The Prague Bulletin of Mathematical Linguistics
NUMBER 119 OCTOBER 2022

EDITORIAL BOARD

Published twice a year by Charles University (Prague, Czech Republic)

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: pbml@ufal.mff.cuni.cz

ISSN 0032-6585

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 119 OCTOBER 2022

CONTENTS

Articles

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 119 OCTOBER 2022 5–22

Exploring Explicitation and Implicitation in Parallel Interpreting and Translation Corpora

Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, Heike Przybyl

Saarland University, Department of Language Science and Technology

Abstract

We present a study of discourse connectives in English-German and German-English translation and interpreting where we focus on the phenomena of explicitation and implicitation. Apart from distributional analysis of translation patterns in parallel data, we also look into surprisal, i.e. an information-theoretic measure of cognitive effort, which helps us to interpret the observed tendencies.

1. Introduction

The present paper deals with the phenomena of explicitation and implicitation (Klaudy and Károly, 2005; Blum-Kulka, 1986) in translation and interpreting. From the existing studies, we know that explicitation patterns differ in translation and interpreting (e.g. Lapshinova-Koltunski et al., 2021; Defrancq et al., 2015; Gumul, 2006, amongst others): while translation seems to share increased explicitness, interpreting rather shows a reduction of cohesive ties.

Existing corpus-based studies of explicitation and implicitation often looked into comparable data, contrasting distributions of discourse connectives in the subcorpora of translations and comparable non-translations in the target language (Puurtinen, 2004; Olohan and Baker, 2000). At the same time, detection of explicitation and implicitation effects is believed to require analysis of parallel corpora, i.e. looking at the aligned source texts and their translations (see e.g. Marco, 2018; Zufferey and Cartoni, 2014; Becher, 2011). In this study, we analyse parallel data to verify reported explicitation and implicitation trends in translation and interpreting. We inspect the

 Corresponding author: e.lapshinova@mx.uni-saarland.de
Cite as: Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, Heike Przybyl. Exploring Explicitation and Implicitation in Parallel Interpreting and Translation Corpora. The Prague Bulletin of Mathematical Linguistics No. 119, 2022, pp. 5–22. doi: 10.14712/00326585.020.

translational pairs of discourse connectives in the sources and in the targets,[1] to detect explicitation patterns and strategies. We pay attention to the degree of the explicitation signal. Not all connectives have the same degree of how they signal a discourse relation (Crible, 2020; Asr and Demberg, 2012), as this depends on the number and frequency of other relations they may express. Number and frequency of discourse relations may also depend on the text register a connective occurs in (see details in Biber et al., 1999, p. 880ff). In general, ambiguous connectives express a weaker relation. For instance, a weak signal connective in the source, such as *aber* in example (1-a), being translated by a strong signal connective in the target, e.g. *however* in example (1-b), would indicate explicitation. No explicitation is observed if connectives hold a signal of the same degree, as *aber* transferred to *but* in interpreting in example (1-c).

(1) a. ***Aber** ich glaube, in einer Hinsicht gibt es Einigkeit...*(source)
b. ***However**, I believe that in one respect there is consensus...* (translation)
c. ***but** euh one thing we agree on...*(interpreting)

We analyse the distribution of the explicitation and implicitation cases for the same connectives in translation and interpreting in English and German to compare these phenomena across translation modes. We also seek for the explanation of these phenomena using the information-theoretical notion of surprisal, which indicates cognitive processing effort elicited in translation or interpreting.

We start from selected connectives, for which we reported explicitation and implicitation effects observed in bilingual semantic spaces in a previous study (Lapshinova-Koltunski et al., 2021). Relying on the connective lexicon Connective-Lex (Stede et al., 2019) and occurrences of connectives in a reference corpus (GECCo, Kunz et al., 2021), we estimate their signal strength. This is challenging, since we are looking at two languages and cross-lingual estimation of signal strength is not an easy task. We suggest a classification of connectives and their equivalents according to their explicitation or implicitation effects for both translation directions and explore the variation in these effects in translation and interpreting. We also pay attention to the type of relation the connectives express. Then, we look into the level of information conveyed by the connectives to interpret the results from a cognitive perspective.

The paper is structured as follows: in Section 2, we provide an overview of related studies, in Section 3, we present our methodology. The results are presented and discussed in Section 4. In Section 5, we conclude and give ideas for future work.

[1] In this way, our study is similar to translation spotting (Cartoni et al., 2013) a technique to disambiguate connective meaning, although we are not aiming at sense disambiguation but analyse transfer patterns in terms of explicitation or implicitation strategies.

2. Theoretical Background and Related Work

2.1. Explicitation and Implicitation

As already mentioned above, corpus-based analyses of explicitation and implicitation either look into comparable corpora, defining explicitation as a higher degree of cohesive explicitness in translations if compared to comparable non-translations (Puurtinen, 2004; Olohan and Baker, 2000), or analyse parallel data to uncover transformations from the source text to the target (Marco, 2018; Zufferey and Cartoni, 2014). Explicitation or implicitation effects are often related to the increased usage of discourse connectives and have been extensively analysed so in both human and machine translation (Shi et al., 2019; Meyer and Webber, 2013; Hoek and Zufferey, 2015; Hoek et al., 2015, amongst others). Discourse connectives have also been addressed within the studies on interpreting. For instance, Gumul (2006, p. 184) stated that explicitation in interpreting is related to adding discourse markers among other means of cohesive explicitness. At the same time, Shlesinger (1995) observed a reduction of cohesive ties in interpreting if compared to the source language input (implicitation). And Kajzer-Wietrzny (2012) showed that there are differences between translation and interpreting in the usage of linking adverbials, with translation being more explicit. Defrancq et al. (2015) found that interpreters reshaped the discourse structure of the source speeches in terms of connectives. For our research purposes, we adopt the definition of explicitation introduced by Klaudy and Károly (2005, p. 15): explicitation takes place when a translation contains more specific linguistic units instead of more general units in the source, or new linguistic units not present in the source. Previous studies (Przybyl et al., 2022a) also show that in marking logical relations, interpreters tend to prefer more general items over more specific ones (e.g. *but* vs. *however*), which is typical of spoken production (Crible and Cuenca, 2017), and use fewer different, but polyfunctional discourse markers.

Explicitation and implicitation effects also depend on the type of relations discourse connectives trigger: cognitively simple relations are more often left implicit than relations that are cognitively more complex (see Hoek et al., 2017). This is also confirmed in a recent study by Blumenthal-Dramé (2021) who showed that the processing of concessive sentences benefits more from the explicit marking than the processing of causal sentences, as causal links are more expected than concessive ones. Hoek et al. (2015) also show that explicitation and implicitation maybe affected by expectedness of discourse relations, as defined on the basis of the continuity hypothesis (Murray, 1997) and the causality-by-default hypothesis (Sanders, 2005). In our study, we also look at (un)expectedness of discourse relations via connectives as indexed by surprisal (see Section 2.2 below). However, in our previous study on translation and interpreting (Lapshinova-Koltunski et al., 2021), we could not find the effect of relation type on the explicitation and implicitation in the analysed interpreting data, as the analysed neural semantic spaces in interpreting contained more implicitation than translation independent of the relation the connectives triggered.

The real driving force of explicitation is not easy to determine as many factors at once may be involved. We hypothesize that explicitation, on the one hand, facilitates processing for the producers (translator/interpreter), and on the other hand, helps to shape the content for the recipient (audience design). We aim to inspect explicitation through discourse connectives in both comparable corpora – translation/interpreting and comparable originals of the target language, and parallel corpora – the aligned source language inputs and target language outputs. We expect the parallel data to show (1) if discourse connectives are used in translation/interpreting simply because the source texts already contain such items and they are transferred into the target (equivalence); (2) if translators/interpreters leave them out or change them from more specific to more general, e.g. *however* to *but* (implicitation); if translators/interpreters change more general items to more specific ones (explicitation). We assume that from the cognitive perspective, equivalence and implicitation occur to facilitate processing for translators or interpreters. At the same time, implicitation in interpreting is used due to time pressure, which is usually not the case in translation. Explicitation is used to better shape the content for the audience.

2.2. Surprisal

Apart from the distributional analysis of discourse connectives, we also use a probabilistic measure based on Information Theory (Shannon, 1948), i.e. surprisal or *unpredictability in context*. Surprisal adds a direct link to cognition (Teich et al., 2020) as it represents a direct indicator for cognitive processing effort elicited, in our cases by either the source text or the translation output. Surprisal is a word-based measure of cognitive effort (Hale, 2001, p. 4), i.e. highly predictable words, that incur low surprisal, require low cognitive processing effort. Surprisal measures help to analyse language use in terms of rational communication, and account for a trade-off between expressiveness and efficiency.

In translation, and especially in interpreting, cognitive constraints (e.g. time pressure) impact cognitive processing, and as a consequence, language shape. Surprisal measures are assumed to shed a light on such constraints. Therefore, we use surprisal to investigate the driving force of explicitation and implicitation attempting to link them to cognitive processing. A few studies have already used similar information-theoretic measures (entropy, perplexity) to compare translated and non-translated texts (Bizzoni and Lapshinova-Koltunski, 2021; Teich et al., 2020; Martínez and Teich, 2017; Rubino et al., 2016). Mostly, these measures were used for a comparable analysis of texts, i.e. translated or interpreted texts and comparable originals of the target language. Some studies (Martínez and Teich, 2017; Schaeffer et al., 2015) also used word translation entropy indicating how many equally likely translations may be produced for a source word in a given context. Higher translation entropy means more lexical choices for the translator and higher cognitive effort on the translator's side. This was confirmed by Schaeffer et al. (2015) with experimental data.

In our study, we look at surprisal of discourse connectives in the source texts, as well as in translated and interpreted texts. We also compare surprisal of the same connectives across translation and interpreting. We assume that connectives with low surprisal in the source texts would require low cognitive effort for a translator or an interpreter as a recipient. Target side surprisal indicates cognitive effort of a recipient of the translated or interpreted texts (reader or listener). As ambiguous connectives offer more translation options, they require a higher cognitive effort for a translator as a transmitter. We expect that translators and interpreters use equivalence or implicitation strategies defined in Section 2.1 above more often in case of ambiguous connectives, as explicitation strategies would require more cognitive effort of the translator/interpreter. We also assume that surprisal of the same connectives would vary in translation and interpreting, as the interpreting environment poses more constraints.

3. Methodology

3.1. Corpus

As data we use the bidirectional, sentence-aligned corpus versions of Europarl-UdS (Karakanta et al., 2018) and EPIC-UdS (Przybyl et al., 2022b),[2] specifically the English↔German supcorpora. Europarl-UdS includes the officially published original speeches held at the European Parliament, as well as their translations. EPIC-UdS is the spoken counterpart, consisting of transcripts of these speeches and their simultaneous interpretation, without any corrections with respect to the spoken signal. The total number of tokens comprises approx. 26 millions (more details on subcorpora size are given in Table 1 in Appendix). Both corpora are sentence-aligned and include rich annotation (lemma, part-of-speech, dependencies, surprisal). The absolute number of the extracted and analysed instances of connectives amounts to 87366 and 1242 for the written and spoken corpora respectively.

3.2. Methods

We classified different translation options for each connective into three groups, depending on which translation strategy (explicitation, implicitation and equivalence) was used (see Section 2.1 above). This was not a trivial task as meaning, function, distribution and stylistic preferences almost never display one-to-one correspondence in two languages. Relying on the assumption that the degree of how connectives signal a discourse relation depends on number, frequency and text register preferences of the relations, we combined several approaches to decide if translation options were equivalent, explicit or implicit compared to the source. First, we used a German-English bilingual dictionary (Deuter, 2019) to look up the semantically equivalent

[2] http://hdl.handle.net/21.11119/0000-0008-F519-8

connectors for each language. Second, we checked if the logical relations signalled by the German and English connectives overlapped, using the web-based multilingual lexical resource Connective-Lex.[3] Thirdly, we used the GECCo corpus (Kunz et al., 2021), which is annotated with conjunctive relations and contains different spoken and written registers, to compare the distribution among registers for the different connectives. For example, Deuter (2019) lists *falls* as equivalent to *if*. However, the other resources show that *falls* signals fewer logical relations and appears in fewer registers compared to *if*. That means that *falls* can be used in considerably fewer contexts than *if*, which is why is was classified as more explicit. Although a certain amount of subjectivity cannot be denied in this approach, we consider it reliable, as all the three linguists involved in this study agreed on the underlying classification.

We use the tool CQPweb (Hardie, 2012) to extract instances of connectives along with their distributional information from parallel data. The CQP query language allows to include restrictions on the searched structures. We queried for source connectives and looked at the alignments to identify the most frequent corresponding connectives in the target. Then, we queried for these source-target-combinations and extracted frequencies. Once all frequent translation options were identified, we extracted all instances of the source connective without corresponding target connective (*implicit_none*). This does not mean that there is no signal at all in the target. Instead, it means that we consider overt connectives to be stronger signals of logical relations compared to other types of signals like the V1-syntactic construction in example (2).

(2) ***If** we had more cooperation with the Member States, we would not have as many problems* (source). - ***Würde** die Zusammenarbeit mit den Mitgliedstaaten besser funktionieren, hätten wir nicht so viele Probleme* (translation).

Some noise in the data was inevitable. We tried to fine-grain some queries as to exclude non-connective usages. For example, we removed *because of* and only queried for sentence-initial *also*. However, to remove all noise, a manual investigation of the data would have been necessary, which would have been outside the scope of this paper and its exploratory aim.

As mentioned in Section 2.2, to estimate the level of information conveyed by the connectives, we use the information theoretical notion of surprisal. Surprisal measures information content of a word (w) in number of bits, calculated as the negative log base 2 probability of the word in context. Context is defined here as three preceding words, cf. Equation (1).

$$S(w_i) = -\log_2 p(w_i|(w_{i-1}w_{i-2}w_{i-3})) \quad (1)$$

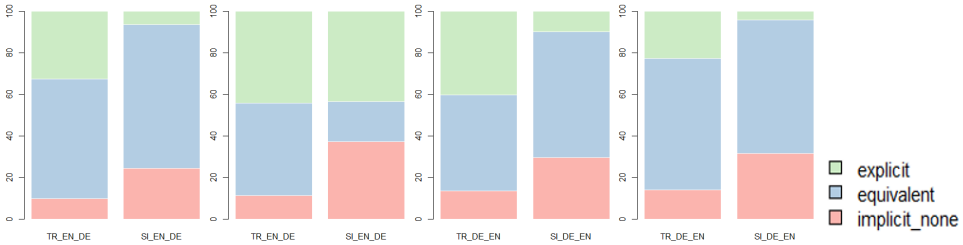
Every word in our corpora is annotated with its surprisal value. We extract the surprisal values for every connective analysed.

[3] http://connective-lex.info/

4. Results

4.1. Translation equivalents

We analyse the frequency distributions of translations as described in Section 3.2 above, paying attention to the translation patterns. An overview of all analysed connectives is given in Table 2 in the Appendix. Due to space restrictions, we visualise translations (explicit=explicitation, equivalent=equivalence, implicit_none=implicitation, i.e. leaving out a connective) of four connectives with stacked bar plots (Figure 1). Generally, we see that there is more explicitation in translation (TR) and more implicitation without overt connectives in interpreting (SI). Equivalence is equally used in both translation modes, with an exception of translation of *because*. In interpreting, we observe more implicitation than equivalence as well as an unusual amount of explicitation for this connective.[4] The amount of explicitation of *because* is the same in both translation modes.



(a) but (b) because (c) aber (d) weil

Figure 1. Distribution for translation/interpreting of but, because, aber and weil

4.2. Surprisal

Now we look into the cognitive processing effort elicited by connectives. First, we compare surprisal of the same connectives in original written and spoken production to translations and interpreting to see if their predictability in context, and thus the cognitive processing load involved, differs within the written and spoken mode. Figures 2 and 3 visualise surprisal values for the four selected connectives in comparable originals (ORG) and translation/interpreting (TR/SI) of the same language. As corpus size and transcription differences (e.g. punctuation in written, no punctuation

[4]The reported differences for the connectives used in translation vs interpreting are confirmed by a Pearson's Chi-squared test: *but*, *because* and *aber* p-value < 2.2e-16, *weil* p-value = 3.965e-08.

in spoken) have an influence on surprisal, we cannot compare written and spoken surprisal values directly.

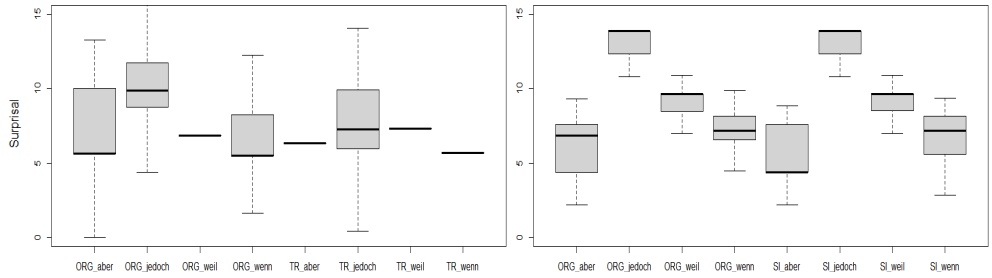


Figure 2. Surprisal in German texts in written (left) and spoken (right) mode

Overall we can observe that for the written mode, most connectives are predictable to the same degree in the given context (visualised by all boxplot quartiles being very close to the median and therefore a line is displayed instead of a box, or narrow boxes with no or short whiskers), which means that processing effort linked to these connectives does not depend on the text production type (translated or not). Translations show less variation concerning surprisal of connectives than written originals, with the exception of *jedoch*. Interestingly, this is the only case where surprisal of a connective is lower in translation than in comparable originals (its processing requires lower cognitive effort). In spoken data, a similar case is observed for *aber* – its surprisal and hence, cognitive processing, is lower in interpreting than in original speech.

In general, surprisal varies more in the spoken data,[5] both in German and English. We do not observe the same tendencies for *however* (equivalent of *jedoch*) in the written and *but* in the spoken data in English). None of the connectives have lower surprisal in either translation or interpreting. Instead, we observe higher surprisal for *because* and *however* in interpreting, which means that they are less expected in interpreting and their occurrence requires a higher processing effort. Comparing surprisal of individual connectives, we observe statistical differences for all connectives studied in both languages for the written mode.[6] Differences are marginal for the spoken mode, with no statistical effect.

We also analyse the parallel data with the focus on the information content of equivalent and explicit use of connectives in the target to see if the strategies used provide any processing bonus for the audience. As stated in Section 2.2, while equiv-

[5]However, differences between the spoken and written mode might be due to mid sentence punctuation being included as context in the written mode which are not transcribed for the spoken data.

[6]Wilcoxon rank sum test with continuity correction for non parametric data: p-value < 2.2e-16.

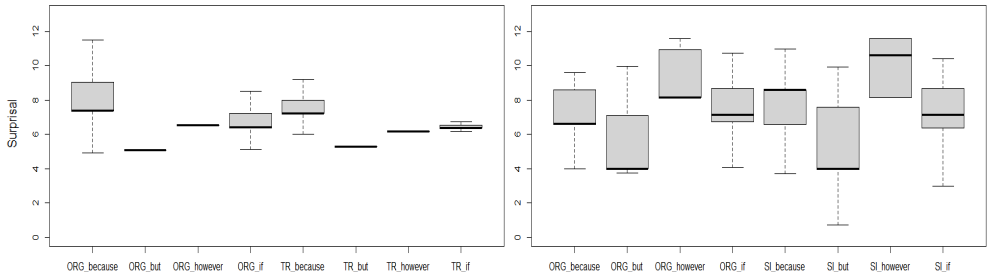
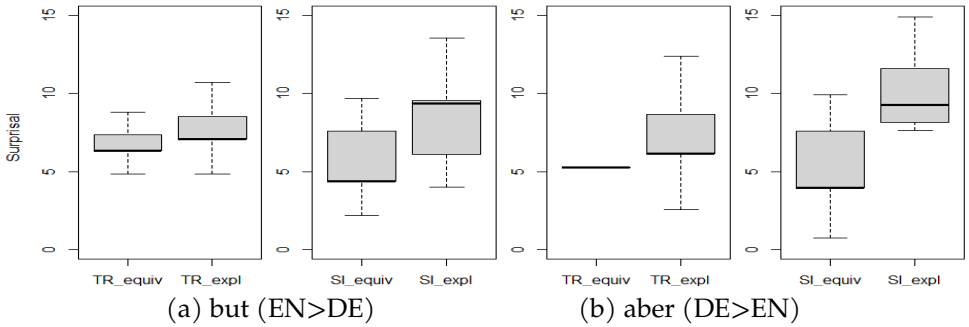


Figure 3. Surprisal in English texts in written (left) and spoken (right) mode

alence is used to reduce the effort on the translator's side, explicitation would mean higher cognitive effort for the translator. However, explicitation in this case could be used for the sake of audience design (to reduce the effort of the reader/listener). Surprisal values are used to compare cognitive processing effort caused by translation patterns (equivalent or explicitation).[7] We extract surprisal of the translations of specific connectives and summarise the surprisal values according to the strategies used. We are not able to analyse processing effort in the cases of implicitation where a connective is left out in the target, as surprisal is calculated for words and we cannot calculate surprisal of a zero.



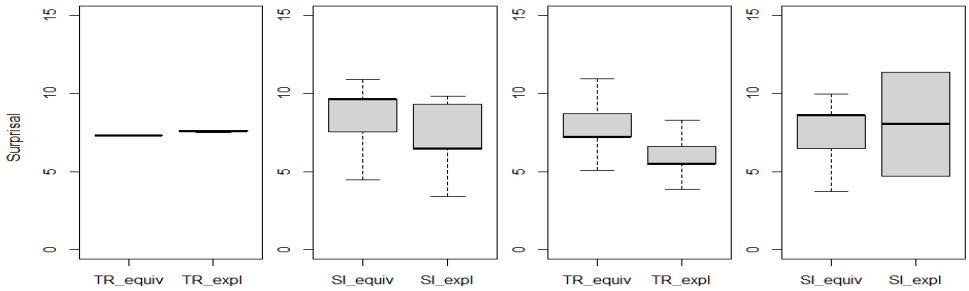
(a) but (EN>DE) (b) aber (DE>EN)

Figure 4. Surprisal for translations of but and aber in in translation and interpreting

We observe differences in processing explicitation across languages for translation of contrast relations triggered by *but* and *aber* in the source (Figure 4): in both trans-

[7]Note that we do not compare surprisal cross-lingually, i.e. between the source connectives and their translations.

lation directions and both modes, explicitation causes a higher cognitive processing effort. The distance between surprisal of equivalent and explicit targets is greater in interpreting than in translation and therefore explicitation causes an even higher processing load for the recipient of the spoken message than the written one. The observed differences are significant for both languages and modes[8].



(a) because (EN>DE) (b) weil (DE>EN)

Figure 5. Surprisal for translations of because and weil in translation and interpreting

For translation of contingency relations triggered by *because* and *weil* in the source, we observe a different tendency (Figures 5). In general, explicitation seems to provide a processing bonus for both translation and interpreting: it either causes a similar processing load as equivalence (German translation and English interpreting)[9] or it even causes a lower processing effort than equivalence (in German interpreting and English translation). This means that processing of this contingency discourse connective benefits more from the explicit marking in both translation modes, and even more in German interpreting and English translation in our data.

This result could appear controversial to the studies showing that cognitively simple relations (e.g. the relation of contingency) are more expected and thus, are more often left implicit than more complex ones (such as comparison). However, in this analysis we miss the cases of explicitation from zero connectives in the source, as well as cases of implicitation to zeros in the target, which does not allow us to report a more comprehensive account of the translation patterns for different relation types.

[8]Wilcoxon rank sum test with continuity correction for non parametric data, TR *but* and *aber*: p-value < 2.2e-16; SI *but*: p-value=4.036e-07, SI *aber*: p-value=2.495e-13.

[9]The reported differences are significant for all subsets apart from for *weil* in the spoken mode. Wilcoxon rank sum test with continuity correction for non parametric data, TR *because* and *weil*: p-value < 2.2e-16; SI *because*: p-value = 0.0004957; SI *weil*: not significant.

