally, for the part-of-speech tagging, we used the Universal Dependencies (UD). Nevertheless, we did not adhere to the UD criterion for abbreviations. Lexical items such as *km* (kilometre) and *m* (meter) were tagged as abbreviations as previously agreed by all ELEXIS team members, rather than as nouns, as UD suggests. It is important to note that we labelled some past participles as adjectives rather than as verbs when they served an adjectival function in the analysed sentences. As for the subtokenisation, contractions were broken into smaller units, for example, *da* (*de + a*) [preposition *de* (of) + the feminine article form *a* (the)]. However, in the case of *desde* (since), which is a contracted form (< prep. Latin *de + ex*), we preferred instead to keep it as a preposition, as recognised by Portuguese grammar and dictionaries.

The sense inventory from the *Dicionário da Língua Portuguesa Contemporânea* (DLPC), published in 2001, from ACL was the dictionary selected. For proper nouns, we used the BabelNet (Navigli and Ponzetto 2012). We had to add a lot of new definitions as we didn't find a large number of matching senses. One of the major difficulties was the fact that the DLPC specifies the meanings in detail, that is, sometimes there are very close meanings and with minimal subtleties of differentiation, so that a meaning detected in a sentence could be associated with two or more meanings in the DLPC.

3.2.9 Slovenian

The process of building the Slovenian sense repository differs from the others in the sense that a pre-existing lexicographic resource was not used for the task. The reason is that Slovenian wordnet (Fišer 2015) is not suitable due to its (semi-)automatic creation, semantic fragmentation and the original English conceptualisation. Furthermore, the Slovenian Academic Dictionary (SSKJ), the only existing monolingual dictionary with semantic description of Slovenian, is not available as an open-access dataset. We have therefore decided to create a new Slovenian semantic repository for this task. For this work we used the Lexonomy dictionary tool (Rambousek 2021), where we imported all the sentences of the



D3.8 Lexical-semantic analytics for NLP - final report

Slovene dataset, which contained approximately 4,000 lexical units. Only nouns, verbs, adjectives and adverbs were semantically disambiguated and defined. We excluded proper names, conjunctions, numerals and other semantically irrelevant elements from the task. Lexonomy tool was used for the creation of the sense repository and for disambiguation of the corpus.

The same sentences from the ELEXIS WSD corpus will also be included in the larger SUK training corpus for Slovenian, formerly known as ssj500k corpus (Krek et al. 2020), and annotation on other levels will be added (Universal Dependency Treebank, Semantic Roles, etc.). Semantic descriptions for all content-word lemmas, which are newly created based on the analysis of the modern standard language corpus, will also be included in the Digital Dictionary Database for Slovenian (Klemenc et al. 2017; Gantar 2020).

A team of 7 lexicographers was involved in the semantic description of analysed headwords. Semantic analysis was carried out on the basis of the Gigafida corpus (Krek et al. 2019). Lexonomy dictionary editor also enabled the use of Word Sketches (SkE) and good examples (GDEX) in several stages (Kilgarriff et al. 2004; Kosem et al. 2019). For the analysis, the lexicographers were given access to the data for headwords that already existed in the lexicographic resources of the Centre for Language Resources and Technologies at the University of Ljubljana, which are available under open access: The Comprehensive Slovenian-Hungarian Dictionary (Kosem et al. 2021), Slovene Lexical Database (Gantar et al. 2013), Thesaurus of Modern Slovene (Krek et al. 2018) etc. This was available for approximately a half of the lemmas. The available data also included semantic types (Kosem and Pori 2021), if this existed for a particular headword in the dictionary database (e.g. adrenalin: 1. BODY-fluid_substance 2. STATE-human-mental) and the synonyms from the Thesaurus of Modern Slovene. In the first phase, the lexicographers' task was to construct a division of senses based on the analysis of the ELEXIS-WSD sentences, to identify possible additional senses of the word in the reference corpus, as well as fixed phrases or phraseological units, and to produce a short description of the identified meanings (sense indicator and/or short explanation). In the second phase, the lexicographer assigned the sentence containing the headword to the



D3.8 Lexical-semantic analytics for NLP - final report

corresponding sense of this headword as an example of usage, as shown in Figure 10.