5. Conclusion, Discussion and Future Work

The present paper deals with translation of connectives involving explicitation and implicitation effects in spoken and written data. We describe various patterns or strategies for a selected number of connectives for English-German and German-English translation and interpreting, reporting on the distributional preferences across the language pairs. We show that equivalence and implicitation are more frequently used in interpreting than translation, as these strategies facilitate cognitive processing in high-time-pressure situations. Moreover, we analysed surprisal of connectives in comparable source and target context. We showed that the same connectives convey a similar level of information, and hence cognitive load, in written originals and translations in the same language for both modes, with some connectives being less expected in interpreting. We also analysed surprisal of translation patterns to discover that explicitation, although challenging for translator or interpreter in case of ambiguous connectives, may provide a bonus in cognitive processing effort for the recipient, especially in case of translation and interpreting of contingency relations. In this case, as assumed, explicitation is used to shape the content for the audience.

In future, we would like to look into cognitive load conveyed by explicitation, when the source does not contain any signal of a relation and a connective is added in translation/interpreting. As our current approach to surprisal calculation has some limitations (calculated on the word level, depending on word sequences), we plan to use a different surprisal measure, i.e. average surprisal of sentences. First of all, this will be more appropriate for the analysis of connectives, as connective usage depends on relations between clauses and sentences, which is out of the scope of the current surprisal calculation. Moreover, comparing average surprisal of sentences with or without a connective would also compensate for the implicitation cases not covered by the current analysis – with the approach used, we are not able to analyse the cases with an omitted connective in translation/interpreting, as surprisal values were calculated on the level of words. Besides, we will explore interaction of such cross-linguistic constraints as function and stylistic preferences and their impact on the transformation patterns.

Acknowledgements

This paper is based on research conducted in a project funded by the Deutsche Forschungsgemeinschaft – SFB 1102/Project B7, ID 232722074.

Bibliography

Asr, Fatemeh Torabi and Vera Demberg. Measuring the Strength of Linguistic Cues for Discourse Relations. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects*, pages 33–42, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL https://aclanthology.org/W12-4703.

Becher, Viktor. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. PhD thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2011.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman Grammar of Spoken and Written English*. Longman, 1999.

Bizzoni, Yuri and Ekaterina Lapshinova-Koltunski. Measuring Translationese across Levels of Expertise: Are Professionals more Surprising than Students? In *Proceedings of the 23rd NoDaLiDa*, pages 53–63, Online, May 31 - June 02 2021. ACL. URL https://aclanthology.org/2021.nodalida-main.6.

Blum-Kulka, Shoshana. Shifts of Cohesion and Coherence in Translation. In House, Juliane and Shoshana Blum-Kulka, editors, *Interlingual and intercultural communication*, pages 17–35. Gunter Narr, Tübingen, 1986.

Blumenthal-Dramé, Alice. The Online Processing of Causal and Concessive Relations: Comparing Native Speakers of English and German. *Discourse Processes*, 0(0):1–20, 2021. doi: 10.1080/0163853X.2020.1855693. URL https://doi.org/10.1080/0163853X.2020.1855693.

Cartoni, Bruno, Sandrine Zufferey, and Thomas Meyer. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue and Discourse*, 4:65–86, 2013. doi: 10.5087/dad.2013.204.

Crible, Ludivine. Weak and strong discourse markers in speech, chat and writing: Do signals compensate for ambiguity in explicit relations? *Discourse Processes*, 2020. doi: 10.1080/0163853X.2020.1786778.

Crible, Ludivine and Maria Josep Cuenca. Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8:149–166, 2017.

Defrancq, Bart, Koen Plevoets, and Cédric Magnifico. Connective Items in Interpreting and Translation: Where Do They Come From? *Yearbook of Corpus Linguistics and Pragmatics*, 3:195–222, 2015. doi: 10.1007/978-3-319-17948-3_9.

Deuter, Margaret, editor. *Das Große Oxford Wörterbuch: Englisch-Deutsch / Deutsch-Englisch*. Oxford University Press, third edition, 2019.

Gumul, Ewa. Explicitation in Simultaneous Interpreting: A strategy or a by-product of language mediation? *Across Languages and Cultures. A Multidisciplinary Journal for Translation and Interpreting Studies*, 7:171–190, 2006. doi: 10.1556/Acr.7.2006.2.2.

Hale, John. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166, Stroudsburg, PA, 2001. Association for Computational Linguistics. doi: 10.3115/1073336.1073357.

Hardie, Andrew. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17:380–409, 2012. doi: 10.1075/ijcl.17.3.04har.

Hoek, Jet and Sandrine Zufferey. Factors Influencing the Implicitation of Discourse Relations across Languages. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK, April 2015. Association for Computational Linguistics. URL https://aclanthology.org/W15-0205.

Hoek, Jet, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. The Role of Expectedness in the Implicitation and Explicitation of Discourse Relations. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 41–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2505. URL https://aclanthology.org/W15-2505.

Hoek, Jet, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, 121:113–131, 2017. doi: 10.1016/j.pragma.2017.10.010.

Kajzer-Wietrzny, Marta. *Interpreting universals and interpreting style*. PhD thesis, Uniwersytet im. Adama Mickiewicza, Poznan, Poland, 2012. Unpublished PhD thesis.

Karakanta, Alina, Mihaela Vela, and Elke Teich. Europarl-UdS: Preserving Metadata from Parliamentary Debates. In Fišer, Darja, Maria Eskevich, and Franciska de Jong, editors, *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018.

Klaudy, Kinga and Krisztina Károly. Implicitation in Translation: Empirical Evidence for Operational Asymmetry in Translation. *Across Languages and Cultures*, 6:13–28, 2005. doi: 10.1556/Acr.6.2005.1.2.

Kunz, Kerstin, Ekaterina Lapshinova-Koltunski, José Manuel Martínez Martínez, Katrin Menzel, and Erich Steiner. *GECCo – German-English Contrasts in Cohesion: Insights from Corpus-based Studies of Languages, Registers and Modes*, volume 355 of *Trends in Linguistics. Studies and Monographs* [*TiLSM*]. De Gruyter Mouton, 2021. ISBN 9783110711073. doi: 10.1515/9783110711073.

Lapshinova-Koltunski, Ekaterina, Heike Przybyl, and Yuri Bizzoni. Tracing variation in discourse connectives in translation and interpreting through neural semantic spaces. In *Proceedings of CODI at EMNLP-2021*, pages 134–142, Punta Cana and Online, 10–11 November 2021. ACL. doi: 10.18653/v1/2021.codi-main.13.

Marco, Josep. Connectives as indicators of explicitation in literary translationA study based on a comparable and parallel corpus. *Target*, 30:87–111, 2018. doi: 10.1075/target.16042.mar.

Martínez, José Manuel Martínez and Elke Teich. Modeling Routine in Translation with Entropy and Surprisal: A Comparison of Learner and Professional Translations. In Cercel, Larissa, Marco Agnetta, and Maria Teresa Amido Lozano, editors, *Kreativität und Hermeneutik in der Translation*. Narr Francke Attempto Verlag, 2017.

Meyer, Thomas and Bonnie Webber. Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-3303.

Murray, John D. Connectives and narrative text: The role of continuity. *Memory and Cognition*, 25:227–236, 1997. doi: 10.3758/BF03201114.

Olohan, Maeve and Mona Baker. Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, 1:141–158, 2000. doi: 10.1556/Acr.1.2000.2.1.

Przybyl, Heike, Alina Karakanta, Katrin Menzel, and Elke Teich. Exploring linguistic variation in mediated discourse: translation vs. interpreting. In Kajzer-Wietrzny, Marta, Silvia Bernardini, Adriano Ferraresi, and Ilmari Ivaska, editors, *Empirical investigations into the forms of mediated discourse at the European Parliament*, Translation and Multilingual Natural Language Processing. Language Science Press, Berlin, 2022a.

Przybyl, Heike, Ekaterina Lapshinova-Koltunski, Katrin Menzel, Stefan Fischer, and Elke Teich. EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1193–1200, Marseille, France, 20-25 June 2022b. ELDA. URL https://aclanthology.org/2022.lrec-1.127.

Puurtinen, Tiina. Explicitation of clausal relations: A corpus-based analysis of clause. Connectives in translated and non-translated Finnish children's literature. In Mauranen, A. and P. Kujamäki, editors, *Translation universals: Do they exist?*, pages 165–76. John Benjamins, Amsterdam/Philadelphia, 2004. doi: 10.1075/btl.48.13puu.

Rubino, Raphael, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification. In *Proceedings of NAACL HT 2006*, pages 960–970, San Diego, California, June 12-17 2016. doi: 10.18653/v1/N16-1110.

Sanders, Ted J. M. Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning*, pages 105–114, 2005.

Schaeffer, Moritz, Barbara Dragsted, Kristian Tangsgaard Hvelplund, Laura Winther Balling, and Michael Carl. Word Translation Entropy: Evidence of Early Target Language Activation During Reading for Translation. In *Technical Committee on Thought and Language (TL)*, Tokyo, Japan, 2015. doi: 10.1007/978-3-319-20358-4_9. URL http://www.ieice.org.

Shannon, Claude E. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Shi, Weijia, Muhao Chen, Yingtao Tian, and Kai-Wei Chang. Learning Bilingual Word Embeddings Using Lexical Definitions. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 142–147, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4316. URL https://aclanthology.org/W19-4316.

Shlesinger, Miriam. Shifts in cohesion in simultaneous interpreting. *The Translator*, 1:193–214, 1995. doi: 10.1080/13556509.1995.10798957.

Stede, Manfred, Tatjana Scheffler, and Amália Mendes. Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours*, 24, 2019. doi: 10.4000/discours.10098.

Teich, Elke, José Martînez Martînez, and Alina Karakanta. Translation, information theory and cognition. In Alves, Fabio and Arnt Lykke Jakobsen, editors, *The Routledge Handbook of Translation and Cognition*, chapter 20. Routledge, London, 2020. ISBN 9781138037007. doi: 10.4324/9781315178127-24.

Zufferey, Sandrine and Bruno Cartoni. A multifactorial analysis of explicitation in translation. *Target*, 26(3):361–384, 2014. doi: 10.1075/target.26.3.02zuf.

Appendix

Europarl UdS	tokens	EPIC UdS	tokens
ORG WR EN	8,693,135	ORG SP EN	67,526
TR EN DE	3,100,647	SI EN DE	57,532
ORG WR DE	7,869,289	ORG SP DE	56,488
TR DE EN	6,260,869	SI DE EN	58,503

Table 1. Corpus overview EUROPARL UdS (written) and EPIC UdS (spoken)

Source	Target	Category	fpm TR	fpm SI
		CONTINGENCY		
if	wenn	equivalent	1,482.27	1,585.48
if	falls	explicit	75.51	30.20
if	ob	equivalent	113.69	45.30
if	none	implicit_none	490.70	875.79
wenn	if	equivalent	787.00	1,831.08
wenn	when	equivalent	32.40	35.55
wenn nicht	unless	equivalent	13.98	0.00
wenn nicht	albeit	explicit	678.23	408.88
auch wenn	even though	equivalent	56.28	17.78
wenn	none	implicit_none	705.69	1,084.43
because	weil	equivalent	574.69	423.00
because	denn	explicit	407.83	921.00
because	da	explicit	161.49	30.00
because	none	implicit_none	147.00	815.39
weil	because	equivalent	852.69	1,084.43
weil	since	explicit	53.89	35.55
weil	for	explicit	77.25	0.00
weil	as	explicit	175.83	35.55
weil	none	implicit_none	191.52	533.33
		TEMPORAL		
finally	schließlich	equivalent	79.76	60.40
finally	abschließend	explicit	104.08	0.00
finally	zu dem Schluss	explicit	14.99	15.10

continued on next page

continued from previous page

Source	Target	Category	fpm TR	fpm SI
finally	zu dem Abschluss	explicit	7.35	15.10
finally	zu guter letzt	explicit	0.85	15.10
finally	dann	implicit	7.92	60.40
finally	letztlich / letztendlich	explicit	7.35	15.10
finally	ein/als letztes	explicit	0.57	0.00
finally	letztens	explicit	2.26	0.00
finally	letzter Punkt / letzte Anmerkung / letzte Bemerkung / letztes Wort	explicit	2.55	0.00
finally	nicht zuletzt	explicit	2.55	0.00
finally	none	implicit_none	92.77	135.90
schließlich	finally	equivalent	34.62	17.78
schließlich	ultimately	equivalent	10.06	0.00
schließlich	in the end	explicit	3.24	0.00
schließlich	after all	explicit	21.32	0.00
schließlich	at the end of the day	explicit	6.48	0.00
schließlich	lastly	explicit	3.58	0.00
schließlich	at last	explicit	0.85	0.00
schließlich	last but not least	explicit	0.85	0.00
schließlich	in the final analysis	explicit	2.05	0.00
schließlich	in conclusion	explicit	0.85	0.00
schließlich	none	implicit_none	20.29	17.78
	EXPANSION			
also	außerdem	equivalent	7.64	0.00
also	auch	equivalent	4.81	45.30
also	ebenfalls	explicit	1.98	0.00
also	ebenso	explicit	1.13	0.00
also	darüber hinaus	explicit	3.11	0.00
also	gleichfalls	explicit	0.28	0.00
also	ferner	explicit	3.11	0.00
also	des Weiteren	explicit	0.57	0.00
also	zudem	explicit	3.39	0.00
also	und	implicit	1.98	30.20
also	hinzu kommt	explicit	1.13	0.00
also	zusätzlich	explicit	0.28	0.00
also	none	implicit_none	6.50	15.10

continued on next page

continued from previous page

Source	Target	Category	fpm TR	fpm SI
außerdem	also	equivalent	47.92	35.55
außerdem	furthermore	equivalent	7.50	0.00
außerdem	moreover	equivalent	17.22	0.00
außerdem	in addition	explicit	19.95	0.00
außerdem	what is more	explicit	1.19	0.00
außerdem	besides	equivalent	1.88	0.00
außerdem	apart from that	explicit	1.02	0.00
außerdem	none	implicit_none	12.79	35.55
		COMPARISON		
but	aber	equivalent	1,558.07	3,503.16
but	sondern	equivalent	597.89	392.60
but	jedoch	explicit	393.97	0.00
but	doch	explicit	684.71	286.90
but	allerdings	explicit	83.43	60.40
but	dennoch	explicit	52.32	15.10
but	none	implicit_none	374.17	1,374.08
aber	but	equivalent	2,007.75	2,328.85
aber	however	explicit	1,081.39	248.88
aber	though	explicit	256.83	53.33
aber	although	explicit	139.67	0.00
aber	while	explicit	113.24	17.78
aber	whilst	explicit	35.64	0.00
aber	nevertheless	explicit	43.66	17.78
aber	yet	equivalent	59.18	0.00
aber	nonetheless	explicit	29.67	35.55
aber	none	implicit_none	598.08	1,137.76
however	aber	implicit	109.45	181.20
however	jedoch	equivalent	486.74	15.10
however	doch	implicit	89.09	0.00
however	allerdings	equivalent	156.40	30.20
however	dennoch	equivalent	36.20	0.00
however	none	implicit_none	128.97	45.30
jedoch	however	equivalent	66.17	35.55
jedoch	but	implicit	142.74	17.78
jedoch	though	implicit	21.32	0.00
jedoch	although	implicit	17.74	0.00
jedoch	while	explicit	11.77	0.00
jedoch	whilst	explicit	2.56	0.00

continued on next page

continued from previous page

Source	Target	Category	fpm TR	fpm SI
jedoch	nevertheless	explicit	5.63	0.00
jedoch	yet	implicit	7.84	0.00
jedoch	nonetheless	explicit	2.39	0.00
jedoch	none	implicit_none	36.50	17.78

Table 2. Distribution of connectives and their translations: fpm=frequency per million, TR=translation, SI=interpreting

Address for correspondence:
Ekaterina Lapshinova-Koltunski
e.lapshinova@mx.uni-saarland.de
Department of Language Science and Technology
Campus A2 2, Room 1.02
D-66123 Saarbrücken, Germany

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 119 OCTOBER 2022 23–36

Translating Argumentation: Distributional Semantic Analysis of Argumentation Resources in Parallel Corpora

Vahram Atayan, Bogdan Babych

Institute for Translation and Interpreting, Heidelberg University

Abstract

In the paper, we report results of our experiments on identifying distributional semantic characteristics of different types of lexical items used in argumentation: connectors, meta-argumentative words, key notions of a given discourse and the evaluative/connotative lexicon. These characteristics are contrasted within monolingual English corpora of different genres (Europarl-EN and Cord COVID-19) and in translation context for German-into-English direction (Europarl-DE and Europarl-EN-from-DE). For the analysis, we propose a number of new methods that better characterize distributional semantic differences between the argumentatively relevant lexical items, such as measuring the knee in the mutual information-ranked list curve, testing categories for span variation and different selection procedures. In our experiments, meta-argumentative lexical items show the biggest differences in their distribution with other word types on several of such measures. The analysis based on word vector allows us to create a selection heuristic for candidate lists for different categories of argumentative lexicon.

1. Introduction

The characterization of linguistic resources used in argumentation has been an important challenge for contrastive linguistics, qualitative and corpus-based translation studies (e.g. Atayan, 2007). In particular, different languages build argumentation structures with different inventories of lexical, morphosyntactic, and discursive means, as well as usage patterns, which involve interaction across various linguistic levels and paradigmatic sub-systems of a language (argumentative connectors, evaluative/connotated lexicon, meta-argumentative constructions, etc.). Also syntagmatically the argumentation patterns may span non-local context, extending through

 Corresponding author: atayan@uni-heidelberg.de
Cite as: Vahram Atayan, Bogdan Babych. Translating Argumentation: Distributional Semantic Analysis of Argumentation Resources in Parallel Corpora. The Prague Bulletin of Mathematical Linguistics No. 119, 2022, pp. 23–36. doi: 10.14712/00326585.021.

several sentences and larger discourse units. However, it is difficult to automatically identify and align multiword argumentation patterns in multilingual corpora with sufficient accuracy. This seriously limits the applicability of standard corpus-based methods, tools and annotation resources for their study, since most traditional approaches primarily target phenomena in the local context of corpus searches (e.g., morphosyntactic or lexical patterns within a window of a few words). Specifically, while there have been several corpus-based studies of systematic differences between original texts and translations (so called 'translationese') on the levels of the general lexicon, modal markers, morphosyntactic patterns, indirect equivalents (e.g., Babych et al., 2007; House, 2011; Hoey, 2011; Kranich and Gast, 2015; Gast, 2022), the range of such studies for argumentation patterns across languages has been limited.

Our paper addresses this methodological gap in contrastive corpus-based analysis of argumentation, namely we suggest a number of new distributional semantic properties of argumentation resources, which allow us to quantify differences in their usage across languages and genres or in original and translation corpora within the same language. In the paper we report the results of our experiments on evaluating relevant distributional parameters of three types of argumentation patterns in multilingual and translation context: (a) meta-argumentative words; (b) key notions of a given discourse; (c) evaluatives/connotated lexicon. Lexical items of these three categories have been manually annotated in two different selections of ca. 1000 word types from each of our corpora: (1) The Europarl corpus of parliamentary proceedings (Koehn, 2005), where we selected original texts that have been authored in (1a) German and (1b) English, as well as (1c) English texts translated from German. (2) The CORD-19 English monolingual corpus of medical research articles about Covid-19 (Wang et al., 2020). Both corpora are POS-tagged and lemmatized (Schmid, 2013).

Therefore, this paper seeks to make two kinds of contribution. Firstly, we propose new methods for distributional analysis of argumentatively relevant lexical items, that extend standard collocation-based approaches used nowadays in corpus linguistics to characterize the general lexicon. Secondly, we identify distributional characteristics of the argumentative lexicon that best distinguish different types of argumentation resources and allow for linguistic interpretations that verify or extend existing theoretical models of argumentation. Finally, we use manually annotated lists of argumentatively relevant lexical items of different types to generate automatically further candidates for the three categories using word vector models.

The methods of distributional semantics are typically based on identification of certain *observable* features and annotations in text corpora, their statistical analysis and linguistic interpretation, which often creates a possibility to test quantitative predictions made by linguistic models and provides new linguistic insights or improved understanding of the phenomena, modelling their structure and interaction, possibly leading to new testable predictions. Because such explicit features are more easily found on the lexical and morphosyntactic levels, i.e., in the local context of linguistic constructions, there is a certain 'street-light effect' within the current research

paradigm, which has mostly focussed on phenomena below the sentence level. This also applies to the study of cross lingual phenomena in corpus-based contrastive linguistics and translation studies (e.g., Kruger et al., 2011).

Still, corpus-based studies of certain discourse-level phenomena (such as discourse particles) have indicated that specific linguistic properties of these resources result in different sets of observable features, which distinguish them from the general lexicon. The research suggests that traditional corpus-based methods are much less effective for the study of phenomena on the discourse level, especially when such resources do not have direct translation equivalents (Gast, 2022, 323-324).

Argumentation is a phenomenon of communicative discourse (van Eemeren, 2018; van Eemeren et al., 2019, 5), so applicability of lexically-oriented corpus methodologies would be limited for modelling its distributional semantic properties. Therefore, the distributional analysis of argumentation resources and their translation equivalents across languages would involve the tasks of (1) identifying their characteristic observable features in corpora, also beyond the local context and (2) developing statistical measures that quantify them and could capture distributional properties of different types of argumentation resources, differences in their usage across genres, across different languages, as well as in original vs. translated discourse.

2. Methodology: distributional measures for argumentation lexicon

To analyze the impact of the distributional semantics on the argumentation, we defined four categories of argumentatively relevant lexical items: connectors (like *since*), evaluative or connotated words (like *dangerous* or *progress*), content-related non-connotated key notions of a given discourse (*Commission* for Europarl or medical terms like *venous* for COVID corpus) and meta-argumentative words (like *disagree* or *reason*). For the present study we are concentrating on the last three lexical groups. Our goal is to identify potential differences between these groups concerning their distributional properties, measured via different parameters related to the mutual information (MI) of a given word (with a cut-off at the frequency of 50 items/corpus) and its collocates (with a cut-off at least 10 tokens/corpus) defined as

$$\mathrm{MI} = \log_2(\mathrm{ObservedFreq}^2/\mathrm{ExpectedFreq}), \quad (1)$$

where ObservedFreq and ExpectedFreq are observed and expected frequencies of a collocate in the context of the word calculated in a given span (Evert (2008, 19) considers this relation as most popular in the MI-calculation). $\mathrm{MI} > 1$ is typically considered as indicator for a collocation relation between two words. In the previous research, collocation analyses have been conducted for different spans, ranging from 1 to dozens or hundreds of words; the choice of the specific span is of essential importance in the research (Evert, 2008, 12).

In our analysis, we are taking into consideration collocation spans from 1 to 9 in order to understand the potential differences in the behaviour of different types of

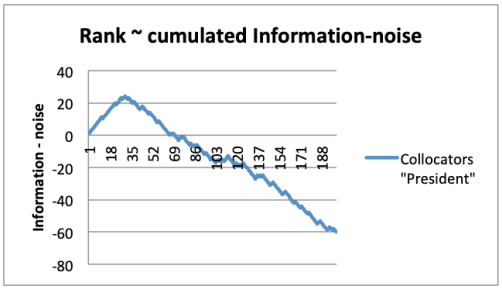


Figure 1: Knee: signal vs. noise in MI-ranked collocates of *president*

argumentative lexemes in local vs. rather clause-level context. Since the number of collocates is normally growing monotonically with the span, we suggest taking into consideration rather the number of most informative collocates measured by MI. Typically, the MI of the collocates of a given word seems to follow a Zipfian distribution with a small number of strong collocates and a wide range of collocates with low MIs, potentially generating more noise than useful information. To improve the information/noise ratio for a data set, it's possible, of course, to simply cut off the list by a fixed number or a selected threshold of MI. Yet both choices are rather arbitrary, so that we suggest instead taking into consideration only the collocates with an above average contribution to the overall MI of the whole collocate set. To do this, we calculate the knee of the discrete curve of decreasing MI values.

As Satopaa et al. (2011) point out, knee points of a curve represent in general the best balance for inherent trade-offs, in our case, between stronger collocates and rather accidental correlations. To illustrate informally the potential usefulness of this concept for our goals, we calculated the first 200 collocates of the word *president* in the English Europarl corpus (using

$$\mathrm{MI}_{lex} = \mathrm{ObservedFreq}/\mathrm{ExpectedFreq} \times \log_2(\mathrm{CandidateFreq}) \qquad (2)$$

as MI measure to select rather fully lexical items). In the next step, we annotated the items clearly related semantically or pragmatically to the lexeme *president*, such as proper names (*Obama*), organizations (*PPE*), countries (*Venezuela*), related political functions (*Chancellor*) and genre related collocates (*madame*) and weighted those with 1. We considered formally as noise non-evident collocates like *too, ask, incoming* etc. and weighted those with -1. Finally, we calculated the rank-dependent overall information as the sum of weights up to the given rank.

In Figure 1 we can easily see that the "noise" (here interpreted as statistically relevant but intuitively non-evident collocates) becomes dominant after first ca. 72 items. The calculation (`kneed` Python package) of the knee of the MI curve for *president* (with 3548 collocates with MI over 1) gives us 74 as knee value, which agrees rather well

with our informal estimation (further studies with more lexemes will be conducted in the future to verify the usefulness of knee as information measure). We consider thus the knee value as a general informative characterization of the collocates set of a given lexeme. To obtain an overall result for our argumentatively relevant items, we calculate the average value of the knees of all elements for every given subset. Linguistically speaking, we are asking for the number of particularly informative collocates for the lexical items and try to understand the behaviour of this parameter, which reflects, for different spans, the interaction of a given word with local and extended context.

In our analysis, the COVID-19 corpus represents a culturally-neutral genre, which has distinct argumentation patterns typical for scientific discourse. The Europarl corpus is more culture-specific and uses argumentation patterns typical for political debates. For the analysis, we focus on genre comparison (COVID vs. Europarl) and on parallel English and German texts, both originally authored and translated. These corpora are used as a benchmark for the proposed methodology. They also allow us to develop linguistic interpretation of the results and to make further testable predictions about the behaviour of argumentation structures in translation. The sizes of the corpora: sub-corpus of COVID: 3.360.000 tokens; EP_EN: 7.281.000, sub-corpus of EP_DE: 2.581.000, EP_ENfDE: 3.854.751.

3. Experiment on knee comparison across different collocation spans

We extracted two different subsets of argumentative lexemes by annotating the categories of evaluative/connotated, key notion and meta-argumentative among ca. 1000 words with the highest average MI of all collocates (linguistically speaking, words with smaller sets of strong collocates) and among ca. 1000 randomly selected items for our four corpora. In general, we would assume a monotonic increase of the knees over spans, because broader spans typically introduce more weak collocates reinforcing indirectly the information contribution of stronger ones, which previously were slightly under the knee point, and then eventually coming to saturation at larger spans. For MI-based selections we expect generally lower knee values, due to the dominance of a small number of strong collocates typical in such cases. This hypothesis is confirmed by our data, cf. Figure 2. As far as genre and translation effects are concerned, we can report the following findings.

1. In the genre comparison, presented in Figure 2, we see the particularity of key notions in the COVID MI-based selection with a steady increase of the knee value over spans 1-9 as compared to evaluative/connotated and meta-argumentative words. This effect is possibly related to the internal inhomogeneity of the corpus texts (scientific articles) with well-defined functional parts (ABSTRACT, METHODS, DISCUSSION etc.). Due to this repartition, some parts of the vocabulary are not uniformly distributed in the texts but concentrated in one or other type of subtext. This could be the case for certain specialized key notions with high

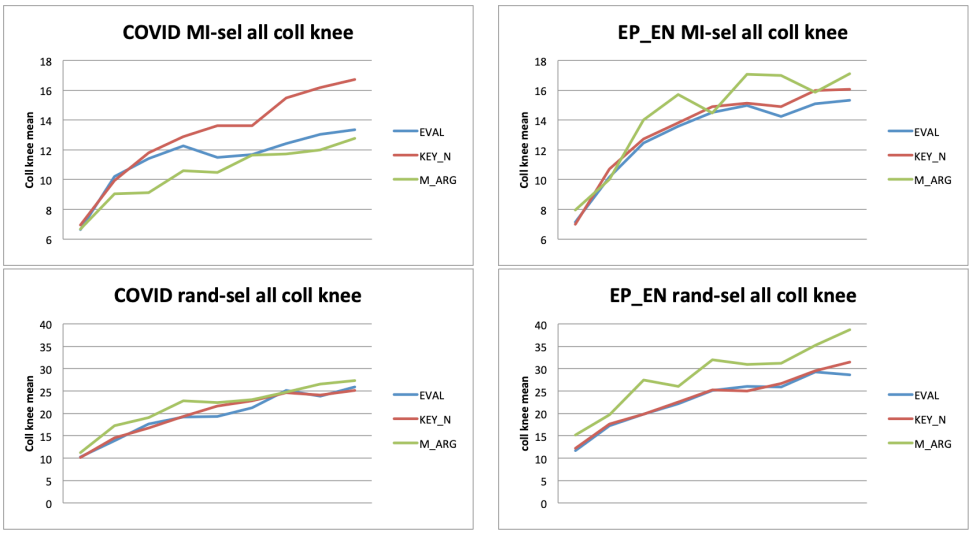


Figure 2: Size of knee in Covid vs. EP corpora: MI-ranked and random selection

average MIs, present in rather technical parts of an article. Now, an extension of the collocation span would in such cases integrate potential new collocates from the same subtext, where a part of the vocabulary would be generally over-represented in comparison with the other parts of the article, even without a specific relation to the original, i.e., collocation-based key notion. So we would have something like keywords of the given subtext type (as compared with the rest of corpus as reference corpus) instead of word collocates. This hypothesis will be tested in our future work by a detailed distributional analysis of the specific subtext types.

2. In the EP corpus, where we don't have a particular internal division of the single typically rather short interventions, such effects are not possible. Instead, it is the category of meta-argumentatives that shows a particular pattern (higher knee values than the other categories and non-monotonic curve with a decrease or plateau around the span of 4). If we analyze specifically the difference in the knee value between the spans 3 and 4 we see that frequent meta-argumentative verbs like *concede*, *reject*, *endorse* or *conclude* have higher collocation knees for span 4, i.e. maintain the monotonic increase. This seems to be related to frequent local sentence initial patterns like "Connector – subject – (negation) – argumentative verb": *Therefore the Commission does not support*, containing 3 or more words with limited variability (personal pronouns, typical political actors like European Commission, Parliament, parliamentary groups, negations etc.).

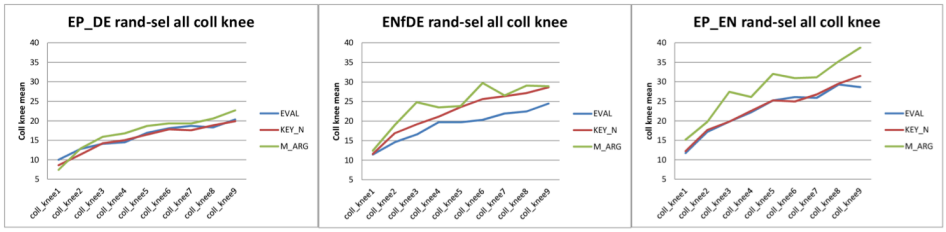


Figure 3: Knee size in Europarl: original DE, EN translated from DE and original EN

On the contrary, adverbs (*clearly*) and meta-argumentative nouns (*stance, contradiction*) with their limited scope and larger diversity of collocates, generate a cumulatively more important decrease of knee value, resulting in the overall non-monotonic pattern.

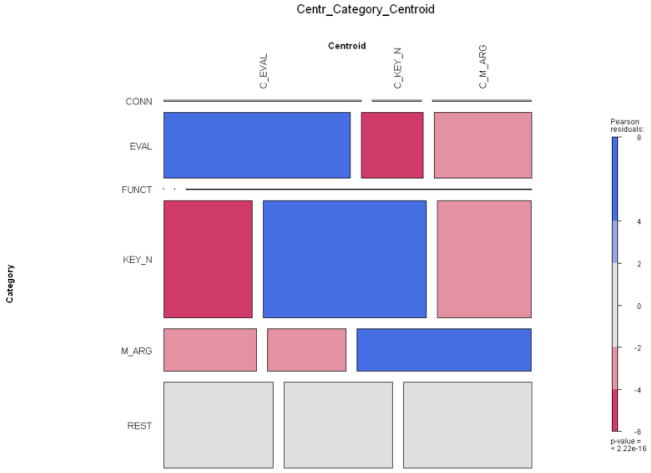
3. The effect concerning the difference between meta-argumentatives and other categories can be observed generally in the English corpora. Figure 3 presents the collocation knee value over spans from 1 to 9 for the original German EP corpus, English translations from German and the English originals for argumentative lexemes annotated in a random selection of ca. 1000 words. Here we see on the one side very similar patterns for the original and translated English corpora, there the meta-argumentatives show the same non-monotonic dynamics. Yet, the knee values are generally lower for the translated corpus and thus more similar to the German originals, possibly due to the general tendency of translation towards reduced variability of linguistic means, which could also reinforce this effect. The more regular pattern of meta-argumentatives in German could be related to the differences in the syntax, in particular the sentence final position of non-finite verb parts (participles and infinitives), as well as finite verbs in subordinate clauses, which reduces the probability of sentence initial pattern building.

4. Experiment on centroid prediction

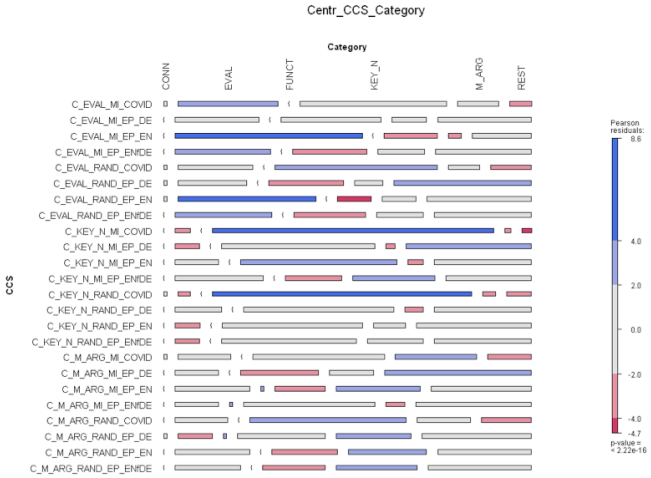
In our further experiment, we are using word embedding models generated with the Python `gensim` package to explore the possibility of semi-automatic identification of argumentatively-relevant lexical items. We use the annotation of the three fully lexical categories (evaluatives/connotated lexemes, key notions and meta-argumentatives) to automatically generate candidate lists and to evaluate the potential improvement in precision as compared to the initial frequencies of lexemes of a given type in our annotations.

			improvement		arg. lex.	
centroid	% initial annot.	% correct predict.	abs.	rel.	overall initial annot.	centroid-based predict.
EVAL_MI_COVID	11.4	32	20.6	2.8	44.4	92
EVAL_MI_EP_DE	11.1	27	15.9	2.4	51.9	70
EVAL_MI_EP_EN	19.5	60	40.5	3.1	54.6	81
EVAL_MI_EP_ENfDE	17.2	31	13.8	1.8	49.5	70
EVAL_RAND_EP_DE	9.4	22	12.6	2.3	48.1	55
EVAL_RAND_EP_EN	12.4	45	32.6	3.6	39.5	67
EVAL_RAND_COVID	11.2	24	12.8	2.1	38.5	86
EVAL_RAND_EP_ENfDE	11.6	31	19.4	2.7	39.4	69
KEY_N_MI_COVID	26.2	90	63.8	3.4	44.4	97
KEY_N_MI_EP_DE	35.5	49	13.5	1.4	51.9	60
KEY_N_MI_EP_EN	31.0	50	19.0	1.6	54.6	69
KEY_N_MI_EP_ENfDE	26.3	18	-8.3	0.7	49.5	72
KEY_N_RAND_EP_DE	29.7	48	18.3	1.6	48.1	69
KEY_N_RAND_EP_EN	18.2	44	25.8	2.4	39.5	62
KEY_N_RAND_COVID	18.9	83	64.1	4.4	38.5	91
KEY_N_RAND_EP_ENfDE	20.1	43	22.9	2.1	39.4	69
M_ARG_MI_COVID	6.7	26	19.3	3.9	44.4	85
M_ARG_MI_EP_DE	5.4	14	8.6	2.6	51.9	53
M_ARG_MI_EP_EN	4.1	27	22.9	6.6	54.6	67
M_ARG_MI_EP_ENfDE	6.0	6	0.0	1.0	49.5	62
M_ARG_RAND_EP_DE	9.0	24	15.0	2.7	48.1	63
M_ARG_RAND_EP_EN	8.9	25	16.1	2.8	39.5	70
M_ARG_RAND_COVID	8.4	17	8.6	2.0	38.5	84
M_ARG_RAND_EP_ENfDE	7.7	26	18.3	3.4	39.4	67

Table 1: Improvements with centroid-based prediction for argumentative lexicon.



(a) Main centroid types



(b) Sub-categories

Figure 4: Pearson's residuals of distributions across categories of centroids

Our starting point is built by the annotations of eight selections of ca. 1000 items each from our corpora ({COVID, EP_EN, EP_DE, EP_ENfDE}x{MI-based selection, random selection}), annotated with 6 categories {CONN, EVAL, FUNCT, KEY_N, M_ARG, REST}. Our three fully lexical argumentative categories cover in different corpora between 9.4% and 19.5% (evauatives/connotated lexicon), 18.2% and 35.5% (key notions) and 4.15 and 9% (meta-argumentatives) of the initially annotated items with an overall percentage of argumentative elements ranging from 38.5% (COVID, random selection) to 54.6% (EP_EN, MI-based selection). We built then word embedding models for four corpora with a window of 5 items to the right and to the left, with vector size 100, taking into consideration only items with a frequency over 10. In the next step we calculated the word vectors for the items in our annotated lists of fully lexical argumentative elements, 24 in total (8 selections x 3 categories) and the centroid for each list (defined as the simple mean of word vectors for all lexemes for the list, cf. e.g. Brokos et al., 2016, 114). Then we extracted from the model the first 300 of most similar word to the centroid for every list and conducted an annotation of 100 words for each centroid (2400 items in total) using the categories defined above. Thus, we are trying to use the centroid for, e.g., evaluatives in EP_EN corpus to generate a list of potential candidates of the same category.

As could be easily seen from Table 1, we find an improvement (i.e., higher percentage of items of given category with respect to the initial annotation) for 22 of 24 centroids (besides KEY_N_MI_EP_ENfDE and M_ARG_MI_EP_ENfDE), with some major improvement like 6.6 times for M_ARG_MI_EP_EN, (though with a rather low frequency in the initial annotation), or 4.4 times for KEY_N_RAND_COVID. It is important to notice, that centroid-based identification of argumentative lexical items generates an overall higher percentage of argumentative lexicon (last two columns), that is, the centroid of evaluatives in a given corpus, for example, generates more evaluatives, but also more key notions and meta-argumentatives than functional or non-argumentative fully lexical items, etc. Still we can argue that the method we developed generates category-specific improvements of the candidate list, distinguishing between these three classes, too. We calculated the distribution of the elements annotated in centroid-based lists with respect to the centroid types as well as single centroids, using mosaic plots (Friendly, 1994) (generated using `vcd` package in `R`, Meyer et al., 2022) to represent Pearson residuals of the distribution. Figure 4a shows a very strong statistically significant correlation of generated evaluative lexical items: evaluatives, key notions and meta-argumentatives with their respective centroids, as well as clear or even strong statistically significant negative correlation between any of these categories and the centroids based on the other categories, e.g. between the key notion candidates and the centroid of evaluatives or meta-argumentatives (dark/light blue rectangles mark observed values exceeding the expected value by more than the four-/twofold of the square root of the expected value, red values show the underrepresentation of the same magnitude). In Figure 4b we can see a more detailed picture for every single configuration of corpus, category and selection procedure.

Here we observe partially mixed results, in particular the generally lower effectiveness of the procedure in ENfDE corpus. This lower performance of the vector model for the translated corpus could be a corollary of the lower number of informative words among the collocates used in words vectors, cf. the discussion in Section 3, point 3. On the other side we observe relevant improvement in the original EP_EN and COVID corpus. Putting it all together, we obtain an increased precision in the identification of candidates for the three classes of argumentatively relevant lexical items, defined above. Thus, our procedure gives us a heuristic to identify items of the given category with higher precision.

5. Discussion

Our results indicate that distributional characteristics of argumentation resources can be modelled with the proposed discourse-level metrics, such as the average size of knee of ranked collocation lists and vector centroids for word embeddings of argumentation lexicon. Specifically, these methods reduce noise in larger collocation spans, which are needed for capturing distributional properties beyond local context.

The dynamics of knee change across different spans indicates that meta-argumentative lexicon has distinct distributional characteristics in comparison with the general lexicon and other types of argumentation resources, (evaluatives, key notions, etc.): for most corpora their curve is flattening or going down in the middle-size spans. A possible linguistic interpretation of this result could be that meta-argumentatives find several 'islands of consistency' both in the local and more distant context within the discourse. This could indicate that meta-argumentatives form part of 'coordinated constructions' (Fillmore et al., 1988), e.g., on the one hand, they are integrated within the local syntactic structure, on the other hand, they have discourse-level valencies that are filled with more distant coordinated argumentation resources. This hypothesis could be experimentally tested, contributing to current research on discourse-level phenomena in construction grammars (e.g., Enghels and Sansiñena, 2021), extending them with construction coordination models for argumentation.

In the translation context, meta-argumentatives are also much stronger influenced by the target language in comparison to other types of argumentation lexicon, possibly because of greater asymmetry in the syntactic structure of English and German clauses, which could have a particular impact on meta-argumentative verbs.

More generally, the proposed methodology highlights interesting distributional characteristics of the argumentation lexicon used in translated corpora. Translations are often found to be influenced by the linguistic structures and usage patterns of the source language on the lexical, syntactic and textual levels, which can be confirmed with the corpus-based statistical analysis (e.g., Baroni and Bernardini, 2006). Such effects are often referred to as 'translationese', and they have been a serious limiting factor for the use of parallel corpora in contrastive linguistics research, as well as for the development and evaluation of modern Machine Translation (MT) systems that

are trained on parallel corpora (e.g., Zhang and Toral, 2019; Graham et al., 2020; Vanmassenhove et al., 2021). The comparison of knee changes across collocation spans provides a better understanding of how different types of argumentation resources vary in this respect. Practical applications of this line of research could lead to improvements in MT training and evaluation procedures that will minimize the influence of the source language patterns found in parallel corpora.

Finally, we would like to emphasize that this work is a pilot study with corresponding important limitations. The annotation of categories of argumentative lexical items should in the future be realized by multiple annotators with control of inter-annotator agreement. The possible explanations of the distributional properties of our categories should be tested for other corpora and compared systematically with qualitative evaluation of the uses of argumentative lexical items in different types of corpora.

Bibliography

Atayan, Vahram. Argumentationsstrukturen–ein Äquivalenzparameter bei der Übersetzung. *Multiperspektivische Fragestellungen der Translation in der Romania. Hommage an Wolfram Wilss zu seinem*, 80:61–93, 2007.

Babych, Bogdan, Anthony Hartley, Serge Sharoff, and Olga Mudraya. Assisting translators in indirect lexical transfer. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 136–143, 2007.

Baroni, Marco and Silvia Bernardini. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, 2006. doi: 10.1093/llc/fqi039.

Brokos, Georgios-Ioannis, Prodromos Malakasiotis, and Ion Androutsopoulos. Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 114–118, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2915. URL `https://aclanthology.org/W16-2915`.

Enghels, Renata and María Sol Sansiñena. Discourse-level phenomena in construction grammars. *Constructions and Frames*, 13(1):3–20, 2021. doi: 10.1075/cf.00045.int.

Evert, Stefan. Corpora and collocations. In Lüdeling, Anke and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin, 2008. URL `http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_manuscript.pdf`.

Fillmore, Charles J, Paul Kay, and Mary Catherine O'connor. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, pages 501–538, 1988. doi: 10.2307/414531.

Friendly, Michael. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, 1994. doi: 10.1080/01621459.1994.10476460.

Gast, Volker. Comparing Annotation Types and n-Gram Sizes. In Schützler, Ole and Julia Schlüter, editors, *Data and methods in corpus linguistics: Comparative approaches*, chapter 11, pages 323–352. Cambridge University Press, 2022. doi: 10.1017/9781108589314.012.

Graham, Yvette, Barry Haddow, and Philipp Koehn. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, 2020. doi: 10.18653/v1/2020.emnlp-main.6.

Hoey, Michael. Lexical priming and translation. *Kruger, Alet, Kim Wallmach & Jeremy Munday (eds.)*, pages 153–168, 2011.

House, Juliane. Using translation and parallel text corpora to investigate the influence of global English on textual norms in other languages. *Corpus-based translation studies: Research and applications*, pages 187–208, 2011.

Koehn, Philipp. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86, 2005.

Kranich, Svenja and Volker Gast. Explicitness of epistemic modal marking: Recent changes in British and American English. *Thinking Modality: English and Contrastive Studies of Modality. Newcastle upon Tyne: Cambridge Scholars Publishing*, pages 3–22, 2015.

Kruger, Alet, Kim Wallmach, and Jeremy Munday. *Corpus-based translation studies: Research and applications*. Bloomsbury Publishing, 2011.

Meyer, David, Achim Zeileis, Kurt Hornik, Florian Gerber, and Michael Friendly. *vcd: Visualizing Categorical Data*, 2022. URL `https://CRAN.R-project.org/package=vcd`. R package version 1.4-10.

Satopaa, Ville, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011. doi: 10.1109/ICDCSW.2011.20.

Schmid, Helmut. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154, 2013.

van Eemeren, Frans H. *Argumentation theory: A pragma-dialectical perspective*. Springer, 2018. doi: 10.1007/978-3-319-95381-6.

van Eemeren, Frans H, Rob Grootendorst, and Tjark Kruiger. Handbook of argumentation theory. In *Handbook of Argumentation Theory*. De Gruyter Mouton, 2019. doi: 10.1515/9783110846096.

Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, 2021. doi: 10.18653/v1/2021.eacl-main.188.

Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.

Zhang, Mike and Antonio Toral. The Effect of Translationese in Machine Translation Test Sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, 2019. doi: 10.18653/v1/W19-5208.

Address for correspondence:
Vahram Atayan
`atayan@uni-heidelberg.de`
Institute for Translation and Interpreting
Heidelberg University, Plöck 57a, Heidelberg 69117, Germany

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 119 OCTOBER 2022 37–66

Reflexives as Part of Verb Lexemes in the VALLEX Lexicon

Václava Kettnerová, Markéta Lopatková, Anna Vernerová

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

Reflexivity represents one of the core research tasks in current linguistics. As the use of reflexives, encoding a variety of meanings, typically brings about changes in verb valency, the description of reflexivity is highly relevant – among others – also for valency oriented studies. In this paper, we address the reflexive in Czech categorized as a derivational morpheme (e.g., *zlomit*^{pf} 'to break something' → *zlomit se*^{pf} 'to break; to crack'), with the focus on valency behavior of reflexive verbs as represented in the valency lexicon of Czech verbs *VALLEX*.

In the data component of the lexicon, reflexive verbs, i.e., verbs with reflexive lexemes, are captured in separate lexicon entries, represented by respective verb lemma(s) containing the free reflexive morpheme *se* or *si*. In *VALLEX*, there are 922 lexical entries for reflexive verbs described in 1 545 lexical units represented by 1 525 verb lemmas (this number covers almost one quarter of lexical units and one third of verb lemmas in the lexicon). Reflexive verbs can be divided into two groups: into those without any non-reflexive counterpart (*reflexiva tantum*, 208 lexical units represented by 177 verb lemmas) and into those for which a non-reflexive base verb can be identified (*derived reflexive verbs*, 1 337 lexical units represented by 1 348 verb lemmas). Those derived reflexive verbs that are directly related to their non-reflexive base verbs are classified into seven types on the ground of their relation to the non-reflexive counterparts, captured in the data component of the lexicon by the value of the attribute reflexverb.

Further, the relation of derived reflexive verbs and their respective non-reflexive counterparts is described by formal rules comprised in the grammar component of the lexicon (19 rules in total), which provide the information on changes in the mapping of semantic participants onto valency complementations.