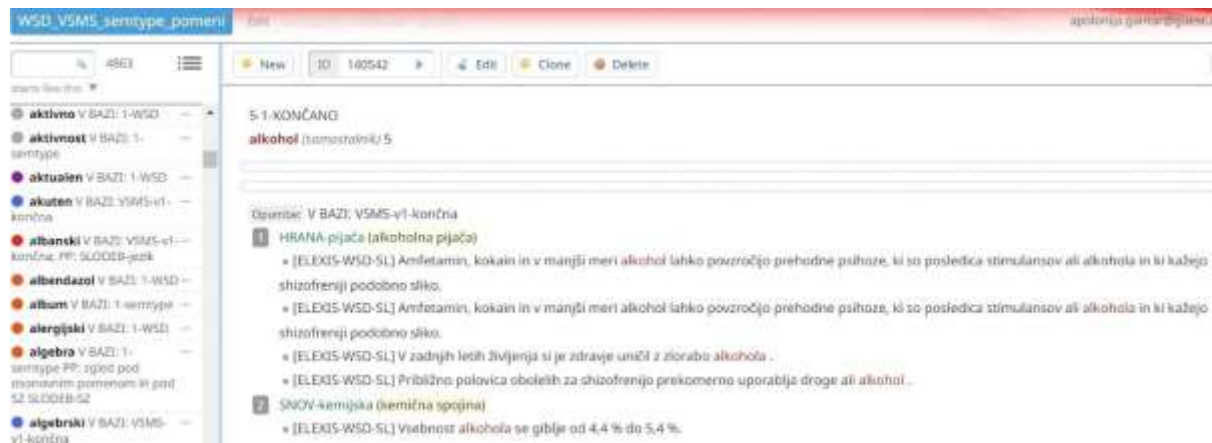


Figure 10: Semantic description of headword alcohol with associated ELEXIS-WSD-sl sentences in Lexonomy dictionary editor.

In the process, the dictionary entry was assigned to another lexicographer, who performed further analysis using the corpus, added collocations to each sense, annotated the words with labels (grammar, domain, register) and added corpus examples to all registered senses. In this way, we were able to speed up the time-consuming lexicographic work and all entries could be produced by a team with consistent decisions in related lexicographic problems. However, due to this arrangement it was not possible to calculate inter-annotator for WSD task.

The biggest challenge - besides the already known sense splitting and lumping problem - was the identification of multi-word units related to the individual senses of the words. In most cases, we tended to identify the multiword unit as a whole (i.e. with a distinctive description of the meaning, e.g. atomska bomba - atomic bomb, blagovna znamka - a brand), but we also assigned the elements of the multiword unit (typically the head of a noun phrase) to the most relevant senses, e.g. svetovna vojna - world war = Concept; svetoven - world = Concept; vojna - war = Concept).



3.2.10 Spanish

All annotation tasks on the Spanish dataset have been done to maintain a high degree of morphosyntactic compatibility with the UD Spanish Ancora Corpus (https://github.com/UniversalDependencies/UD_Spanish-AnCora). For instance, subtokenisation has only been applied to verb+clitic compounds and not to comitatives or to the amalgamated remnants "al", "del". Likewise, perfective tenses, passives, numerical and calendrical expressions have not been joined although a few units have received a compound sense disambiguation, such as "tabla periódica" 'periodic table' and "llevar a cabo" 'to carry out'.

Given the multilingual context of the task and the nature of the ES-Wiktionary a handful of additional criteria had to be applied. WSD has been done following a minimal approach where generic meanings are preferred to very specific ones, when semantically possible. Hence, all verb "poder" 'can, could, be able to' examples have been assigned to the basic notions of "posibilidad" 'ability, probability' and "permiso" 'permission'. Similarly, all auxiliary and lexical uses of "ser", "estar" (both 'be') and "haber" ('have') have been assigned to a few definitions instead of to the vast array of (chiefly contextual) definitions usually found in dictionaries for these items.