 Corresponding author: kettnerova@ufal.mff.cuni.cz
Cite as: Václava Kettnerová, Markéta Lopatková, Anna Vernerová. Reflexives as Part of Verb Lexemes in the VALLEX Lexicon. The Prague Bulletin of Mathematical Linguistics No. 119, 2022, pp. 37–66.
doi: 10.14712/00326585.022.

1. Introduction

The importance of the study of reflexivity is widely acknowledged in contemporary linguistics (see esp. Genuišienė, 1987; Kemmer, 1993; Frajzyngier and Walker, 2000; König and Gast, 2008). Reflexivity – in a broad sense – covers all uses of verbs (sporadically nouns and adjectives) marked by a reflexive. As the reflexive, we mean "an element in the verb (affix, ending, etc.) or its environment (particle, pronoun etc.) which has (or once had) a reflexive meaning (of coreference of two semantic roles) as its only or one of many function" (Genuišienė, 1987, p. 25). As the reflexive is involved in a variety of meanings, which are typically associated with valency changes, their description belongs to primary tasks of valency oriented research. In Czech, two functions of the reflexive are distinguished: (i) the function of the personal pronoun and (ii) the function of a free morpheme that is seen (a) as part of the reflexive verb form or (b) as part of reflexive verb lexemes (see esp. Kopečný, 1954; Komárek et al., 1986; Panevová, 2008). From this follows that while the reflexive of the type (iia) is categorized as an inflectional morpheme, the reflexive of the type (iib) is treated as a derivational morpheme. Here we primarily focus on the valency changes brought about by the reflexive categorized as the derivational morpheme. We describe these changes as captured in the valency lexicon of Czech verbs *VALLEX* (esp. Lopatková et al., 2016).[1]

The paper is structured as follows. The background valency theory of the *VALLEX* lexicon is described in Section 2. The lexicographic representation of the reflexive is outlined in Section 3, with an emphasis on the position of different functions of the reflexive in the whole architecture of the *VALLEX* lexicon. Different types of verbs with reflexive lexemes (henceforth *reflexive verbs*) were identified in the data; their treatment in the lexicon is then introduced. As our main contribution in this paper, we concentrate on a classification of derived reflexive verbs (Section 4). In this classification, we adopt a taxonomic approach similar to the one proposed by Genuišienė (1987) and for Czech outlined by Pergler (2020). We categorize different types of derived reflexive verbs on the basis of changes in their valency structure compared to their non-reflexive base verbs. It should be emphasized at the very beginning that these verbs "constitute a semantic continuum and discrete ... types distinguished are only points along this continuum" (Genuišienė, 1987, p. 59). We thus delimit individual types of derived reflexive verbs as focal points on this continuum.

[1] The changes in the valency structure of verbs associated with the reflexive pronoun, including syntactic reflexivity and reciprocity, and those brought about by the reflexive verb form, comprising deagentive and dispositional diathesis, have been thoroughly described by Kettnerová et al. (2021) and Lopatková et al. (2016), respectively.

2. Basic Concepts of the *VALLEX* Lexicon

2.1. Valency Theory in *VALLEX*

The *VALLEX* lexicon makes use of the valency theory formulated within the Functional Generative Description as its theoretical background (see esp. Sgall et al., 1986; Panevová, 1974–75, 1994; Panevová et al., 2014). Valency is primarily related to the so-called tectogrammatical layer in this approach, roughly corresponding to the deep syntactic layer, with a specific impact on the surface syntactic layer and the morphemic layer as well. A predicate (typically verbs, but also some nouns, adjectives and adverbs in their individual senses), is characterized by a certain set of *valency complementations*. Two types of valency complementations are distinguished: actants and free modifications.[2]

Actants, roughly corresponding to arguments of other syntactic theories, modify only restricted groups of predicates that can be listed and they occur in a single predicate only once. On the surface syntactic layer, they are expressed as the subject or as (direct and indirect) objects. Five actants are distinguished for verbs: Actor (ACT), Addressee (ADDR), Patient (PAT), Origin (ORIG) and Effect (EFF). *Free modifications*, roughly corresponding to adjuncts of other theories, can modify any predicate and they can appear in a single predicate more than once. On the surface, they are realized as adverbials. Unlike actants, they are distinguished primarily on a semantic basis. Both actants and free modifications can be either obligatory or optional on the tectogrammatical layer; this distinction is determined by the so-called dialogue test (Panevová, 1974–75).

Actants (be they obligatory or optional) and obligatory free modifications constitute a *valency frame* of a predicate. The valency frame is a set of valency positions, each standing for one valency complementation, labeled by a functor (i.e., a label representing the relation of the valency complementation to its governing predicate), and by the information on its obligatoriness. For actants, their morphemic forms are provided as well, determining their surface realization; morphemic forms of free modifications follow from their functors.

For semantic characterization of a verb and its valency complementations, the concept of a situation has appeared to be beneficial (see esp. Mel'čuk, 2004). Each verb in a given sense denotes a situation with a certain set of participants; their number, types (characterized by semantic roles)[3] and relations then characterize the verb in a unique way. Any changes in this set typically indicate a change of the situation, hence a different sense of the verb. Each participant of the verb typically corresponds to one of its valency complementation.

[2] Plus quasi-actants, having some properties in common with actants while others with free modifications; quasi-actants thus represent the borderline category.

[3] Semantic roles – despite some criticisms, see a summary in Levin and Rappaport Hovav (2005) – have proved to be a useful tool in the description of various valency phenomena.

2.2. Structure of the *VALLEX* lexicon

The central concept of the *VALLEX* lexicon[4] is a *lexeme*, an abstract two-fold unit associating all verb forms with their *lexical units*, i.e., with their individual senses. Unlike traditional dictionaries, *VALLEX* treats perfective and imperfective aspectual counterparts within a single lexeme since they typically share their valency properties. Each lexeme is captured as a separate lexicon entry, represented by verb lemma(s). In the lexicon entry, individual lexical units are assigned their valency frames, examples and a gloss. Further, each lexical unit can be provided with additional syntactic and semantic information.

In *VALLEX*, the emphasis is put on analyzing the full spectrum of valency-related phenomena, including the syntactic structures that affect the surface expression of valency. For the description of changes in valency structure of verbs brought about by diatheses, syntactic reflexivity and reciprocity, referred to as *grammaticalized alternations*, the lexicon has been divided into two parts: a *data component* and a *grammar component* (see esp. Lopatková et al., 2016). The data component stores valency frames capturing the active, non-reflexive and non-reciprocal uses of verbs, while the grammar component stores formal rules making it possible to derive valency frames underlying passive, reflexive and reciprocal constructions. Besides these formal rules, the grammar component contains rules associating pairs of lexical units of verbs characterized by systemic shifts in their meaning, referred to as *lexicalized alternations*.

3. Reflexives in the Structure of VALLEX

In Czech, the reflexive has the clitic forms *se*, *si* and the full forms *sebe*, *sobě*, *sebou*. Whereas the full forms are undoubtedly classified as the reflexive personal pronoun, the status of the clitic forms is questionable. The clitic forms of the reflexive can be employed either as the reflexive personal pronoun[5] or as a free morpheme representing either part of the reflexive verb form or part of reflexive verb lexemes (see esp. Kopečný, 1954; Komárek et al., 1986; Panevová, 2001; Panevová, 2008; similarly Štícha et al., 2021).

Here we only briefly (and just for completeness) mention the reflexive personal pronoun and the reflexive classified as part of the reflexive verb form and their treat-

[4] https://ufal.mff.cuni.cz/vallex

[5] Let us stress, however, that the pronominal status of the clitic reflexive is not accepted without reservation. For example, Oliva (2000, 2001), Karlík (1999) and Veselý (2018) assign the pronominal function only to the full forms of the reflexive; esp. changes in agreement of predicative complements are taken by these scholars as strong evidence against the pronominal status of the reflexive *se*.
In contrast, some scholars argue that it cannot be disregarded that the clitic and the full forms of the reflexive are functionally equivalent in some constructions, and the choice of their form is determined by topic-focus articulation (see esp. Komárek, 2001; Fried, 2004, 2007; Panevová, 2001). In *VALLEX*, this position, making it possible to treat functionally equivalent uses of the reflexive in the same manner, has been adopted.

ment in *VALLEX*, as these types have been already thoroughly described (see the references in Sections 3.1 and 3.2.1). We primarily focus on the clitic forms of the reflexive *se*, *si* that are classified as part of reflexive verb lexemes and their representation in *VALLEX* (Sections 3.2.2 and 4).

3.1. Reflexive Personal Pronoun

In Czech, the reflexive personal pronoun, filling one valency position of a verb, encodes referential identity of some of valency complementations of the verb in conventionalized constructions expressing either *reflexivity* or *reciprocity*. In *VALLEX*, these constructions are referred to as *syntactic reflexivity* and *reciprocity*, respectively (esp. Kettnerová et al., 2021). See the reflexive construction in example (1-a) and the reciprocal one in example (1-b). In this case, the clitic forms of the reflexive *se*, *si* are substitutable by the full forms *sebe*, *sobě*, respectively, if topic-focus articulation is changed.

(1) a. ***Generál Peckem se považoval*** *za estéta a intelektuála.*[6]
'General Peckem considered himself an esthete and an intellectual.'
b. ***Oba se*** *pak vzájemně* ***považují*** *za lháře ...*
'They then both view each other as a liar ...'

Reflexive and reciprocal constructions are derived by the syntactic operation of reflexivization and reciprocalization, respectively. In the *VALLEX* lexicon, syntactic reciprocity and reflexivity are treated as grammaticalized alternations, the description of which relies on both the data and the grammar component. As it was thoroughly discussed by Kettnerová et al. (2021), we leave it aside here.

3.2. Reflexive Free Morpheme

The clitic forms of the reflexive can also have the function of a free morpheme, which is distinguished into that representing part of the reflexive verb form (only the clitic reflexive *se*, Section 3.2.1) and into that standing for part of verb lexemes (the clitic reflexive *se* and *si*, Section 3.2.2); as such, they are not substitutable by the full forms.

3.2.1. Reflexive Verb Form

The clitic reflexive *se* combines with the 3rd person of indicative or conditional of a verb, constituting together the reflexive verb form. This verb form is characteristic of deagentive and dispositional constructions, see examples (2-a) and (2-b), respectively, traditionally subsumed under diatheses (see esp. Panevová et al., 2014).

[6] Unless explicitly stated differently, examples are extracted from the Czech National Corpus, subcorpus SYNv10 (available at `https://www.korpus.cz/`).

(2) a. *Druhý poločas **se dohrával** takřka z povinnosti, ...*
'The second half-time was played out almost out of duty, ...'
b. *Zpěvákovi **se** chata špatně **prodávala** ...*
'The singer's cottage sold poorly ...'

Similarly to syntactic reflexivity and reciprocity (see Section 3.1), deagentive and dispositional diatheses are represented in *VALLEX* as two types of grammaticalized alternations. The information on these diatheses is then divided between the data and the grammar component. As the representation of different types of diatheses (including formal rules describing changes in valency structure of verbs) has been thoroughly discussed in Lopatková et al. (2016), we leave them aside here.

3.2.2. Reflexive Verb Lexemes

The clitic forms of the reflexive *se* or *si* can be an obligatory or optional part of verb lexemes as well. Henceforth, we refer to verbs with reflexive lexemes for simplicity as *reflexive verbs* here. If the reflexive is an *obligatory part of a verb lexeme*, the reflexive *se* or *si* combines with all forms (including the infinitive one) and with all lexical units of this verb,[7] see, e.g., the two lexical units of the verb *rozplývat se* ^{impf} – *rozplynout se* ^{pf} 'to dissolve; to gush' in examples (3-a) and (3-b) and the lexical unit of the verb *vážit si* ^{impf} 'to appreciate' in example (3-c). If the reflexive is an *optional part of a verb lexeme*, some lexical units of the verb can be marked or unmarked by the reflexive and the reflexive does not bring about any syntactic and semantic change, see example (3-d) with the verb *naříkat*$_1$ *(si)* ^{impf} 'to complain' and example (3-e) with the verb *šplhat (se)* ^{impf} 'to rise'.

(3) a. *Nepříjemné pocity **se rozplývají**. /*
Nepříjemné pocity **rozplývají.* [modified]
'Unpleasant feelings dissolve.'
b. ***Rozplývejte se** nad krásami Prahy a budete mít pokoj. /*
****Rozplývejte** nad krásami Prahy a budete mít pokoj.* [modified]
'Gush about the beauties of Prague and you will have peace.'
c. *Svobody **si váží** ten, kdo zažil nesvobodu. /*
Svobody **váží ten, kdo zažil nesvobodu.* [modified]
'Freedom is appreciated by those who have experienced unfreedom.'
d. *Nebudu **si** však **naříkat** na život.* [modified] /
*Nebudu však **naříkat** na život.*
'But I will not complain about life.'

[7] The only rare exceptions are represented by the passive participle of reflexive verbs, if available, it does not have to be marked by the reflexive, compare, e.g., *Rodiče se o děti dobře postarali.* 'The parents took good care of the children.' and *O děti bylo dobře postaráno.* 'The children were well taken care of.'

Reflexive	Lexical units	Lemmas	Lexemes
se	1 363	1 311	785
si	182	214	137
(*se*)	37	41	25
(*si*)	115	137	74
se or *si*	1 545	1 525	922
(*se*) or (*si*)	152	176	97
all[8]	1 697	1 701	1 019

Table 1. The basic statistics on reflexive verbs in the VALLEX lexicon. The parentheses indicate optionality of the reflexive.

e. *Rtuť teploměru* ***se*** *prudce* ***šplhala*** *vzhůru.* /
Rtuť teploměru prudce ***šplhala*** *vzhůru.* [modified]
'The mercury in the thermometer was rising up sharply.'

As to the representation of reflexive verbs in the *VALLEX* lexicon, those reflexive verbs whose lexemes are obligatorily marked by the reflexive constitute a separate lexicon entry, represented by respective verb lemma(s) containing the reflexive as the headword of the entry, as e.g., *bát se* [impf] 'to be afraid of; to fear', *přichystat se* [pf] 'to make ready; to prepare', *libovat si* [impf] 'to enjoy; to be pleased', *děkovat si* [impf] 'to thank each other'. In contrast, the verbs whose verb lexemes are only optionally marked by the reflexive are comprised in the same lexicon entry as their non-reflexive counterparts and the respective reflexive in their lemma headwords is recorded in parentheses; e.g., the verbs *spoléhat* [impf] – *spolehnout* [pf] and *spoléhat se* [impf] – *spolehnout se* [pf] 'to rely' are subsumed under the same lexical entry headed by the lemmas *spoléhat (se)* [impf] – *spolehnout (se)* [pf].

VALLEX contains 1 697 lexical units of reflexive verbs that are represented by 1 701 lemmas forming 1 019 lexemes. This number covers almost 25% of all the lexical units and more than 36% of all the verb lemmas comprised in the lexicon. Table 1 breaks down these counts by the form of the reflexive. We can see that the verbs with lemmas that are obligatorily marked by the reflexive *se* heavily outweigh those with lemmas obligatorily marked by the reflexive *si*. However, in the case of the verbs in which the reflexive is an optional part of verb lexemes, the situation is reversed: the verbs with the reflexive *si* prevail over those with the reflexive *se*.

[8] Note that the counts in the "Lemmas" and "Lexemes" columns do not sum up to the numbers in the "all" row as in two cases, a single verb is marked by both the optional *se* and the optional *si* (and thus it is counted twice, both in the (*se*) and in the (*si*) row), namely the verb *počínat (se)* [impf] 'to begin', *počít*

Reflexive	Tantum			Derived		
	LUs	Lemmas	Lexemes	LUs	Lemmas	Lexemes
se	188	157	107	1 175	1 154	678
si	20	20	12	162	194	125
all	208	177	119	1 337	1 348	803

Table 2. The basic statistics on reflexiva tantum and derived reflexive verbs in the VALLEX lexicon (LUs stands for "Lexical units").

Reflexive verbs whose lexemes are obligatorily marked by the reflexive split into two main subtypes: into those for which a non-reflexive counterpart can be identified (henceforth called *derived reflexive verbs*) and into those for which no non-reflexive counterpart can be found (*reflexiva tantum*). The former type is more common than the latter: 1 337 lexical units of derived reflexive verbs represented by 1 348 reflexive lemmas in 803 lexemes vs. 208 lexical units of reflexiva tantum represented by 177 reflexive lemmas in 119 lexemes. Table 2 provides the basic statistics on the distribution of the reflexive *se* and *si* in these verbs.

Reflexiva Tantum

Reflexiva tantum, by some authors also called deponents (see esp. Kemmer, 1993; Haspelmath, 2007) or inherently reflexive verbs (Karlík et al., 2016), can completely lack a non-reflexive counterpart (e.g., *narodit se* ^{pf} 'to be born' but **narodit*, *potesknout si* ^{pf} 'to complain' but **postesknout*; the so-called strong deponents in Haspelmath, 2007) or they can have a seemingly non-reflexive counterpart to which they are not, however, semantically and/or syntactically related from a synchronic point of view (e.g., *hádat se* ^{impf} 'to quarrel' but *hádat* ^{impf} 'to guess', *hodit se* ^{pf} 'to match' but *hodit* ^{pf} 'to throw'; the so-called weak deponents, ibid.).

In the *VALLEX* lexicon, reflexiva tantum are represented by their respective verb lemmas containing the reflexive *se* or *si*. Those of them that have seemingly non-reflexive counterparts are distinguished from the respective non-reflexive verbs by Roman numerals, indicating that these reflexive and non-reflexive verbs are homographs, e.g., *hádat*$_{II}$ *se* ^{impf} 'to quarrel' but *hádat*$_{I}$ ^{impf} 'to guess'. Furher, each lexical unit of reflexiva tantum is assigned the attribute reflexverb with the value tantum. For their statistics see Table 2.

(*si*) ^{pf} 'to do' (belonging to the same lexeme as they are semantically related in Czech) and the verb *rozmýšlet* (*si*) ^{impf}, *rozmýšlet* (*se*) ^{impf} 'to consider; to contemplate'.

Derived Reflexive Verbs

Derived reflexive verbs are derivationally related – from a synchronic perspective – to non-reflexive verbs. The word-formation process deriving these reflexive verbs from the base non-reflexive ones is referred to as reflexivization (see esp. Genuišienė, 1987 and for Czech Petr et al., 1986). Some of lexical units of a derived reflexive verb are then typically *directly semantically and/or syntactically related* to a lexical unit of its non-reflexive counterpart (called reversible reflexives by Genuišienė, 1987). Their mutual relation can be described in terms of regular changes in their semantic and/or syntactic features. For example, the derived reflexive verb *třást se* [impf] 'to shiver' in example (4-b) has its base verb in the non-reflexive verb *třást* [impf] 'to shake', see example (4-a).

The lexical units of derived reflexive verbs that are directly semantically and/or syntactically related to their non-reflexive base verbs can, however, undergo subsequent semantic and syntactic shifts, resulting in separate lexical units. The resulting lexical units then have only an *indirect semantic and syntactic relation* to their non-reflexive counterparts (these are referred to as non-reversible reflexives by Genuišienė, 1987). See examples (4-c) and (4-d) with two different lexical units of the derived reflexive verb *třást se* [impf] with the meaning 'to be eager' and 'to worry about; to be afraid of', respectively.

(4) a. *Zima mnou* ***třásla****, až mi začaly cvakat zuby.*
'The cold was shaking me so much that my teeth began to chatter.'
b. ***Třásl jsem se*** *zimou.*
'I was shivering with cold.'
c. *... ale o to více* ***se třeseme*** *na novinky ze světa celebrit.*
'... but we are all the more keen on the news from the world of celebrities.'
d. *Na 40 turistů* ***se třáslo*** *o život v kabině lanovky nedaleko Lago Maggiore ...*
'About 40 tourists were worried about their life in the cabin of the cable car near Lake Maggiore ...'

Derived reflexive verbs are captured in the *VALLEX* lexicon as separate lexemes in their own lexicon entries represented by respective verb lemmas (with the reflexive *se* or *si* being their part). Each lexical unit of a derived reflexive verb is assigned the attribute reflexverb with the value derived, with a suffix indicating whether the lexical unit of the reflexive verb is directly or indirectly related to a particular lexical unit of the base verb. When directly related, the suffix identifies a type of the reflexive verb, as addressed in Section 4, and the attribute provides also a reference to the respective lexical unit of the base verb. Those lexical units that have only the indirect relation to their non-reflexive base verbs are indicated by the suffix nonspecific, without any reference to a lexical unit of the non-reflexive counterpart. Basic statistics on derived reflexive verbs contained in the data of the lexicon show that the verbs that

Reflexive	Directly Related			Indirectly Related		
	LUs	Lemmas	Lexemes	LUs	Lemmas	Lexemes
se	693	902	521	482	500	318
si	52	73	48	110	126	80
all[9]	745	975	569	592	626	398

Table 3. The basic statistics on derived reflexive verbs in the VALLEX lexicon (LUs stands for "Lexical units").

are directly related to their non-reflexive base verbs slightly outweigh those that are indirectly related, see Table 3.

Further, the grammar component of the *VALLEX* lexicon contains a set of formal rules (19 rules in total) specifying the relation of the derived reflexive verbs that are directly related to their non-reflexive base verbs and their respective non-reflexive counterparts in terms of changes in the mapping of semantic participants onto valency complementations, as discussed in Section 4.

4. Types of Derived Reflexive Verbs in *VALLEX*

In the case of the derived reflexive verbs that are directly semantically and/or syntactically related to their non-reflexive base verbs, this relation can be described in terms of systemic differences in semantic and/or syntactic properties of these verb pairs. Four types of changes can be identified:

(i) Systemic changes in a situation denoted by a reflexive verb which result in the reduction of the number of semantic participants of the reflexive verb compared to its non-reflexive base verb; as a result, the number of its valency complementations is reduced as well, compare, e.g., the non-reflexive verb *naklánět* ^{impf} – *naklonit* ^{pf} 'to tilt' in example (5-a) and the derived reflexive verb *naklánět se* ^{impf} – *naklonit se* ^{pf} 'to lean' in example (5-b). In this case, the semantic participant ‚Agent', expressed in the subject in the non-reflexive verb, is deleted in the reflexive verb, and the number of its valency complementations is then reduced by one complementation.

[9] Note that the "Lemmas" and the "Lexemes" counts on the line "all" here do not sum up to the numbers provided in Table 2 for derived reflexive verbs as a single lemma may simultaneously represent a lexical unit that is directly related and another lexical unit that is indirectly related to its non-reflexive counterpart (and thus it is counted twice); the same is valid for lexemes.

(5) a. (*Naďa a Nika*).ACT-Agent *trhali květy dál, stále víc* ***nakláněli*** *loďku*.PAT-Theme.
'(Nada and Nika).ACT-Agent plucked flowers, tilting the boat.PAT-Theme more and more.'

b. *Muž řeku za hodinu překoná, ovšem jeho loďka*.ACT-Theme ***se*** *cestou povážlivě* ***naklání***.
'The man crosses the river in an hour but his boat.ACT-Theme leans alarmingly along the way.'

(ii) Systemic changes in a situation denoted by a reflexive verb are limited to relations between semantic participants of the situation, while their number and type typically remain preserved. These changes result in the change in the surface syntactic expression of one of the involved semantic participants, compare, e.g., the non-reflexive verb *nadávat* ^{impf} 'to swear' in example (6-a) with the reflexive verb *nadávat si* ^{impf} 'to swear at each other' in example (6-b): both semantic participants ‚Agent' and ‚Recipient' characterizing the non-reflexive verb are retained in the reflexive verb but their relation is presented as mutual. The mutual relation between these semantic participants is mirrored in the surface expression of the valency complementation ADDR, which is expressed by the dative case in the non-reflexive verb, see example (6-a), while in the reflexive verb it has the form of the prepositional group s+7,[10] see example (6-b).

(6) a. *Republikáni*.ACT-Agent ***nadávají*** *demokratům*.ADDR-Recipient ... [modified]
'Republicans.ACT-Agent swear at Democrats.ADDR-Recipient ...'

b. *Republikáni*.ACT-Agent ***si nadávají*** *s demokraty*.ADDR-Recipient ...
(lit. Republicans.ACT-Agent swear with Democrats.ADDR-Recipient ...)
'Republicans and Democrats swear at each other ...'

(iii) A situation denoted by a reflexive verb, i.e., a number and a type of semantic participants as well as their relations, remains the same compared to its non-reflexive base verb but the mapping of the semantic participants onto valency complementations is changed (hence the surface expression of the involved semantic participants). As a result, the perspective from which the situation is viewed changes. Compare, e.g., the mapping of the semantic participants ‚Donor' and ‚Recipient' onto valency complementations in the non-reflexive verb *nakažovat/nakazovat*$_{II}$ ^{impf} – *nakazit* ^{pf} 'to infect' in (7-a) and in the reflexive verb *nakažovat se/nakazovat*$_{II}$ *se* ^{impf} – *nakazit se* ^{pf} 'to become infected' in (7-b). As a result, the situation denoted by these verbs is presented either from the perspective of

[10] The Arabic numerals stand for morphological cases: 1 - nominative, 2 - genitive, 3 - dative, 4 - accusative, 6 - locative, and 7 - instrumental; in the case of prepositional groups, the preposition precedes the number indicating the respective case (prepositions are not translated as they can have various interpretations depending on their governing verbs).

the ‚Donor' (the non-reflexive verb, example (7-a)) or from the perspective of ‚Recipient' (the reflexive verb, example (7-b)).

(7) a. *Když Kolumbus doplul do Ameriky, jeho námořníci*.ACT-Donor ***nakazili*** *domorodce*.PAT-Recipient *neštovicemi* ...
'When Columbus reached America, his sailors.ACT-Donor infected the natives.PAT-Recipient with smallpox ...'
b. *Domorodci*.ACT-Recipient ***se nakazili*** *od Kolumbových námořníků*.PAT-Donor *neštovicemi* ... [modified]
'The natives.ACT-Recipient became infected with smallpox from Columbus' sailors.PAT-Donor ...'

(iv) In limited cases, changes between reflexive verbs and their non-reflexive counterparts are restricted to the surface expression of the participant expressed in non-reflexive verbs as the accusative direct object. Compare, e.g., the verb $dohadovat_{I}(si)/dohodovat_{II}(si)^{impf} - dohodnout\ (si)^{pf}$ 'to negotiate; to fix' in example (8-a) with the reflexive verb $dohadovat_{I}se/dohodovat_{II}se^{impf} - dohodnout\ se^{pf}$ 'to negotiate; to fix' in example (8-b). In the non-reflexive verb, the semantic participant ‚Information' mapped onto PAT is expressed in the direct object while in the reflexive verb, it is realized as an indirect object with the form of the prepositional group na+6.

(8) a. *Domácí oddíl chce s rozhodčím* ***dohodnout*** *termín*.PAT-Information ... [modified]
'The home team wants to fix the date.PAT-Information with the referee ...'
b. *Aliance* ***se*** *chce s Moskvou* ***dohodnout*** *na společné chartě*.PAT-Information ...
'The Alliance wants to agree with Moscow on a collective charter.PAT-Information ...'

Based on the four types of changes introduced above, we distinguish seven types of the derived reflexive verbs that are directly related to their non-reflexive base verbs, namely decausative, autocausative, 'partitive object', reciprocal, converse, quasiconverse, and deaccusative reflexive verbs,[11] see Table 4. Basic statistics on individual types of these derived reflexive verbs can be found in Table 5 at the end of Section 4 (page 64.).

In *VALLEX*, the relation between derived reflexive verbs and their non-reflexive base verbs is described by general rules provided in the grammar component. These rules capture the correspondence between the affected semantic participants and valency complementations in a simple form of a table: Column I introduces the correspon-

[11] We refer to these types in line with Genuišienė (1987), as it was further used for Czech by Pergler (2020) and for Polish by Wiemer (2007).

Type of reflexive verbs	Type of changes	Semantic	Syntactic	
			Deep	Surface
decausative	(i)	+	+	+
autocausative	(i)	+	+	+
'partitive object'	(i)	+	+	+
reciprocal	(ii)	+	–[12]	+
converse	(iii)	–	+	+
quasiconverse	(iii)	–	+	+
deaccusative	(iv)	–	–	+

Table 4. Individual types of the derived reflexive verbs that are directly related to their non-reflexive counterparts as identified in the VALLEX lexicon. The + sign in the columns "Semantic" and "Syntactic" indicates changes in semantic participants and in their deep and surface syntactic expression, respectively; the — sign indicates that semantic participants and their deep and surface syntactic expression are preserved.

dence in constructions of non-reflexive base verbs and column II in constructions of derived reflexive verbs (with the variables X and Y standing for valency complementations), see below. The information on changes in surface positions of the involved valency complementations then follows from their morphemic forms provided in respective valency frames. The symbol ∅ indicates that a semantic participant is not mapped onto any valency complementation, and hence it is realized neither in the deep syntactic structure nor in the surface structure. The rule is applied if the condition specified there is met, namely, if the attribute reflexverb, captured in the data component for individual lexical units of derived reflexive verbs, contains a required value, indicating the type and the subtype of a derived reflexive verb (and, if necessary, the affected valency complementations as well).

Type of reflexive verbs Subtype of reflexive verbs		`name of the rule`
condition	reflexverb: derived-type_subtype_X_Y	
	I	II
‚Semantic participant 1' **‚Semantic participant 2'**	X Y	Y ∅

[12] In limited cases, the semantics change in reciprocal reflexive verbs involves the change of the number of their participants; then this change is reflected in their deep and surface structure as well, see footnote 19.

4.1. Decausative Reflexive Verbs

Decausative reflexive verbs (referred to as anticausative, inchoative or spontaneous as well, see e.g., Haspelmath, 1993 and Fried, 2004) are typically derived from transitive causative verbs by the reflexive *se* (e.g., *navazovat se* ^{impf} – *navázat se* ^{pf} 'to bind; to be attached together' ← *navazovat* ^{impf} – *navázat* ^{pf} 'to tie; to fasten sth to sth', *otvírat se/otevírat se* ^{impf} – *otevřít se* ^{pf} 'to open up' ← *otvírat/otevírat* ^{impf} – *otevřít* ^{pf} 'to open', *uskutečňovat se* ^{impf} – *uskutečnit se* ^{pf} 'to come into being' ← *uskutečňovat* ^{impf} – *uskutečnit* ^{pf} 'to make sth happen'). Their non-reflexive base verbs are characterized (in addition to other possible semantic participants) by two participants. The first semantic participant, mapped onto the nominative ACT expressed in the subject position, has the semantic role of ‚Agent' or of ‚Causator'.[13] These two roles alternate in individual uses of a single non-reflexive base verb, compare the two examples of the non-reflexive verb *obnovovat* ^{impf} – *obnovit* ^{pf} 'to restore' in (9-a). The second semantic participant has the role of ‚Theme' or of ‚Patient'.[14,15] It mostly corresponds to PAT realized by the prepositionless accusative in the direct object position, see examples in (9-a) and in (10-a).

(9) a. *obnovovat* ^{impf} – *obnovit* ^{pf} 'to restore; to renew sth'
ACT$_1$ PAT$_4$ BEN$_3$
A Haremheb.ACT-Agent ***obnovil*** *také pořádek*.PAT-Theme *ve Vesetu* ...
'And Haremheb.ACT-Agent also restored order.PAT-Theme in Veset ...'
... *potěšení*.ACT-Causator *z pohybu* ***obnovilo*** *příjemný pocit*.PAT-Theme *v jeho nitru* ...
'... the pleasure.ACT-Causator of movement restored a pleasant feeling.PAT-Theme inside him ...'

b. *obnovovat se* ^{impf} – *obnovit se* ^{pf} 'to be renewed'
ACT$_1$ BEN$_3$ MEANS$_7$
Pořádek.ACT-Theme *ve Vesetu* ***se obnovil.*** [modified]
'Order.ACT-Theme in Veset was renewed.'
Příjemný pocit.ACT-Theme *v jeho nitru* ***se obnovil.*** [modified]
'A pleasant feeling.ACT-Theme inside him was renewed.'

---

[13] ‚Causator', in contrast to ‚Agent', lacks volitional features. Distinguishing these two roles is justified by the fact that decausative reflexive verbs have a systemic relation to quasiconverse reflexive verbs that operate only with ‚Causator', see Section 4.6.

[14] ‚Patient' differs from ‚Theme' in animate features. These two roles split the group of non-reflexive base verbs into two subgroups, syntactic constructions of which exhibit a different type of ambiguity, see the remark below.

[15] For a small group of decausative verbs, the second semantic participant, being of the propositional character, has the role ‚Phenomenon'. These verbs follow the same pattern as those decausative verbs with the ‚Theme' participant. However, in contrast to these decausative verbs, ‚Phenomenon' can be realized by the infinitive or by dependent content clauses as well. In the grammar component, these decausative verbs are described by a separate rule.

(10) a. *utápět* ^{impf} – *utopit* ^{pf} ‘to drown sb/sth’
ACT$_1$ PAT$_4$
... *spodní zpětné proudy*.ACT-Causator *mohou* ... ***utopit*** *neopatrné plavce*.PAT-Patient ...
‘... lower backcurrents.ACT-Causator can drown unwary swimmers.PAT-Patient ...’

b. *utápět se* ^{impf} – *utopit se* ^{pf} ‘to drown; to be drowned’
ACT$_1$ CAUS$_7$
... *neopatrní plavci*.ACT-Patient ***se*** *mohou* ***utopit*** ... [modified]
‘... unwary swimmers.ACT-Patient may drown ...’
... *mladík*.ACT-Patient ***se utopil*** ... [made-up]
‘... the youth.ACT-Patient drowned ...’

In decausative reflexive verbs, the first semantic participant ‚Agent‘ or ‚Causator‘ is deleted. The remaining semantic participant ‚Theme‘ or ‚Patient‘ is then realized as the nominative ACT in the subject position. As a result, the valency structure of decausative reflexive verbs – compared to the valency structure of their non-reflexive counterparts – is reduced by one valency complementation, typically by PAT expressed in non-reflexive verbs by the accusative, see the valency frames and the examples of the decausative reflexive verb *obnovovat se* ^{impf} – *obnovit se* ^{pf} ‘to be renewed’ in (9-b) and of the verb *utápět se* ^{impf} – *utopit se* ^{pf} ‘to drown; to be drowned’ in (10-b).

As a consequence of the changes in semantic participants, decausative reflexive verbs are deprived of the causative feature and the event expressed by these verbs appears to be uncontrolled, spontaneous or accidental (see esp. Genuišienė, 1987; Fehrmann et al., 2014; Haspelmath, 1993 and for Czech Fried, 2004).

Remark on ambiguity. Decausative reflexive verbs are the source of two types of ambiguity, systematically related to the type of the second affected semantic participant. The first type is tied to the participant ‚Theme‘. It can be illustrated with the first sentence in example (9-b): this construction can be interpreted either as a construction of the decausative reflexive verb *obnovovat se* ^{impf} – *obnovit se* ^{pf} in the sense “the order was renewed by itself” or as a deagentive construction of the non-reflexive base verb *obnovovat* ^{impf} – *obnovit* ^{pf} with the generalized ACT, which is not expressed on the surface, in the sense “somebody restored the order” (see Section 3.2.1).

The second type of ambiguity is associated with ‚Patient‘, see, e.g., the second sentence in example (10-b): this sentence can be interpreted either as a construction of the decausative reflexive verb *utápět se* ^{impf} – *utopit se* ^{pf} in the sense “the youth drowned accidentally” or as a syntactically reflexive construction of the non-reflexive base verb *utápět* ^{impf} – *utopit* ^{pf} in the sense “the youth drowned himself (on purpose)”.

Representation of Decausative Reflexive Verbs in *VALLEX*

In *VALLEX*, there are 374 lexical units of derived reflexive verbs classified as decausative reflexive verbs. These lexical units are contained in 300 lexemes, represented

by 549 verb lemmas. In the data component, each lexical unit of a decausative reflexive verb is assigned the value derived-decaus provided in the attribute reflexverb. In addition, this attribute provides the link to a respective lexical unit of the non-reflexive base verb. The value derived-decaus is then further specified by the suffix identifying the second participant affected by the change (i.e., ‚Theme', ‚Patient',[16] and ‚Phenomenon', see footnote 15), and thus this value uniquely determines the rule describing the relation between the decausative reflexive verb and its non-reflexive base verb.

In the grammar component, three formal rules describe differences in semantic and syntactic properties of decausative reflexive verbs. For example, rule R1, `decaus_theme`, captures the relation between decausative reflexive verbs with ‚Theme' (the column II) and their non-reflexive base verbs (the column I): this rule stipulates that ‚Theme' mapped onto PAT in non-reflexive base verbs corresponds to ACT in decausative reflexive verbs and that ‚Agent' or ‚Causator' corresponding to ACT in non-reflexive verbs is not mapped onto any valency complementation in decausative reflexive verbs.

Decausative reflexive verbs Theme		Rule R1 `decaus_theme`
condition	reflexverb: derived-decaus_theme	
	I	II
‚Agent | Causator'	ACT	∅
‚Theme'	PAT	ACT

4.2. Autocausative Reflexive Verbs

Non-reflexive base verbs of autocausative reflexive verbs (e.g., *odlišovat se*^{impf} – *odlišit se*^{pf} 'to become different' ← *odlišovat*^{impf} – *odlišit*^{pf} 'to differentiate', *otáčet se*^{impf} – *otočit se*^{pf} 'to rotate; to be turning' ← *otáčet*^{impf} – *otočit*^{pf} 'to turn', *ulevovat si*^{impf} – *ulevit si*^{pf} 'to be relieved of' ← *ulevovat*^{impf} – *ulevit*^{pf} 'to ease; to relieve sb') are characterized (besides other possible participants) by two semantic participants. The first participant ‚Agent' is mapped onto ACT expressed in the nominative subject. The second participant can have either the role of ‚Patient' or of ‚Recipient', both having animate features. ‚Patient' mostly corresponds to PAT, less often to ADDR, realized predominantly in the direct object expressed by the prepositionless accusative, see an example of the non-reflexive verb *oženit*^{pf} 'to marry sb to sb' in (11-a).[17] If the second seman-

[16] In the case of ‚Patient', the valency complementation onto which this participant is mapped in non-reflexive verbs has to be provided in the suffix as well.

[17] In a limited number of cases, this valency complementation can be expressed as an indirect object by the prepositionless dative, by the prepositionless instrumental or by the prepositional group s+7.

tic participant is represented by ‚Recipient', it corresponds to ADDR realized as an indirect object expressed by the prepositionless dative, see an example of the verb *připomínat*^{impf} – *připomenout*^{pf} 'to remind sb of sth' in (12-a).

The morphemic form of the second affected valency complementation (either PAT or ADDR) determines the choice of the reflexive deriving autocausative reflexive verbs. If it has the form of the prepositionless accusative (sporadically, of the instrumental or of the prepositional group s+7, see footnote 17) the reflexive *se* is selected, as illustrated with the verb *oženit se*^{pf} 'to get married', see example (11-b). The prepositionless dative underlies the choice of the reflexive *si* as the verb *připomínat si*^{impf} – *připomenout si*^{pf} 'to remember' shows, see example (12-b).

In autocausative reflexive verbs, either ‚Agent' and ‚Patient' or ‚Agent' and ‚Recipient' are conflated into a single semantic participant, involving features of both of them: the participant does an activity as ‚Agent' and at the same time it is either affected by this activity as ‚Patient', see example (11-b), or this activity is directed to him as to ‚Recipient', see example (12-b). As a result, the number of semantic participants of autocausative reflexive verbs, and hence the number of their valency complementations, is reduced by one participant and by one valency complementations, respectively, compared to their non-reflexive base verbs. The affected participant with the features of both ‚Agent' and ‚Patient', or of both ‚Agent' and ‚Recipient' is then mapped onto ACT expressed in the nominative subject, see examples (11-b) and (12-b), respectively.

Autocausative reflexive verbs are not syntactically distinguished from decausative reflexive verbs (Section 4.1). However, they differ in the characteristics of participants mapped onto their ACT: while ACT of decausative reflexive verbs corresponds to ‚Theme' or to ‚Patient', ACT of autocausative reflexive verbs corresponds to the participant combining the features of both ‚Agent' and ‚Patient' or ‚Recipient', preserving thus agentivity in these reflexive verbs. The fact is evidenced, e.g., by their compatibility with adverbials expressing intentionality (e.g., *úmyslně, záměrně, schválně* 'on purpose').

(11) a. *oženit*^{pf} 'to marry sb to sb'
ACT$_1$ ADDR$_{s+7}$ PAT$_4$
Boženka.ACT-Agent *před léty Arnošta*.PAT-Patient ***oženila*** *s nepraktickou bohatou Helgou* ...
'Years ago, Boženka.ACT-Agent married Arnošt.PAT-Patient to impractical rich Helga ...'

b. *oženit se*^{pf} 'to get married'
ACT$_1$ PAT$_{s+7}$
Před léty se Arnošt.ACT-Agent+Patient ***oženil*** *s nepraktickou bohatou Helgou* ... [modified]
'Years ago, Arnošt.ACT-Agent+Patient got married to impractical rich Helga ...'

(12) a. *připomínat*^{impf} – *připomenout*^{pf} 'to remind sb of sth'
ACT$_1$ ADDR$_3$ PAT$_{4,dcc}$

Trhovci.ACT-Agent *návštěvníkům*.ADDR-Recipient ***připomenou*** *stará, mnohdy už zapomenutá řemesla.*
'Traders.ACT-Agent will remind visitors.ADDR-Recipient of old, often forgotten crafts.'

b. *připomínat si* ^{impf} – *připomenout si* ^{pf} 'to remember'
ACT$_1$ PAT$_{4,dcc}$
Návštěvníci.ACT-Agent+Recipient ***si připomenou*** *stará, mnohdy už zapomenutá řemesla.*
[modified]
'Visitors.ACT-Agent+Recipient will remember old, often forgotten crafts.'

Representation of Autocausative Reflexive Verbs in *VALLEX*

In the data component of *VALLEX*, autocausative reflexive verbs are indicated by the value derived-autocaus provided in the attribute reflexverb (249 lexical units contained in 208 lexemes, which are represented by 372 verb lemmas). Autocausative reflexive verbs are further subclassified with respect to the second participant affected by the change, uniquely identifying the respective rule.[18]

In the grammar component, two rules describe the relation between autocausative reflexive verbs and their non-reflexive base verbs. For example, rule R2, `autocaus_recipient`, captures the changes in the mapping of semantic participants ‚Agent' and ‚Recipient' onto valency complementations in autocausative reflexive verbs (column II) and in their non-reflexive counterparts (column I). It determines that ‚Agent' is mapped onto ACT and ‚Recipient' onto ADDR in non-reflexive verbs while in autocausative reflexive verbs, these two participants are conflated into the single one (the symbol +) that corresponds to ACT, as examples (12-a) and (12-b) above illustrate.

Autocausative reflexive verbs Recipient		Rule R2 `autocaus_recipient`
condition	reflexverb: derived-autocaus_recipient	
	I	II
‚Agent'	ACT	∅
‚Recipient'	ADDR	∅
‚Agent + Recipient'	∅	ACT

4.3. 'Partitive Object' Reflexive Verbs

'Partitive object' reflexive verbs are derived from non-reflexive verbs by the reflexive *se* (e.g., *odvracet se* ^{impf} – *odvrátit se* ^{pf} 'to turn away' ← *odvracet* ^{impf} – *odvrátit* ^{pf} 'to turn sth away', *ovládat se* ^{impf} – *ovládnout se* ^{pf} 'to control oneself' ← *ovládat* ^{impf} – *ovládnout* ^{pf} 'to

[18] In the case of ‚Patient', the suffix of the value further specifies whether the changes involve ADDR or PAT.

control sth', *zaměřovat se impf – zaměřit se pf* 'to focus on sth' ← *zaměřovat impf – zaměřit pf* 'to focus on st'). The non-reflexive verbs from which 'partitive object' reflexive verbs derive are characterized (in addition to other possible semantic participants) by the participants ‚Agent' and ‚Theme'. The latter participant represents inalienable possession of ‚Agent', coming from several semantically restricted domains: it can be ‚Agent's body parts, characteristic features, emotions, ideas etc. ‚Agent' is mapped onto the nominative ACT expressed in the subject position and ‚Theme' typically corresponds to PAT, which is predominantly realized by the prepositionless accusative as the direct object, see an example of the verb *soutřeďovat impf – soustředit pf* 'to focus sth on sth' in (13-a). In limited cases, PAT is realized as an indirect object by the prepositionless instrumental, see an example of the verb *kroutit impf* 'to twist sth' in (14-a).

In 'partitive object' reflexive verbs, the mapping of the semantic participant ‚Agent' is preserved: it still corresponds to the nominative ACT expressed as the subject. The participant ‚Theme' does not, however, correspond to any valency complementation, despite being implied by these reflexive verbs. As a consequence, the corresponding PAT complementation is deleted from valency frames of 'partitive object' reflexive verbs. These frames – compared to valency frames of their non-reflexive base verbs – are thus reduced by one valency position, see examples of the 'partitive object' verb *soustřeďovat se impf – soustředit se pf* 'to focus on sth' in (13-b) and of the verb *kroutit se impf* 'to twist' in (14-b).

(13) a. *soustřeďovat impf – soustředit pf* 'to focus sth on sth'
ACT$_1$ PAT$_4$ EFF$_{k+3,na+4}$
Společnost Heineken.ACT-Agent ***soustředí*** *svou pozornost*.PAT-Theme *na ochranu vodních zdrojů ...*
'Heineken.ACT-Agent focuses its attention.PAT-Theme on the protection of water resources ...'

b. *soustřeďovat se impf – soustředit se pf* 'to focus on sth'
ACT$_1$ PAT$_{k+3,na+4}$
Společnost Heineken.ACT-Agent ***se soustředí*** *na ochranu vodních zdrojů ...* [modified]
'Heineken.ACT-Agent focuses on the protection of water resources ...'

(14) a. *kroutit impf* 'to twist sth'
ACT$_1$ PAT$_{7,s+7}$ DIR
Tanečnice.ACT-Agent ***kroutí*** *pánví*.PAT-Theme, *jako by se zrovna ocitla v Riu na karnevalu.*
'The dancer.ACT-Agent twists her pelvis.PAT-Theme like she is at the carnival in Rio.'

b. *kroutit se impf* 'to twist'
ACT$_1$ CAUS$_7$
Tanečnice.ACT-Agent ***se kroutí****, jako by se zrovna ocitla v Riu na karnevalu.* [modified]
'The dancer twists like she is at the carnival in Rio.'

Representation of ‘Partitive Object’ Reflexive Verbs in *VALLEX*

In the data component of *VALLEX*, 54 lexical units of reflexive verbs in 51 lexemes, which are represented by 94 verb lemmas, are assigned the value derived-partobject in the attribute reflexverb. To the value, the respective valency complementation onto which ‚Theme‘ is mapped in non-reflexive base verbs is suffixed.

In the grammar component, a single rule R3, `partobject`, describes the relation between ‘partitive object’ reflexive verbs and their respective base verbs. This rule states that ‚Theme‘ mapped onto the valency complementation of a non-reflexive base verb, represented in the rule by the variable Y, does not correspond to any complementation in the respective ‘partitive object’ verb. Further, it stipulates that the mapping of ‚Agent‘ onto ACT remains the same in both partitive object verbs and their non-reflexive base verbs, compare the pair of examples (13-a) and (13-b) and of examples (14-a) and (14-b).

‘Partitive object’ reflexive verbs		Rule R3 `partobject`
condition	reflexverb: derived-partobject_Y	
	I	II
‚Agent‘ **‚Theme‘**	ACT Y	ACT ∅

4.4. Reciprocal Reflexive Verbs

Reciprocal reflexive verbs belong to the so-called inherently reciprocal verbs, i.e., to the verbs that express mutuality between some of their participants in their lexical meaning. Their non-reflexive counterparts are characterized (besides other semantic participants) by two participants. The first semantic participant, ‚Agent‘, is mapped onto the nominative ACT expressed in the subject position. The latter has either the role of ‚Patient‘ or of ‚Recipient‘. ‚Patient‘ is mapped onto PAT, mostly expressed by the prepositionless accusative in the direct object position, as exemplified by the verb *nenávidět* ^{impf} ‘to hate’ in (15-a); in fewer cases, PAT is expressed by the prepositionless dative as an indirect object. ‚Recipient‘, as the latter possible participant, is mapped onto ADDR (sporadically onto PAT), typically expressed by the dative in an indirect object, as, e.g., the verb *půjčovat* ^{impf} – *půjčit* ^{pf} ‘to lend sth to sb’ in (16-a) illustrates.