Overall, the ES-Wiktionary has a high coverage for the WSD task, with some help from BabelNet. However, a number of additions and modifications has been made, especially on technical and biomedical lemmas. Some definitions, mainly of demonyms and other denominal or deadjectival lemmas related to human or animated beings, have been slightly adjusted as a convenient way of levelling lexicographical anisomorphisms found in the resource.



4 Conclusion

In this deliverable, we illustrated the research activities carried out within task 3.3 (work package 3) focused on lexical analytics for NLP and particularly on sense clustering, domain labeling and the diachronic distribution of senses.

Specifically, we summarized the contributions of the previous deliverables, i.e. a knowledge-based approach to sense clustering called Clusty and discussed the integration of Transformer-based representations to drop the requirement as ongoing and future work. As for domain labeling, we specifically developed an m-BERT-based architecture which takes as input a dictionary definition and outputs the most appropriate domain label such as biology, medicine or law. In this context, we demonstrated the crucial role of the lexicographic data made available within the ELEXIS project, which allows higher performance to be achieved both in English and low-resource language settings. As far as the diachronic distribution of senses is concerned, we conducted a novel diachronic analysis in which we show the evolution of language in terms of most frequent senses over time in multiple languages. With this aim in view, we created corpora in which sentences are annotated with an indication of time. Subsequently, we disambiguated such corpora and investigated the most frequent senses over the course of centuries.

Finally, as an additional contribution to this task and, more generally, to the NLP and lexicographic communities, we introduced the ELEXIS parallel sense-annotated corpus, an entirely manually-curated corpus available in 10 European languages featuring 5 annotation layers.



References

Barba, E., Procopio, L., & Navigli, R. (2021, November). ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1492-1503).

Bel, N., Busa, F., Calzolari, N., Gola, E., Lenci, A., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M. and Zampolli, A. 2000. "SIMPLE: A General Framework for the Development of Multilingual Lexicons". In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019, pages 4171–4186 Minneapolis, Minnesota, June 2 - June 7, 2019*.

Fišer, D. (2015). Semantic lexicon of Slovene sloWNet 3.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1026>.

Gantar, P. (2020). Dictionary of modern Slovene: from Slovene lexical database to digital dictionary database. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 46 (2), pp. 589-602.

Gantar, P. et al. (2013). Slovene lexical database 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1030>.

Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D. (2004) *The Sketch Engine*. Information Technology.



D3.8 Lexical-semantic analytics for NLP - final report

Klemenc, B., Robnik-Šikonja, M., Fürst, L., Bohak, C. and Krek, S. (2017). Technological Design of a State-of-the-art Digital Dictionary. V: GORJANC, Vojko (ed.), et al. Dictionary of modern Slovene : problems and solutions. 1st ed., e-ed. Ljubljana: Ljubljana University Press, Faculty of Arts, 9-22.

Kosem, I. et al. (2021). Comprehensive Slovenian-Hungarian Dictionary 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1453>.

Kosem, I. and Pori, E. (2021). Slovenske ontologije semantičnih tipov: samostalniki. V: Kosem, I. (ed.). Kolokacije v slovenščini. 1. izd. Ljubljana, Znanstvena založba Filozofske fakultete, pp. 159–202.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J., Brank, J. (2020). The ssj500k training corpus for Slovene language processing. V: FIŠER, Darja (ed.), ERJAVEC, Tomaž (ed). Proceedings of the Conference on Language Technologies and Digital Humanities : September 24th - 25th 2020, Ljubljana, Slovenia. 1st ed. Ljubljana, Institute of Contemporary History, 2020, 24-33.