The morphemic form of the second affected valency complementation (either PAT or ADDR) determines the form of the reflexive as a derivational means in reciprocal reflexive verbs. If it has the form of the accusative, the reflexive *se* is applied in the derivation, as exemplified by the reciprocal reflexive verb *nenávidět se* ^{impf} ‘to hate each

other' in example (15-b). In contrast, the dative conditions the choice of the reflexive *si*, see the verb *půjčovat si* ^{impf} – *půjčit si* ^{pf} 'to lend sth to each other' in example (16-b).

In reciprocal reflexive verbs, the two semantic participants, ‚Agent' and ‚Patient' or ‚Agent' and ‚Recipient', are retained as well as their mapping onto valency complementations.[19] However, the relation between them changes. In contrast to non-reflexive base verbs, these two participants are involved in a mutual relation. As a result, an event denoted by a reciprocal reflexive verb is conceived as a mutual action of the affected participants. The mutuality of the two participants is formally manifested by the change of the form of the second valency complementation, onto which ‚Patient' or ‚Recipient' is mapped: in reciprocal reflexive verbs, this valency complementation has uniformly the form of the prepositional group s+7 and it is thus realized on the surface as an indirect object, see the verb *nenávidět se* ^{impf} 'to hate each other' in example (15-b) and the verb *půjčovat si* ^{impf} – *půjčit si* ^{pf} 'to lend sth to each other' in example (16-b).

(15) a. *nenávidět* ^{impf} 'to hate'
ACT$_1$ PAT$_{4,inf,dcc}$
Manžel.ACT-Agent ***nenávidí*** *všechny moje kamarádky*.PAT-Patient.
'My husband.ACT-Agent hates all my friends.PAT-Patient.'

b. *nenávidět se* ^{impf} 'to hate each other'
ACT$_1$ PAT$_{s+7}$
Manžel.ACT-Agent ***se nenávidí*** *se všemi mými kamarádkami*.PAT-Patient. [modified]
(lit. My husband.ACT-Agent hates with all my friends.PAT-Patient.)
'My husband and all my friends hate each other.'

(16) a. *půjčovat* ^{impf} – *půjčit* ^{pf} 'to lend sth to sb'
ACT$_1$ ADDR$_3$ PAT$_4$ AIM$_{k+3,na+4}$ BEN$_{pro+4}$
Kamarádky.ACT-Agent *jí*.ADDR-Recipient ***půjčují*** *masky.* [modified]
'Her friends.ACT-Agent lend masks to her.ADDR-Recipient.'

b. *půjčovat si* ^{impf} – *půjčit si* ^{pf} 'to lend sth to each other'
ACT$_1$ ADDR$_{s+7}$ PAT$_4$
(*Ona*).ACT-Agent ***si půjčuje*** *masky s kamarádkami*.ADDR-Recipient.
(lit. She.ACT-Agent lends masks with her friends.ADDR-Recipient.)
'She and her friends lend masks to each other.'

[19] In limited cases, the valency structure of reciprocal reflexive verbs is changed more significantly and one valency complementation is either added (e.g., *bít se* ^{impf} 'to fight with sb') or deleted (e.g., *házet si* ^{impf} 'to throw a ball with each other'), and the mapping of semantic participants onto valency complementations then typically changes. In these cases, reciprocal reflexive verbs – compared to their non-reflexive base verbs – are typically subject to semantic shifts and they thus border on the derived reflexive verbs indirectly semantically and syntactically related to their non-reflexive counterparts (see Section 3.2.2).

Representation of Reciprocal Reflexive Verbs in *VALLEX*

In the data component of *VALLEX*, reciprocal reflexive verbs are assigned the value derived-recipr in the attribute reflexverb (85 lexical units in 84 lexemes, represented by 118 verb lemmas). This value is supplemented with the suffix identifying the respective rule that describes the relation between a reciprocal reflexive verb and its non-reflexive base verb. The suffix consists of the second involved participant (either ‚Patient' or ‚Recipient') followed by the valency complementation onto which the respective participant is mapped.[20]

Two rules handle the relation between reciprocal reflexive verbs and their non-reflexive counterparts, one for each of the pairs ‚Agent'-‚Patient' and ‚Agent'-‚Recipient'.[21] For example, rule R4, `recipr_recipient`, applies to reciprocal reflexive verbs with ‚Recipient', as illustrated here with the verb *půjčovat si* ^{impf} – *půjčit si* ^{pf} 'to lend sth to each other' in (16-b). The variable Y in the rule stands for the same valency complementation onto which ‚Recipient' is mapped in both reciprocal reflexive verbs and their non-reflexive base verbs (mostly for ADDR, sporadically for PAT). The rule states that both ‚Agent' and ‚Recipient' do not change their mapping onto valency complementations in reciprocal reflexive verbs (column II) compared to their non-reflexive base verbs (column I), as illustrated by the examples in (16-a) and in (16-b). The change in the morphemic form of ADDR or PAT, indicating the change in its surface expression, is captured in valency frames of reciprocal reflexive verbs provided in the data component.

Reciprocal reflexive verbs Recipient		Rule R4 `recipr_recipient`
condition	reflexverb: derived-recipr_recipient_Y	
	I	II
‚Agent' **‚Recipient'**	ACT Y	ACT Y

[20] Typically, there is only one valency complementation to which the respective participant corresponds in both non-reflexive base verbs and derived reflexive verbs. When more complex changes take place, see footnote 19, two complementations come into play; then the first complementation in the suffix comes from the valency frame of a non-reflexive base verb and the latter from the frame of the respective reciprocal reflexive verb.

[21] Two additional rules are necessary to handle more complex changes of the valency structure in the reciprocal verbs mentioned in footnote 19.

4.5. Converse Reflexive Verbs

Converse reflexive verbs and their non-reflexive counterparts denote the same situation, characterized by the same set of semantic participants. Two of these participants, however, change their mapping onto valency complementations; as a consequence, they are expressed in different surface positions. More specifically, in converse reflexive verbs and in their non-reflexive base verbs, the prominent subject position is occupied each time by a different semantic participant from the affected pair and the situation expressed by these verbs is thus presented from the perspective of the relevant participant. For example, the non-reflexive verb *naučit*^{pf} 'to teach', presents the situation denoted by the verb from the perspective of ‚Speaker', see example (17-a). In contrast, the converse reflexive verb *naučit se*^{pf} 'to learn' adopts the perspective of ‚Recipient', see example (17-b).

Converse reflexive verbs split into several semantic subtypes; in *VALLEX*, they are subclassified according to the semantic participants affected by the changes in the mapping onto valency complementations: ‚Speaker'-‚Recipient', see examples (17-a) and (17-b), ‚Donor'-‚Recipient', see examples (7-a) and (7-b) in Section 4, ‚Experiencer'-‚Stimulus', ‚Locatum'-‚Location', see examples (18-a) and (18-b), ‚Bearer of action'-‚Location'.[22]

(17) a. *naučit*^{pf} 'to teach'
ACT$_1$ ADDR$_4$ PAT$_{3,4,inf,dcc}$ MANN
Svému umění uzdravovat Asklepia.ADDR-Recipient ***naučil*** *Cheirón*.ACT-Speaker. [modified]
'Chiron.ACT-Speaker taught Asclepius.ADDR-Recipient his art of healing.'

b. *naučit se*^{pf} 'to learn'
ACT$_1$ PAT$_{3,4,inf,dcc}$ ORIG$_{od+2,z+2}$ MANN
Svému umění uzdravovat ***se*** *Asklepios*.ACT-Recipient ***naučil*** *od Cheirona*.ORIG-Speaker.
'Asclepius.ACT-Recipient learned his art of healing from Chiron.ORIG-Speaker.'

(18) a. *plnit*^{impf} 'to fill sth'
ACT$_1$ PAT$_4$ EFF$_7$
... děti rády pozorují, jak nádoby.PAT-Location ***plní*** *dešťová voda*.ACT-Locatum. [modified]
'... children like watching rainwater.ACT-Locatum filling the containers.PAT-Location'

b. *plnit se*^{impf} 'to fill with sth' ACT$_1$ PAT$_7$
... děti rády pozorují, jak ***se*** *nádoby*.ACT-Location ***plní*** *dešťovou vodou*.PAT-Locatum.
'... children like watching the containers.ACT-Location filling with rainwater.PAT-Locatum'

[22] In *VALLEX*, there are other 7 lexical units of verbs that are so semantically heterogeneous that they are difficult to be classified. These cases draw attention to the fact that the converse function of the reflexive is not hypothetically limited to the semantic types listed above but it may have a broader scope.

In *VALLEX*, there are 67 lexical units of reflexive verbs in 64 lexemes, represented by 103 verb lemmas, that are assigned the value derived-conv in the attribute reflexverb. This value is suffixed with the semantic participants that are subject to the changes in their mapping onto valency complementations; where necessary, they are supplemented with the affected valency complementations (or the variables representing them).

Six rules describe the relations between converse reflexive verbs and their non-reflexive base verbs: one rule for each semantic type listed above (the only exception being the converse reflexive verbs of the ‚Locatum' and ‚Location' type, which are described by two rules). In addition, one general rule captures the relation between verbs which are difficult to be semantically classified.[23] For example, rule R5, `conv_speaker_recipient`, applies to the converse reflexive verbs in which the changes in the mapping onto valency complementations involve the semantic participants ‚Speaker' and ‚Recipient', as exemplified by the verb *naučit se*^{pf} 'to learn'. The rule stipulates that in the non-reflexive verb *naučit*^{pf} 'to teach' (column I), ‚Speaker' corresponds to ACT and ‚Recipient' to the valency complementation ADDR. In the reflexive verb *naučit se*^{pf} 'to learn' (column II), it is ‚Recipient' that is mapped onto ACT while ‚Speaker' corresponds to ORIG, compare examples (17-a) and (17-b).

Converse reflexive verbs		Rule R5
Speaker-Recipient		`conv_speaker_recipient`
condition	reflexverb: derived-conv_speaker_recipient	
	I	II
‚Speaker'	ACT	ORIG
‚Recipient'	ADDR	ACT

4.6. Quasiconverse Reflexive Verbs

Similarly to converse reflexive verbs (Section 4.5), quasiconverse reflexive verbs express the same situation as their non-reflexive base verbs, consisting of the same set of semantic participants, which are, however, mapped each time onto different valency complementations. In contrast to converse reflexive verbs, one of the affected valency complementations in quasiconverse reflexive verbs is represented either by an optional free modification or by a quasi-actant. Depending on the presence or

[23] See footnote 22. Moreover, there are two lexical units of verbs in which the changes in the mapping onto valency complementations involve the participants ‚Substance' and ‚Source'. The description of these verbs would require a separate rule. However, with respect to their sparseness, we leave them aside.

the absence of this complementation in the surface structure, the sentence has either a converse interpretation (if present), or a decausative interpretation (if absent) (see Section 4.1).

In quasiconverse reflexive verbs and their non-reflexive base verbs, the changes in the mapping affect the pair of the semantic participants ‚Causator'-‚Theme' or the pair ‚Causator'-‚Patient'. In non-reflexive verbs, ‚Causator' corresponds to the nominative ACT and ‚Theme' or ‚Patient' to PAT, mostly expressed by the prepositionless accusative in the direct object position,[24] see the non-reflexive verb *lámat impf* 'to break sth' in example (19-a). In contrast, in quasiconverse reflexive verbs, it is the semantic participant ‚Theme' or ‚Patient' that is mapped onto the nominative ACT expressed in the subject while ‚Causator' corresponds to an optional free modification or to a quasi-actant. For instance, in the first example of the verb *lámat se impf* 'to break; to crack' in (19-b), ‚Causator' is mapped onto the optional free modification CAUS or LOC, and in the second example, onto the quasi-actant OBST. If such a complementation is not expressed on the surface, and ‚Causator' is thus not present, the event is interpreted as spontaneous (then the verb is interpreted as a decausative reflexive verb, see Section 4.1).

(19) a. *lámat impf* 'to break sth'
ACT$_1$ PAT$_4$ EFF$_{na+4}$ OBST$_{o+4}$ BEN$_3$
... *vítr*.ACT-Causator ***láme*** *strom*.PAT-Theme *s korunou zlatých listů* ... [modified]
'... the wind.ACT-Causator breaks the tree.PAT-Theme with a crown of golden leaves ...'
... *vlnolam*.ACT-Causator ***láme*** *vlny*.PAT-Theme ... [modified]
'... the breakwater.ACT-Causator breaks waves.PAT-Theme ...'

b. *lámat se impf* 'to break; to crack'
ACT$_1$ PAT$_{na+4}$ OBST$_{o+4}$ CAUS$_7$ LOC
... *strom*.ACT-Theme *s korunou zlatých listů* ***se láme*** *větrem*.CAUS-Causator / *ve větru*.LOC-Causator ... [modified]
'... the tree.ACT-Theme with a crown of golden leaves breaks in the wind.LOC-Causator ...'
... *vlny*.ACT-Theme ***se lámou*** *o vlnolam*.OBST-Causator ... [modified]
'... the waves.ACT-Theme break on the breakwater.OBST-Causator ...'

Representation of Quasiconverse Reflexive Verbs in *VALLEX*

In *VALLEX*, 225 lexical units of reflexive verbs in 195 lexemes, represented by 369 verb lemmas, are indicated as quasiconverse reflexive verbs by the value derived-quasiconv, provided in the attribute reflexverb. The suffix of the value consists of those optional free modifications and/or the quasi-actant OBST onto which ‚Causator' can

[24] In limited cases, it is realized by the prepositionless instrumental or by the prepositional group s+7 in an indirect object position.

be mapped in quasiconverse reflexive verbs (separated with the symbol | if ‚Causator' can correspond to more than one valency complementation in a single verb).

The relation of quasiconverse reflexive verbs to their non-reflexive counterparts describes a single rule R6, `quasiconv`, where ‚Theme' and ‚Patient' are subsumed under the general role ‚Object'. The rule stipulates that ‚Causator' mapped onto ACT in non-reflexive verbs (column I) corresponds to the complementation represented by the variable Y in quasiconverse reflexive verbs (column II); this variable can stand for the optional complementations CAUS, MEANS, LOC or DIR3, or for the quasi-actant OBST. Further, the rule states that ‚Theme' or ‚Patient' corresponding to PAT in non-reflexive verbs changes its mapping onto ACT in reflexive verbs.

Quasiconverse reflexive verbs		Rule R6 `quasiconv`
condition	reflexverb: derived-quasiconv_Y	
	I	II
‚Causator'	ACT	Y
‚Object'	PAT	ACT

4.7. Deaccusative Reflexive Verbs

The change between deaccusative reflexive verbs and their non-reflexive counterparts is limited to the surface expression of the participant expressed in the accusative direct object in non-reflexive verbs. For this semantic participant, the semantic roles of ‚Information', ‚Theme', ‚Phenomenon', and ‚Recipient' are attested in *VALLEX*. It corresponds mostly to PAT, rarely to ADDR (in the case of ‚Recipient'). In non-reflexive verbs, this valency complementation is expressed on the surface as the direct object while in reflexive verbs, it is demoted, being realized as an indirect object.

For example, in the non-reflexive verb *svěřovat* ^{impf} – *svěřit* ^{pf} 'to confide', the participant ‚Information' mapped onto PAT is realized as the direct object while in the reflexive verb *svěřovat se* ^{impf} – *svěřit se* ^{pf} 'to confide', it is expressed as an indirect object, as the change of the morphemic form of PAT from the prepositionless accusative into the prepositional group s+7 indicates, compare examples (20-a) and (20-b).

(20) a. *svěřovat* ^{impf} – *svěřit* ^{pf} 'to confide'
ACT$_1$ ADDR$_3$ PAT$_{4,dcc}$
Před šesti lety mladá žena ***svěřila*** *svému muži tajemství*$_{PAT\text{-}Information}$ *z dětství.*
[modified]
'Six years ago, a young woman confided a secret from her childhood to her husband.'

b. *svěřovat se* ^{impf} – *svěřit se* ^{pf} ‘to confide’
ACT$_1$ ADDR$_3$ PAT$_{s+7,dcc}$
Před šesti lety ***se*** *mladá žena* ***svěřila*** *svému muži s tajemstvím*$_{\text{PAT-Infomation}}$ *z dětství.*
‘Six years ago, a young woman confided a secret$_{\text{PAT-Information}}$ from her childhood to her husband.’

Representation of Deaccusative Reflexive Verbs in *VALLEX*

In *VALLEX*, 38 lexical units of reflexive verbs, contained in 37 lexemes, represented by 70 lemmas, have the value derived-deaccus in the attribute reflexverb. The suffix of the value indicates PAT or ADDR onto which the affected participant is mapped.

The relation of deaccusative reflexive verbs and their non-reflexive counterparts is described by a single rule R7, `derived-deaccus`, where the role ‚Object‘ subsumes ‚Information‘, ‚Theme‘, ‚Phenomenon‘, and ‚Recipient‘. The rule stipulates that this participant preserves the same correspondence in deaccusative reflexive verbs as in their non-reflexive base verbs (represented in the rule by the variable Y, standing mostly for PAT, rarely for ADDR). Morphemic forms of this valency complementation, provided in valency frames, indicate changes in its surface expression.

Deaccusative reflexive verbs		Rule R7 `deaccus`
condition	reflexverb: derived-deaccus_Y	
	I	II
‚Object‘	Y	Y

5. Conclusion

Reflexive verbs, i.e., verbs with the clitic reflexive *se* or *si* that is classified as a derivational morpheme, being an obligatory or an optional part of lexemes of these verbs, represent a substantial portion of data in a lexicon. In the *VALLEX* lexicon, reflexive verbs cover one quarter of all lexical units of verbs (1 697 lexical units out of 6 859) and more than one third verb lemmas (1 701 lemmas out of 4 664). Here we have introduced their representation in this lexicon. As the main contribution of this paper,

[25] Let us stress that the sums for individual types of reflexive verbs in the row “all” do not add up to the total numbers for directly related derived reflexive verbs (as indicated in Table 3) due to the fact that a single lexical unit of a reflexive verb can enter into different relations with a lexical unit of its non-reflexive base verb (as exemplified, e.g., in Section 4.6 by lexical units with either a converse interpretation, or a decausative interpretation); in such cases, it is counted more times. As this issue deserves further analysis going out of scope of this paper, we leave it aside here.

Type of reflexive verbs	Data			Grammar
	Lexical Units	Lemmas	Lexemes	Rules
decausative	374	549	300	3
autocausative	249	372	208	2
'partitive object'	54	94	51	1
reciprocal	85	118	84	4
converse	67	103	64	7
quasiconverse	225	369	195	1
deaccusative	38	70	37	1
all[25]	1 092	1 675	939	19

Table 5. The basic statistics on different types of the derived reflexive verbs directly related to their non-reflexive base verbs.

derived reflexive verbs have been classified on the basis of changes in their semantic properties and in their valency behavior compared to their respective non-reflexive base verbs. Seven types of derived reflexive verbs are distinguished, which have not been discussed so far in detail in Czech linguistics: decausative, autocausative, 'partitive object', reciprocal, converse, quasiconverse, and deaccusative reflexive verbs.

We have described how individual types of reflexive verbs are represented in the *VALLEX* lexicon, making use of the division of the lexicon into the data component and the grammar component. In the data component, each reflexive verb is assigned the attribute reflexverb, the value of which distinguishes whether the reflexive verb belongs to the reflexiva tantum (the value tantum) or to the derived reflexive verbs (the value derived). Within derived reflexive verbs, a line is drawn between those verbs that have a direct semantic and/or syntactic relation to their non-reflexive base verbs and those with just an indirect relation to the base verbs. The derived reflexive verbs with a direct relation to their non-reflexive base verbs are then categorized into the 7 types, indicated in the suffix of the value derived. In each lexical unit of a derived reflexive verb, the suffix further specifies all the information necessary for the identification of the respective formal rule describing the relation of the reflexive verb to its non-reflexive base verb. In the grammar component, 19 rules are stored, providing the information on changes in the mapping of semantic participants of derived reflexive verbs onto valency complementations.

Acknowledgements

Markéta Lopatková's contribution to the work described herein has been supported by the grant *Language Understanding: From Syntax to Discourse* (*LUSyD*) of the

Czech Science Foundation (GAČR), No. GX20-16819X. The work on the *VALLEX* lexicon has been also supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 *LINDAT/CLARIAH-CZ*.

The work on the *VALLEX* lexicon has been using data and tools provided by the *LINDAT/CLARIAH-CZ Research Infrastructure* (`https://lindat.cz`), supported by the Ministry of Education, Youth and Sports of the Czech Republic (project No. LM2018101).

Bibliography

Fehrmann, Dorothee, Uwe Junghanns, and Denisa Lenertová. Slavic Reflexive Decausative. *Russian Linguistics*, 38(3):287–313, 2014. doi: 10.1007/s11185-014-9133-2.

Frajzyngier, Zygmunt and Traci Walker, editors. *Reflexives: Forms and functions*. John Benjamins, Amsterdam – Philadelphia, 2000.

Fried, Miriam. Czech Reflexivization and the Invariance Principle Revisited. *The Slavic and East European Journal*, 48(4):627–653, 2004.

Fried, Miriam. Constructing grammatical meaning: Isomorphism and polysemy. *Studies in Language*, 31(4):721–764, 2007.

Genuišienė, Ema. *The Typology of Reflexives*. Mouton de Gruyter, Berlin – New York – Amsterdam, 1987.

Haspelmath, Martin. More on the typology of inchoative/causative verb alternations. In Comrie, Bernard and Maria Polinsky, editors, *Causatives and transitivity*, page 87–120. John Benjamins, Amsterdam/Philadelphia, 1993. doi: 10.1075/slcs.23.05has.

Haspelmath, Martin. Further remarks on reciprocal constructions. In Nedjalkov, Vladimir P., editor, *Reciprocal Constructions*, page 2087–2115. John Benjamins, Amsterdam – Philadelphia, 2007.

Karlík, Petr. Reflexiva v češtině. In Rusínová, E., editor, *Přednášky a besedy z 32. běhu Letní školy slovanských studií*, page 44–52, Brno, 1999. Filozofická fakulta Masarykovy univerzity.

Karlík, Petr, Marek Nekula, and Jana Pleskalová, editors. *Nový encyklopedický slovník češtiny*. Nakladatelství Lidové noviny, Praha, 2016.

Kemmer, Suzanne. *The Middle Voice*. John Benjamins, Amsterdam – Philadelphia, 1993. doi: 10.1075/tsl.23.

Kettnerová, Václava, Markéta Lopatková, and Anna Vernerová. Reflexives in the VALLEX Lexicon: Syntactic Reflexivity and Reciprocity. *The Prague Bulletin of Mathematical Linguistics*, 117:17–60, 2021.

Komárek, Miroslav. Několik poznámek k reflexi reflexivity reflexiv. *Slovo a slovesnost*, 62(3): 207–209, 2001.

Komárek, Miroslav, Jan Kořenský, Jan Petr, and Jarmila Veselková, editors. *Mluvnice češtiny 2*. Academia, Praha, 1986. (Co-authors: Kořenský, J. – Petr, J. – Veselková, J.).

König, Ekkerhard and Volker Gast, editors. *Reciprocals and Reflexives. Theoretical and Typological Explorations*. Mouton de Gruyter, Berlin – New York, 2008.

Kopečný, František. Pasívum, reflexivní forma slovesná a reflexivní sloveso. In Bělič, Jaromír, Miloš Dokulil, Karel Horálek, and Alois Jedlička, editors, *Studie a práce linguistické I: K šedesátým narozeninám akademika Bohuslava Havránka*, pages 224–247. ČSAV, Praha, 1954.

Levin, Beth C. and Malka Rappaport Hovav. *Argument Realization*. Cambridge University Press, Cambridge, UK, 2005. doi: 10.1017/CBO9780511610479.

Lopatková, Markéta, Václava Kettnerová, Eduard Bejček, Anna Vernerová, and Zdeněk Žabokrtský. *Valenční slovník českých sloves VALLEX*. Nakladatelství Karolinum, Praha, 2016.

Mel'čuk, Igor A. Actants in Semantics and Syntax I. *Linguistics*, 42(1):1–66, 2004.

Oliva, Karel. Hovory k „sobě/si/sebe/se". In Karlík, Petr and Zdeňka Hladká, editors, *Čeština – univerzália a specifika 2.*, page 167–171, Brno, 2000. Masarykova univerzita.

Oliva, Karel. Reflexe reflexivity reflexiv. *Slovo a slovesnost*, 62(3):200–207, 2001.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description I–II. *The Prague Bulletin of Mathematical Linguistics*, 22:3–40; 23, s. 17–52, 1974–75.

Panevová, Jarmila. Valency Frames and the Meaning of the Sentence. In Luelsdorff, Philip A., editor, *The Prague School of Structural and Functional Linguistics*, page 223–243. John Benjamins Publishing Company, Amsterdam – Philadelphia, 1994.

Panevová, Jarmila. Problémy reflexivního zájmena v češtině. In Hasil, J. and J. Kuklík, editors, *Sborník přednášek z 44. běhu Letní školy slovanských studií*, page 81–88, Praha, 2001. Filozofická fakulta Univerzity Karlovy.

Panevová, Jarmila. Problémy se slovanským reflexivem. *Slavia*, 77(1-3):153–163, 2008.

Panevová, Jarmila, Eva Hajičová, Václava Kettnerová, Markéta Lopatková, Marie Mikulová, and Magda Ševčíková. *Mluvnice současné češtiny 2, Syntax na základě anotovaného korpusu*. Karolinum, Praha, 2014.

Pergler, Jiří. *Česká reflexivní deagentizace v diachronním pohledu*. Filozofická fakulta Univerzity Karlovy, Praha, 2020.

Petr, Jan, Miloš Dokulil, Karel Horálek, Jiřina Hůlková, and Miloslava Knapppová, editors. *Mluvnice češtiny 1.* Academia, Praha, 1986.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.

Veselý, Vojtěch. K slovotvorné funkci reflexivních morfémů se, si. *Naše řeč*, 101:138–157, 2018.

Štícha et al., František. *Velká akademická gramatika spisovné češtiny II. Morfologie – Morfologické kategorie / Flexe*. Academia, Praha, 2021.

Wiemer, Bjorn. Reciprocal and Reflexive Constructions in Polish. In Nedjalkov, Vladimir P., editor, *Reciprocal Constructions*, page 513–559. John Benjamins, Amsterdam – Philadelphia, 2007. doi: 10.1075/tsl.71.18wie.

Address for correspondence:
Václava Kettnerová
kettnerova@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha 1, Czechia

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 119 OCTOBER 2022 67–92

Enhancing Derivational Information on Latin Lemmas in the LiLa Knowledge Base. A Structural and Diachronic Extension

Matteo Pellegrini,[a] Marco Passarotti,[a] Eleonora Litta,[a]
Francesco Mambrini,[a] Giovanni Moretti,[a] Claudia Corbetta,[a]
Martina Verdelli[b]

[a] CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milano
[b] Università di Pavia

Abstract

In this paper[1] we document both the structural and the diachronic extension of the derivational information provided in the LiLa Knowledge Base of interoperable linguistic resources for Latin. Structurally, to the flat information on families (i.e., groups of lemmas that share the same base) and affixes that is already available for the collection of lemmas of the LiLa Lemma Bank, we add hierarchical information on derivation processes provided by the Word Formation Latin (WFL) lexical resource, which in turn is characterised by a step-to-step morphotactic approach, where lexemes that are directly derived from one another are connected through word formation rules of different kinds. This is done by modelling WFL data into an ontology that adheres to the principles of the Linked Data paradigm, and connecting these data to the LiLa Lemma Bank. From a diachronic point of view, while the previous version of WFL only took Classical Latin lemmas into account, in this paper we describe the work conducted to produce a new version of WFL that is enhanced with derivational information on Medieval Latin lemmas. We then show how the data of this new version of WFL were used to extract derivational information in the format required by the LiLa Lemma Bank.

[1] This paper is an extended version of the work presented by Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini and Giovanni Moretti at the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo), 9-10 September 2021, Nancy, France.

 Corresponding author: matteo.pellegrini@unicatt.it
Cite as: Matteo Pellegrini, Marco Passarotti, Eleonora Litta, Francesco Mambrini, Giovanni Moretti, Claudia Corbetta, Martina Verdelli. Enhancing Derivational Information on Latin Lemmas in the LiLa Knowledge Base. A Structural and Diachronic Extension. The Prague Bulletin of Mathematical Linguistics No. 119, 2022, pp. 67–92. doi: 10.14712/00326585.023.

1. Introduction

In recent years, the principles of the so-called Linked Data paradigm[2] have increasingly been applied to language data and metadata, with the aim of improving interoperability between resources that were originally developed for different purposes, and are therefore characterised by different formalisms and conceptual models. As a consequence, several resources are continuously being added to the so-called Linguistic Linked Data Cloud (Cimiano et al., 2020). Within this framework, the aim of the *LiLa* project[3] is to add Latin to this cloud, by creating a Knowledge Base (KB) of interlinked resources using a common vocabulary for knowledge description for the existing textual (i.e. corpora) and lexical (e.g. dictionaries and lexica) resources, as well as for Natural Language Processing (NLP) tools such as morphological analysers and part-of-speech taggers.

To do so, LiLa adopts the data model of the Resource Description Framework (Lassila and Swick, 1998), making use of a series of Semantic Web and Linked Data standards, including ontologies, to describe linguistic annotation (OLiA, cf. Chiarcos and Sukhareva 2015), corpus annotation (NIF, cf. Hellmann et al. 2013; CoNLL-RDF, cf. Chiarcos and Fäth 2017) and lexical resources (Lemon, cf. Buitelaar et al. 2011; OntoLex, cf. McCrae et al. 2017). According to the RDF data model, information is coded in terms of triples, that connect a subject – a labelled node – to an object – another labelled node or a literal – by means of a property – a labelled edge.

The backbone of the architecture of the LiLa KB is the Lemma Bank, a large collection of lemmas – i.e. citation forms – to which tokens from textual resources and entries from lexical resources, as well as the output of NLP tools, are connected. The Lemma Bank initially included a limited amount of derivational information on Classical Latin lemmas, taken from the Word Formation Latin (WFL) lexical resource (Litta and Passarotti, 2019). Initially, a choice was made not to include the entire information provided by WFL. That, however, might prove useful in certain circumstances.

In this paper, we start by detailing the organisation of derivational information in the Lemma Bank – in contrast with the one of the source from which it is extracted, namely WFL – and its coverage with respect to all the lemmas in the Lemma Bank (Section 2). We then show how we have extended the derivational information available in the LiLa Knowledge Base in two directions: structurally, and in terms of diachronic coverage. As for the former (Section 3), we describe a new ontology to model WFL data, so as to include it into the Lila KB in its entirety and link it to the Lemma Bank, thus having its information available in both formats within the same framework. We also discuss how our model interacts with other models developed by the Linked Data community – namely, the OntoLex-Lemon vocabulary for describing

[2] https://www.w3.org/DesignIssues/LinkedData.html.

[3] https://lila-erc.eu.

lexical resources (McCrae et al., 2017; Buitelaar et al., 2011) and, more specifically, its Morphology Module (Klimek et al., 2019; Chiarcos et al., 2022). As for the latter (Section 4), we document the work done to produce a new version of WFL, enhanced to incorporate new derivational relations regarding Medieval Latin lemmas, in addition to the Classical Latin ones of the first version. We also show how these new relations have been exploited to provide additional "flat" derivational information in the Lemma Bank for the same Medieval Latin lemmas. Lastly, we draw a number of conclusions and highlight a few directions for future work (Section 5).

2. Derivational information in the LiLa Knowledge Base

The intuition behind the way in which LiLa connects different resources and tools is based on the central role of words: the idea is that textual resources are made of occurrences of words, lexical resources describe some properties of words, and NLP tools process words. As a consequence, in LiLa's architecture, a pivotal role is played by the class `Lemma` in LiLa's ontology,[4] a subclass of the class `Form` from OntoLex-Lemon. A lemma is defined as the canonical form of a lexical item, i.e. the one that is used for citation purposes by dictionaries and lemmatisers. The core of the LiLa KB is its Lemma Bank, a collection of around 200,000 Latin lemmas taken from the database of the morphological analyser Lemlat (Passarotti et al., 2017). Lemlat's database includes Classical Latin lemmas taken from Glare (2012), Georges (1998) and Gradenwitz (1904), Medieval Latin lemmas taken from du Cange et al.'s (1883-1887) glossary and proper names taken from Forcellini's (1965) *Onomasticon*. Through the Lemma Bank, the entries of the various lexical resources represented in LiLa and the tokens of the corpora included therein can be linked to the appropriate lemma, thus achieving the desired interoperability.

WFL is a word formation lexicon of Latin, characterised by a step-to-step morphotactic approach. This means that lexemes that are considered as deriving from one another are connected via word formation rules (WFR) of different kinds, by the application of one affix or one part-of-speech change at a time (note that circumfixation is not productive in Latin). There are compounding rules – with two, or more input lexemes and one output lexeme – and derivation rules – with only one lexeme as input and one as output. Among derivation rules, depending on the presence or not of affixes and their nature, there are affixal rules (more specifically, prefixal and suffixal) and conversion, when only a change of part of speech is involved. Furthermore, rules are classified according to the part of speech of the lexemes they take as input and output. All these features are illustrated in the examples of Table 1.

In WFL all the members of the same word formation family are grouped in a hierarchical structure, resembling that of a directed tree-graph, taking root from the ancestor – the lexeme from which all the members of the family ultimately derive –

[4] `https://lila-erc.eu/lodview/ontologies/lila/`.

input lexeme(s)	output lexeme	prefix	suffix	WFR
FELIX$_A$ 'happy'	FELICITAS$_N$ 'happiness'	-	-tas	A-To-N -tas
FELIX$_A$ 'happy'	INFELIX$_N$ 'unhappy'	in-	-	A-To-A in-
MALUS$_A$ 'bad'	MALUM$_N$ 'bad thing'	-	-	A-To-N
AGER$_N$ 'field' + COLO$_V$ 'cultivate'	AGRICOLA$_N$ 'farmer'	-	-	N+V=N

Table 1. Examples of Word Formation Rules in WFL

and branching out to all derivatives by means of the successive application of individual WFRs. For example, Figure 1 shows a portion of the family taking root from the ancestor lexeme FELIX$_A$ 'happy' in WFL: the four lexemes are linked by edges labelled by the affix involved in the WFR at work.

The Lemma Bank of the LiLa KB includes only a selection of the word formation information contained in WFL. Whenever a lemma is considered "derived", it is accompanied by information related to its morphological segmentation. So each derived lemma has a relation to one or more affixes and one or more (in case of compounding) bases, merely defined as abstract connectors between lemmas that belong to the same family. Hence besides `Lemmas`, two other classes are involved, namely `Affixes` – in their turn divided into `Prefixes` and `Suffixes` – and `Bases`. Each lemma is linked to the base to which it is related by means of the property `hasBase`, and to the affixes it contains by means of the property `hasPrefix` or `hasSuffix`.[5] As a result, the organization of derivational information in the Lemma Bank is flat, rather than hierarchical. Figure 2 shows how the four lexemes in the portion of the word formation family of FELIX$_A$ of Figure 1 are linked to the same base and to their affixes in the Lemma Bank, without any representation of both the WFR and the derivational hierarchical order.

Two different perspectives on derivational morphology are thus taken by WFL and by the Lemma Bank. In the 4-way classification of resources specialized in word formation operated by Kyjánek (2020), WFL can be considered as lexeme-oriented, since it describes the relationship among individual derivationally related lexemes. The approach of the Lemma Bank, on the other hand, is family-oriented, since it identifies groups of derivationally related lexemes sharing the same base.[6]

As is argued by Litta et al. (2020), the choice of a flat organization of derivational information in the Lemma Bank is due to its compatibility with more recent, Word-and-Paradigm theoretical approaches, such as Construction Morphology (Booij, 2010). Furthermore, this approach allows for a more natural treatment of cases that were

[5]These properties are all defined in LiLa's ontology.

[6]Kyjánek's (2020) classification also identifies morpheme-oriented resources – that decompose morphologically complex words into sub-word units – and paradigm-oriented resources – that aim at a modelling consisting of aligned morphological relations.

Figure 1. Word Formation in WFL

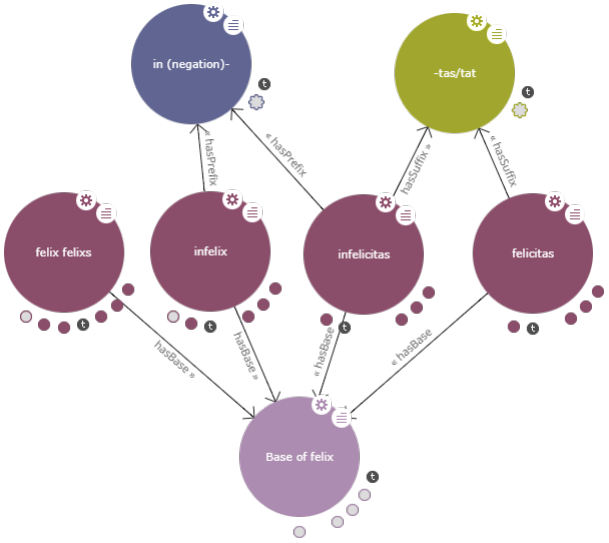


Figure 2. Word Formation in the Lemma Bank

problematic for the rigidly hierarchic structure in WFL (Litta and Budassi, 2020). For instance, WFL is forced to take a stance on the directionality of conversion processes, even when cases are not clear-cut, for instance ADVERSARIUS$_A$ 'opposed' vs. ADVERSARIUS$_N$ 'opponent'. An even more significant phenomenon is exemplified by a word like EXAQUESCO 'to become water': in this case, the step-by-step procedure of WFL requires the application of one affixation process at a time, but since neither *EXAQUO nor *AQUESCO are actually attested as intermediate steps, it has been necessary to add one of them (namely, *AQUESCO) as a fictional entry, so to comply with the requirements of WFL's general structure.

On the other hand, LiLa's flat representation of Latin word formation overlooks many details on the order of derivation. Since such information can still be potentially useful, we have decided to model the data from WFL so that it could be added to the LiLa KB. In this way, we achieve a structural enrichment of the knowledge of word formation represented in LiLa: both the flat organization of the Lemma Bank and the hierarchical organization of WFL are made available within a unified framework, leaving up to data users the choice about the kind of information that proves to be more appropriate for their specific needs. The details of the architecture of the WFL ontology designed for this purpose are the topic of Section 3.

One further direction to increase the degree of informativity of the LiLa KB on word formation concerns its diachronic coverage. At the time when WFL was com-

piled, entries from the Du Cange's medieval latin glossary had not yet been added into the Lemlat database, so WFL revolved around Classical and Late Latin only. Because the lexical basis of the LiLa KB is richer, we have felt the need to enrich its coverage even for what word formation information is concerned, and decided to keep this information in both theoretical formats, in order to offer the same level of flexibility as for the Classical Latin data. Since the bases, prefixes and affixes listed in the Lemma Bank are ultimately derived from WFL, that needed to remain the starting point for this enrichment phase. In Section 4, we describe the procedure that we followed to enhance WFL with new relations whose output is a Medieval Latin lemma, and to exploit this information to infer the base and the prefix and/or suffix of the corresponding lemmas in the Lemma Bank.

3. Modelling WFL with LiLa and Morph

The full inclusion of a lexical resource into the LiLa KB involves the modellisation of its data into an ontology that respects the Linguistic Linked Open Data (LLOD) standards. Figure 3 illustrates the details of our proposed ontology for WFL. Properties are represented as labelled directed arrows, and Classes as boxes. Boxes are colour-coded, according to the ontology where they are defined. This information is also expressed in the portion of the name that precedes the colon (e.g. `morph:Rule` means that "Rule" is a Class described in the "Morph" module of OntoLex). The arrows that are not labelled and have a white head are shortcuts for subclass relations.

Consistently with the spirit of Linked Data, our model makes use of classes and properties already defined in other ontologies. The most relevant for our purpose is OntoLex (cf. above in Section 1), both in its core model – where the class `LexicalEntry` is defined – and in more specific modules. In particular, we use the properties `source` and `target` from the Variation & Translation module (`vartrans`),[7] devised to handle relations of different kinds between lexical entries and senses, and several classes (the ones in blue in Figure 3) defined in the above-mentioned (cf. Section 1) Morphology module (`morph`). Furthermore, we refer to the classes already used in LiLa to treat derivational information (the ones in light green in Figure 3). Besides the ones taken from existing ontologies, we had to define some new classes and properties – identifiable by the `wfl` prefix and their white colour in Figure 3 – in order to properly model the information contained in WFL, as we will detail below.

There is one instance of the class `ontolex:LexicalEntry` for each lexeme contained in WFL. The entries of WFL that are directly derived from one another are linked by a specific instance of the class `morph:WordFormationRelation`, through properties taken from the `vartrans` module of OntoLex, having the entry of the base as `source` and the one of the derivative as `target`. Each relation is then connected to the WFR it instantiates (`wfl:WFLRule`) by means of the property `wfl:hasWordFormationRule`.

[7] `https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans`.

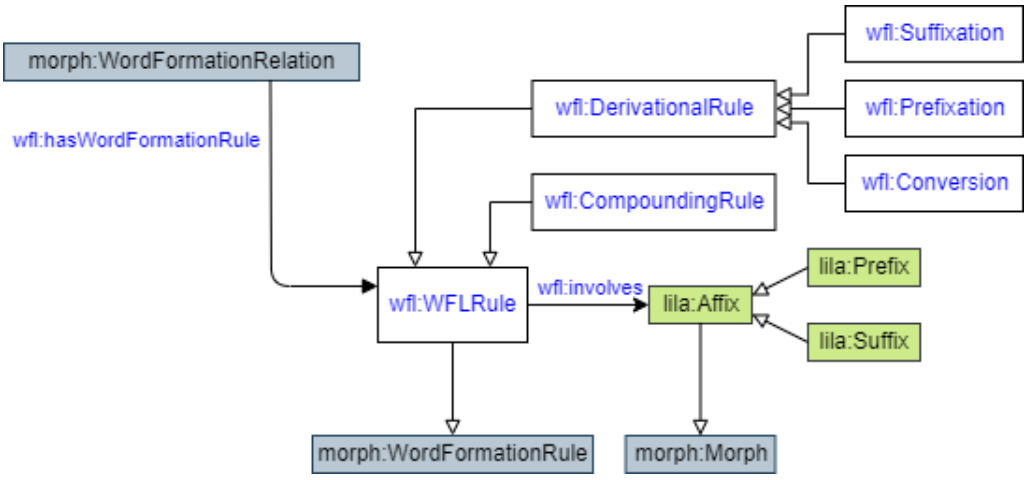


Figure 3. Architecture of the WFL ontology

The class `wfl:WFLRule` has two subclasses `wfl:DerivationalRule` and `wfl:CompoundingRule`, with the former having in its turn three subclasses `wfl:Suffixation`, `wfl:Prefixation` and `wfl:Conversion`, to reflect the organization of WFL.[8] Lastly, an object property `wfl:involves` links affixal rules to the prefix or suffix they display, as they are coded in LiLa – i.e. to an instance of either `lila:Prefix` or `lila:Suffix`, both subclasses of `lila:Affix`. Besides the use of `morph:WordFormationRelation`, the integration with the Morphology Module (`morph`)[9] of OntoLex is achieved by establishing a subclass relation between the rules of WFL and the ones of `morph` (`morph:WordFormationRule`) on the one hand, and between the affixes of the Lila ontology and the ones of `morph` (`morph:Morph`) on the other hand.

Figure 4 shows the model at work with specific pairs of related words with the Linked Data treatment of the derivation of INFELIX$_A$ 'unhappy' from FELIX$_A$ 'happy' on the one hand (left side of the image), of INFELICITAS$_N$ 'unhappiness' from INFELIX$_A$ 'unhappy' on the other hand (right side of the image).

[8]For the sake of completeness, we should mention that there is also a class `wfl:Backformation`, to account for a few cases of words that have been (probably) created by analogy, having been interpreted as the base of an already existing complex word that, however, has actually been formed by a different process. A clear example is the word CONSUEO 'to be used to', back-formed from CONSUESCO 'to become used to', that has actually been created by prefixing *con-* to SUESCO 'to become used to'. Since this phenomenon is very marginal in our data (there are only 5 cases in WFL), we do not go into more detail here.

[9]Note that this module is still the object of discussion in the Linked Data community: our proposal reflects its current state, but some details might change in the future.

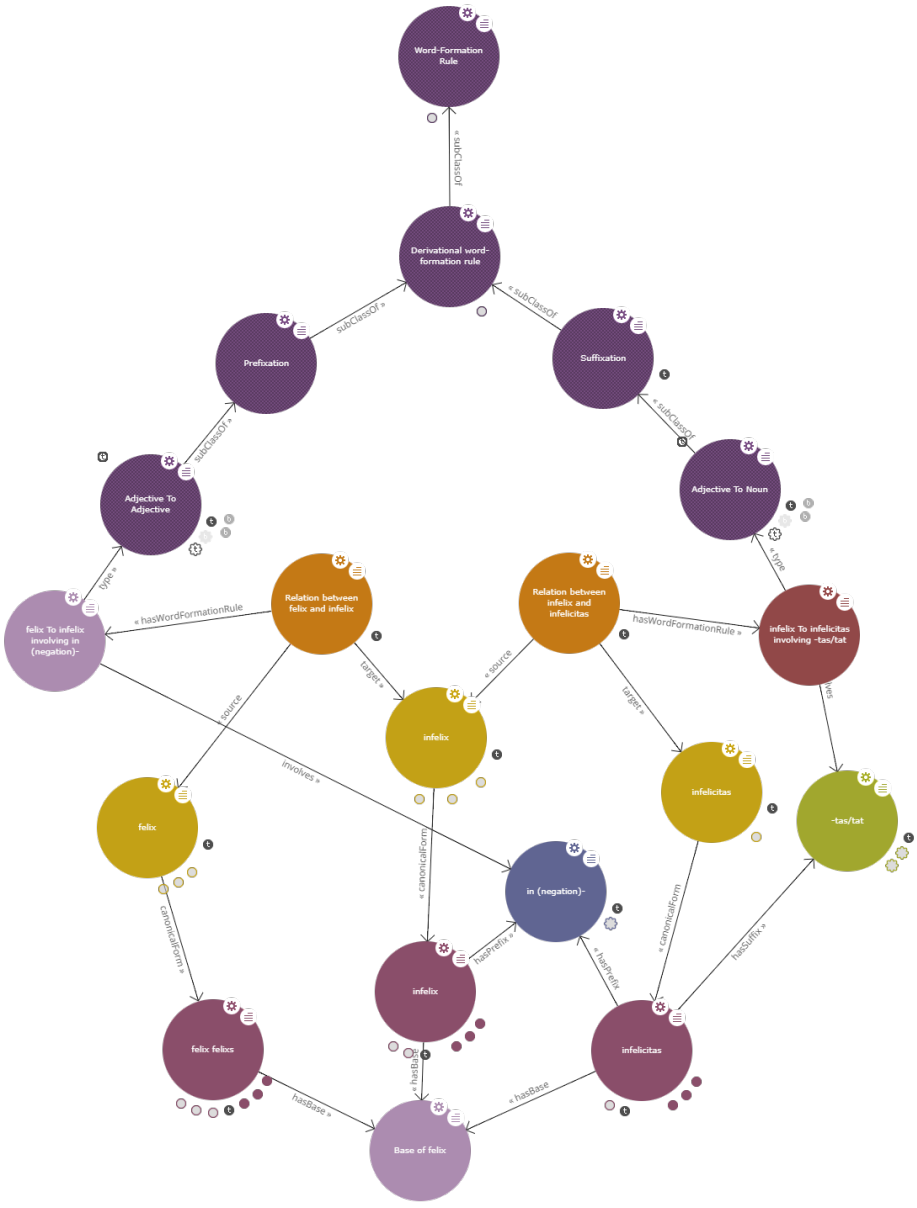


Figure 4. Modelling of prefixation and suffixation in the WFL ontology. Colours represent the classes in the LiLa ontology. E.g., dark purple: Classes; yellow: Lexical Entries; orange: WFL Relations; brownish purple: Lemmas, etc.