Krek, S., Erjavec, T., Repar, A., Čibej, J.aka, Arhar Holdt, Š., Gantar, P., Kosem, I., RObnik Šikonja, M., Ljubešić, N., Dobrovolc, K., LASKOWSKI, C., Grčar, M., Holozan, P., Šuster, S., Gorjanc, V., Stabej, M., Logar, N. (2019). Corpus of written standard Slovene Gigafida 2.0. Ljubljana: Centre for Language Resources and Technologies, University of Ljubljana. CLARIN.SI data & tools. <http://hdl.handle.net/11356/1320>.

Krek, Simon; et al. (2018), Thesaurus of Modern Slovene 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1166>.



Lafon, Pierre. 1980. Sur la variabilite de la frequence des formes dans un corpus. *Mots*, 1:127–165.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli A. 2000. "SIMPLE: A General Framework for the Development of Multilingual Lexicons", *International Journal of Lexicography*, 13(4): 249-263.

Magnini, B., & Cavaglia, G. (2000, May). Integrating Subject Field Codes into WordNet. In *LREC* (Vol. 1413).

Martelli, F., Kalach, N., Tola, G., & Navigli, R. (2021, August). SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 24-36).

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Sandford Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R., Sancho-Sánchez, J., Lipp, V., Váradi, T., Györfy, A., László, S., Quochi, V., Monachini, M., Frontini, F., Tiberius, C., Tempelaars, R., Costa, R., Salgado, A., Čibej, J., & Munda, T. (2021). Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In *Proceedings of eLex 2021*.

Maru, M., Scozzafava, F., Martelli, F., & Navigli, R. (2019, November). SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3534-3540).



D3.8 Lexical-semantic analytics for NLP - final report

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

Navigli, R., & Martelli, F. (2019). An overview of word and sense similarity. *Natural Language Engineering*, 25(6), 693-714.

Orlando, R., Conia, S., Brignone, F., Cecconi, F., & Navigli, R. (2021, November). AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 298-307).

Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of NAACL-HLT* (pp. 1267-1273).

Procopio, L., Barba, E., Martelli, F., & Navigli, R. (2021). MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation. In *IJCAI* (pp. 3915-3921).

Navigli, Roberto. 2009. *Word Sense Disambiguation: A Survey*, *ACM computing surveys (CSUR)*, 41(2), 10.

Procopio, L., Barba, E., Martelli, F., & Navigli, R. (2021). MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation. In *IJCAI* (pp. 3915-3921).

Rambousek, A., Jakubíček, M., Kosem, I. (2021). 'New developments in Lexonomy' in Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. 5–7 July 2021, virtual. Brno: Lexical Computing CZ, s.r.o.



D3.8 Lexical-semantic analytics for NLP - final report

Roventini, A. et al. 2003. “ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian”. In A. Zampolli, N. Calzolari, L. Cignoni, (Eds.) *Linguistica Computazionale*, vol. XVIII-XIX, Pisa-Roma, IEPI. Tomo II, pp. 745-791.

Roventini, A., Ruimy, N., Marinelli, R., Ulivieri, M. & Mammini, M. 2007. “Mapping Concrete Entities from PAROLE-SIMPLE-CLIPS to ItalWordNet: Methodology and Results”. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 161–164, Prague, Czech Republic. Association for Computational Linguistics.

Roventini, A. & Ruimy, N. 2008. “Mapping Events and Abstract Entities from PAROLE-SIMPLE-CLIPS to ItalWordNet”. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002, February). Multiword expressions: A pain in the neck for NLP. In *International conference on intelligent text processing and computational linguistics* (pp. 1-15). Springer, Berlin, Heidelberg.

SSKJ: Slovar slovenskega knjižnega jezika = Dictionary of Slovene Literary Language. 2nd ed. Ljubljana: SAZU and Fran Ramovš Institute of the Slovenian Language ZRC SAZU, 2014. Also available at: www.fran.si.

Tedeschi, S., Martelli, F., & Navigli, R. (2022, July). ID10M: Idiom Identification in 10 Languages. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 2715-2726).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.



D3.8 Lexical-semantic analytics for NLP - final report

Vial, L., Lecouteux, B., & Schwab, D. (2019). Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. *arXiv preprint arXiv:1905.05677*.