There is a specific word formation relation – in orange in the picture – between each of the entries of WFL that are considered as derived from one another, i.e. one between FELIX$_A$ and INFELIX$_A$ and one between INFELIX$_A$ and INFELICITAS$_N$. Each relation is instantiated by a specific WFR: see the nodes labelled as "felix To infelix involving in (negation)-"[10] and "infelix To infelicitas involving -tas/tat",[11] respectively. Starting from the one that forms INFELIX$_A$ from FELIX$_A$, it belongs to the class of prefixation rules creating adjectives from other adjectives: see the node with label "Adjective to Adjective" connected to the node with label "Prefixation" by means of the property `subClassOf` in Figure 4. Furthermore, this rule is also said to involve the prefix "in (negation)-". As for the WFR that forms INFELICITAS$_N$ from INFELIX$_A$, it belongs to the class of suffixation rules creating deadjectival nouns, and it involves the suffix "-tas/tat". Both prefixation and suffixation are sub-classes of the class of (affixal) derivational word formation rules, that on its turn is a sub-class of the class including all the rules of WFL. The bottom part of Figure 4 shows the connection with the Lemma Bank and the derivational information included therein. The lexical entries of WFL (above, in yellow) are connected to the lemmas of the Lemma Bank (below, in purple) by means of the OntoLex-Lemon property `canonicalForm`, and lemmas are connected to their shared base and to all the prefixes and suffixes they display, through the properties `hasBase`, `hasPrefix` and `hasSuffix` respectively.

There is one fact that is worth stressing in the description of this model: word formation relations always link a single source to a single target in our model. This restriction is inherited from the class of which `morph:WordFormationRelation` is stated to be a subclass, i.e. `LexicalRelation` from the `vartrans` module, that has been defined as connecting exactly two lexical entries. This has consequences on the treatment of compounding, as illustrated by Figure 5, showing the case of AGRICOLA$_N$ 'farmer' (from AGER$_N$ 'field' + COLO$_V$ 'to cultivate'). In this case, two relations are needed (one between the compound and its first member, one between the same compound and its second member), both of them pointing to the same WFR. A last remark should be made on the order of constituents, that is explicitly coded on each relation by means of the property `wfl:positionInWFR`: for instance, in the case of AGRICOLA$_N$ the value of this property is 1 for the relation between AGER$_N$ and AGRICOLA$_N$, 2 for the relation between COLO$_V$ and AGRICOLA$_A$.

For the sake of completeness, we also exemplify the treatment of noun-to-adjective conversion in Figure 6 below. It can be observed that the picture is similar to the one of affixal derivation (see Figure 4 above), the only difference being that the rule is not stated to involve any affix, consistently with the definition of conversion.

[10]The negative meaning of the prefix *in-* is specified to distinguish it from its omograph meaning 'entering', appearing for instance in INEO 'to go into, enter' from EO 'to go'.

[11]The notation of the shape of the suffix reflects the presence of different stem allomorphs in different forms, e.g. NOM.SG *infelici-tas* vs. GEN.SG *infelici-tat-is*.

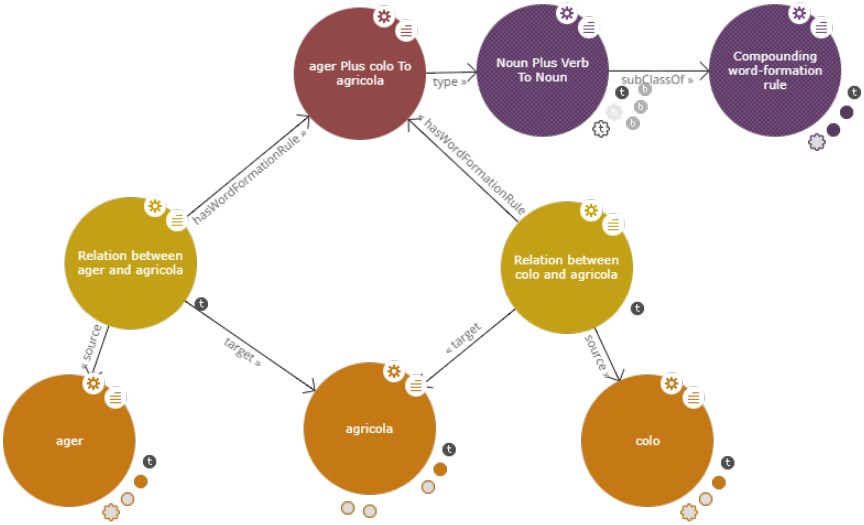


Figure 5. Modelling of compounding in the WFL ontology

4. Extracting derivational information for Medieval Latin lemmas

This section describes the procedure that we followed to enhance WFL with relations regarding Medieval Latin lemmas, and consequently to provide derivational information on those lemmas also in the Lemma Bank.

Our procedure is articulated (a) in an automatic extraction of words that might be derivationally related with one another based on their form – described in Section 4.1 – and (b) in a manual validation of the pairs that have been identified – addressing a number of issues raised during the process, some of which are discussed in Section 4.2. We then add the newly created relations to WFL and to Lemlat's database and we use them to establish new triples having an existing lemma of the Lemma Bank as subject and `hasBase`, `hasPrefix` and `hasSuffix` as properties, as detailed in Section 4.3. This section concludes by presenting some quantitative data on the outcome of this procedure in Section 4.4.

4.1. The methodology

The first step in our procedure is to identify the derivational processes that are most relevant in Classical Latin in terms of frequency: these are our starting point for the addition of new derivational information. To this end, we exploit the data of WFL, looking at the number of relations instantiating each WFR. For instance, the

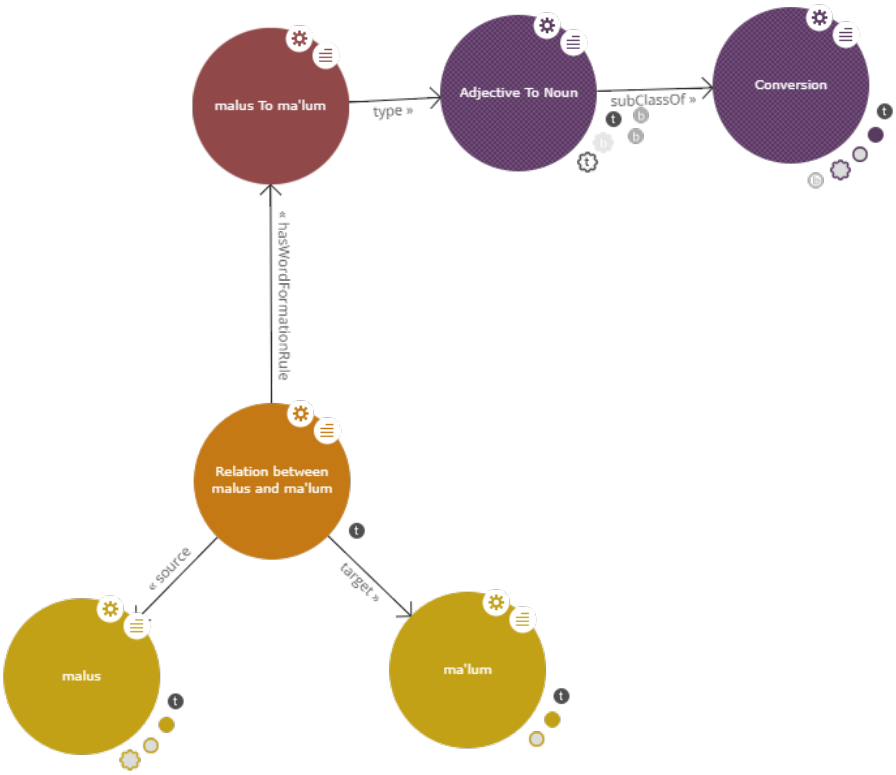


Figure 6. Modelling of conversion in the WFL ontology

single most frequent rule proves to be the one that derives (mostly action) nouns from verbs using the suffix labelled as -(*t*)*io*(*n*)- in WFL: there are 2,555 relations that instantiate this rule – among them, DUCTIO$_{N}$ 'leading' from DUCO$_{V}$ 'lead'.

The following step is to look for Medieval Latin lemmas that are potentially the output of one of these rules, based on their form. In cases of suffixation, this is done by first identifying Medieval Latin lemmas that end with a sequence of segments matching the suffix at hand, and then going through all lemmas – both Classical and Medieval Latin – and checking if they can be the input of the rule introducing that suffix, on the basis of the form of the potential input and output.

For instance, as for the above-mentioned action nouns ending in -(*t*)*io*(*n*)-, we can see that these are normally derived by taking the Third Stem (cf. Aronoff 1994) of the verb – in the example above, *duct-* – and adding the suffix *-ion-*, followed by the appropriate case endings – with the final nasal segment being regularly dropped in the nominative and vocative singular, thus yielding citation forms in *-io*. Our procedure consists in selecting all Medieval Latin lemmas whose citation form ends in *-io*, and keeping only the ones for which, among all verbs, there is at least one that can potentially be the base according to the regular formal procedure described above.

To give an example, among the Medieval Latin lemmas provided by Lemlat we find RETRUSIO$_{N}$ 'inclusion'; and among all lemmas from Lemlat (both Classical and Medieval), we find RETRUDO$_{V}$ 'thrust back'. The latter has *retrus-* as its Third Stem, as is shown by its perfect participle *retrusus*, that is listed in the Lemma Bank as a member of the class `Hypolemma`.[12] Hence, we identify a potential derivational relation between these two lemmas, as it is plausible that RETRUSIO$_{N}$ is an action noun formed from RETRUDO$_{V}$.

It should be noted that in some cases we also extract bases belonging to a part of speech different than the one that is required as input by the relevant rule. For instance, as potential inputs for adjectives in *-alis* we extract not only nouns, as in the most frequent case (e.g. ABYSSUS$_{N}$ 'bottomless pit' for ABYSSALIS$_{A}$ 'bottomless'), but also adjectives (e.g. ASSIDUUS$_{A}$ 'constantly present' for ASSIDUALIS$_{V}$ 'assiduous'). This is justified by the fact that adjectives are also attested – although much less frequently – as bases of adjectives in *-alis* already in Classical Latin (e.g. NOVALIS 'that is ploughed anew' from NOVUS 'new').

As for prefixation, we rely on a similar intuition, but follow a procedure that goes in the opposite direction: we start from all the lemmas, both Classical and Medieval, and look for the Medieval ones that, based on their form, can be analysed as being derived by adding one of the several Latin prefixes to it. For instance, for MELIORO$_{V}$ 'improve' we find three potential prefixed verbs that come Du Cange's glossary, namely

[12]This makes it possible to accomodate the different ways in which a participle like *retrusus* can be lemmatised, i.e. as an adjective (hypolemma RETRUSUS, `http://lila-erc.eu/data/id/hypolemma/38805`) or as the verb from which the participial form is created (lemma RETRUDO$_{V}$ `http://lila-erc.eu/data/id/lemma/122792`). For further details on the architecture of the LiLa KB and its use of hypolemmas, see Passarotti et al. (2020).

EMELIORO$_V$ 'improve/correct' (with prefix *e(x)-*), IMMELIORO$_V$ 'improve' (with prefix *in-*) and REMELIORO$_V$ 'correct' (with prefix *re-*). The application of this different procedure is partly due to practical reasons. The fact that inflection in Latin is mostly suffixal makes it difficult to predict the shape of the outcome of the application of a suffixation rule, as it is preliminarily necessary to strip the inflectional endings in order to identify the base to which the suffix should be added, and allomorphy plays a relevant role due to inflection class distinctions and other less systematic facts – for instance, the different stem variants of 3rd declension nouns, usually due to phonological adjustments triggered by the addition of particular endings (e.g. NOM.SG *dent-s* > *dens* 'tooth' vs. GEN.SG *dent-is*), but also simply suppletive stems (e.g. NOM.SG *iecur* 'liver' vs. GEN.SG *iecinoris*). On the left side of words, instead, there is much less allomorphy. As a consequence, it is much easier to predict the shape of the outcome of a prefixation rule, at least as long as we are dealing with category-preserving rules, for which we do not even need to guess the final part of the derived word, and can simply assume it to be the same as in the base. From a theoretical standpoint, this choice is also motivated by the fact that different preverbs in Latin, and in Indo-European languages in general (cf. Lehmann 1983, Booij and van Kemenade 2003) appear to have common characteristics that make it reasonable to treat them as a unique process – for instance, most of them have a basic spatial or temporal meaning, besides other metaphoric or idiosyncratic uses. On the other hand, suffixation rules differ markedly from each other in their morphological and semantic behaviour.

For the time being, we do not deal with compounding, as it would require a completely different treatment, which we leave for future work. Conversion is also currently out of the picture. By definition, in cases of conversion there is no formal marker of the derivation process except (possibly) for the different inflectional endings, hence it would not be possible to apply the procedure outlined above as it is. The only exception is the process of conversion that takes the Third Stem of a verb as input and produces a noun as output. The inclusion of such cases in our procedure is made possible by the fact that we can exploit LiLa's `Hypolemma` information (cf. Footnote 12 above) to extract lemmas that can plausibly be considered as potential bases of the conversion process, in a way similar to the one we have described above for suffixes that apply to the Third Stem like *-(t)io(n)-*. For instance, the Medieval Latin lemma DISPLICITUS$_N$ 'offence' can be presumed to be a conversion from the Third Stem *displicit-* of the verb DISPLICEO 'displease'. The reason for the inclusion of this conversion process is that among the rules that we consider there are also denominal nouns in *-atus*; but many of the nominal lemmas of LiLa that end in *-atus*, given their meaning, seem to be better analysed as conversions from the Third Stem of a verb rather than as obtained by adding the suffix *-at-* to a nominal or adjectival base. Hence, by extracting all the possible bases we are in a position to manually select the most appropriate one on a case-by-case basis (cf. Section 4.2 below).

Among the pairs of lemmas extracted following the procedure described above there are also false positives, either because the formal similarity between the two

lemmas is coincidental, or due to the noisy nature of the data we are dealing with (cfr. Section 4.4 below). Since we aim at having high-quality data, we also performed a manual validation of the base-derivative pairs that were extracted in this first phase, as detailed in the next section.

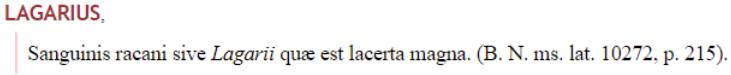
4.2. Issues in the annotation process

This section concerns issues related to the manual validation of the base-derivative pairs automatically extracted from the procedure described above. We discuss two issues in particular: in Section 4.2.1 we focus on the choice between two or more candidates that are automatically extracted as input for the same output; in Section 4.2.2, we deal with problematic entries of Du Cange's glossary.

4.2.1. Manual selection among two or more potential inputs for the same output lemma

First of all, alongside cases in which for a given output only one input was extracted, which were checked by looking at their semantics, in several cases the automatic selection detected two or more candidates as possible input lemmas. Therefore, we manually identified the most appropriate input lemma among all the ones that had been extracted. For instance, for the Medieval derivative noun COALITIO$_N$ 'assemblage, meeting', we automatically extracted two candidate inputs, COALO$_V$ 'feed, sustain, nourish' and COALESCO$_V$ 'grow together with something, unite', both of them coming from the Classical Latin section of Lemlat's database. This is motivated by the fact that both these verbs have *coalit-* as their Third Stem. In this case, COALESCO$_V$ is semantically closer to the derivative COALITIO$_N$ than COALO$_V$. Consequently, we selected the former as the input.

We also had to deal with several cases in which we automatically extracted two or more candidates with a similar meaning. These cases usually involved a Classical and a Medieval Latin lemma, as we can see in the noun COLLEGARIUS$_N$ ('one of the colleagues'). For this lemma, the automatic selection detected both COLLEGA$_N$ 'colleague' and COLLEGUS$_N$ 'companion' as possible inputs. The first one is a Classical Latin lemma, while the second one is attested only in Du Cange's glossary. Because of their semantic similarity, both candidates are potentially the input from which COLLEGARIUS$_N$ is formed. In such cases, we choose the Classical lemma, for practical reasons: as we will see in more detail in Section 4.3, when the input of the relation is a Classical Latin lemma that is present in WFL – as happens in this case – we are often in a position to attach the new relation to the corresponding tree – in this case, the one having LEGO$_V$ 'read' as its root. As a consequence, we can also infer information on the base and prefixes/affixes to which the corresponding lemma in LiLa should be linked. If we had selected the Medieval Latin lemma as input, on the other hand, we would have missed the relation of COLLEGARIUS$_N$ with the verb LEGO$_V$ and with the prefix *con-*, as no information on the derivational history of COLLEGUS$_N$ is provided in WFL and in the LiLa Lemma Bank.


LAGARIUS,

Sanguinis racani sive *Lagarii* quæ est lacerta magna. (B. N. ms. lat. 10272, p. 215).

Figure 7. Entry for LAGARIUS *in Du Cange's glossary*

CONSTABILITOR, in Charta Roberti II. Principis Capuani apud Michaelem Monachum in Sanctuario Capuano pag. 643.

Figure 8. Entry for CONSTABILITOR *in Du Cange's glossary*

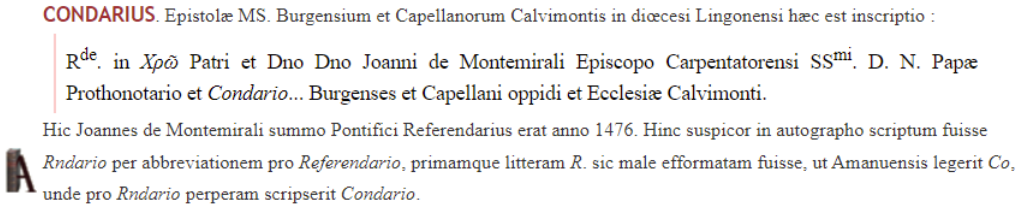
4.2.2. Issues related to the nature of Du Cange's entries

In addition to the previous issue, we also had to deal with some peculiarities of our source of Medieval Latin lemmas, which is a glossary, rather than a dictionary, and also has a complex publishing history (Géraud, 1839): for these reasons, there are many issues related to the descriptions or definitions of lemmas provided by glossators. More specifically, the glossary presents several cases of lemmas with quotes or bibliographic references, but without a definition, as we can see in the entries reported in Figures 7 and 8.

In both cases the glossator does not provide a definition, but just quotes the text where the lemma derives from (as we can see in 7) or provides the bibliographic reference where the lemma was found (as we can see in 8). Since we did not have an explanation of the meaning of the derivatives LAGARIUS$_{A}$ and CONSTABILITOR$_{N}$, we cannot be sure on their derivation history, therefore we decided not to include them in our validation process.

Moreover, there are philological issues taken into account by the glossator in the definition of the lemmas. An example is given in Figure 9 below.

In the entry, the glossator provides the source of the lemma and quotes the text where the lemma is taken from, without giving an explanation of the meaning. He actually proposes to amend the form *condario* to *Rndario* (hypothesizing an abbreviation for the word Referendario 'referendum'), identifying the former as an amanuensis' mistake. Therefore, it is not even sure that such a word ever existed in Latin: hence, also in such cases we decided not to consider relations involving those lemmas as valid.



CONDARIUS. Epistolæ MS. Burgensium et Capellanorum Calvimontis in diœcesi Lingonensi hæc est inscriptio :

> R^de^. in *Xpõ* Patri et Dno Dno Joanni de Montemirali Episcopo Carpentatorensi SS^mi^. D. N. Papæ Prothonotario et *Condario*... Burgenses et Capellani oppidi et Ecclesiæ Calvimonti.

Hic Joannes de Montemirali summo Pontifici Referendarius erat anno 1476. Hinc suspicor in autographo scriptum fuisse *Rndario* per abbreviationem pro *Referendario*, primamque litteram *R*. sic male efformatam fuisse, ut Amanuensis legerit *Co*, unde pro *Rndario* perperam scripserit *Condario*.

Figure 9. Entry for CONDARIUS *in Du Cange's glossary*

4.3. Organizing the information

As an outcome of the procedure described above in Sections 4.1 and 4.2, we obtain data in the form of a series of derivational relations between an input and output lemma. To release this information, first of all we need to include it into the relational database that contains the information displayed by WFL. One of its tables reports a list of relations, with fields for the identifier of the WFR at play and input and output lemmas. After adding our new relations to this table, we can release WFL 2.0, a new version of WFL that also covers Medieval Latin.[13] This is the version of WFL that is modelled as an ontology as described in Section 3 above and consequently included into the LiLa KB.[14] Its data can also be accessed with the same tree-based graphical interface used for the previous version of WFL.[15]

Furthermore, the morphological analysis of a wordform performed by Lemlat exploits this knowledge to provide the user with derivational information on the lemma that is assigned to the analysed wordform, if such information is available. Specifically, the analysis outputs i) the lemma (or lemmas, in cases of compounding) from which the wordform's lemma is derived, ii) the derivational process at play, iii) its type and iv) the affix (prefix or suffix) involved in this process, as illustrated in Figure 10 below. Hence, thanks to the work described in the previous sections, such derivational information will be provided in Lemlat's analyses also for some Medieval Latin lemmas, and not only for Classical Latin lemmas.

A caveat needs to be made on the structure of the data of WFL 2.0. While the output of the newly added relations is always a Medieval Latin lemma, the input can be either a Classical Latin or a Medieval Latin lemma. In the latter case, it is not possible to attach it to any of the trees of the previous version of WFL, as only Classical Latin is

[13] https://github.com/CIRCSE/WFL.

[14] http://lila-erc.eu/data/lexicalResources/WFL/Lexicon.

[15] https://wfl.marginalia.it/

```
SEGMENTATION:  felicitat -em
---------- morphological feats ----
--afs--

Case:   Accusative
Gender: Feminine
Number: Singular
        =================LEMMA ================
        felicitas         N3B f0322 f
        --------------morphological feats-------
        NcC

        PoS:   Noun
        Type:  Common
        Inflexional Category:  III decl
        --------------derivational info--------
        IS DERIVED: YES
        --------------rule id: 48--------------
        Lexical Basis:
                felix           N3A f0327 * Af-
        Derivational Type: Derivation_Suffix
        Derivational Category: A-To-N
        Affix: tas/tat
```

Figure 10. Morphological analysis of a wordform by Lemlat

represented there: hence, we end up with unconnected branches.[16] In some cases, this reflects the fact that both input and output are Medieval Latin innovations, that cannot be related to any Classical Latin lemma: for instance, the Medieval Latin lemma ACUNYDO$_V$ 'be suspicious', that we identify as the input of the derivational process forming ACUNYDAMENTUM$_N$ 'suspiciousness', cannot be linked to any Classical Latin lemma – it is tagged as *vox catalonica* in Du Cange's glossary. In other cases, it would be possible to attach these unconnected branches to an existing WFL tree, by identifying the intermediate step(s) in the derivational history. For instance, we find a derivational relation between the two medieval Latin lemmas VICINO$_V$ 'be close' and ABVICINO$_V$ 'separate'. The base VICINO$_V$ could be connected to a WFL tree by establishing a relation between it and the Classical Latin adjective VICINUS$_A$ 'close' as input – relation that is not identified by our automatic procedure because it is a conversion process. Hence, these missing links will eventually be found when progressing on the coverage of Medieval Latin lemmas.

In addition to releasing this new version of WFL and enhancing the informativity of Lemlat's analyses, we have also decided to release our data in the format required by the flat structure of the Lemma Bank, using our new relations to infer new triples

[16] Unless we have identified also a relation where the same lemma appears as output in our procedure, and whose input is a Classical Latin lemma for which we do have a tree in WFL.

Lemma	hasBase	hasPrefix	hasSuffix
DEFECTIVUS$_A$	Base of FACIO$_V$	*de-*	*-(t)iu-*
INDEFECTIVUS$_A$	**Base of FACIO$_V$**	***de-, in-***	***-(t)iu-***

Table 2. New derivational information in the Lemma Bank: transfer of base, prefix and suffix from input to output lemma

lemma	hasBase	hasPrefix	hasSuffix
ACUNYDO$_V$	**Base of ACUNYDO$_V$**		
ACUNYDAMENTUM$_N$	**Base of ACUNYDO$_V$**		***-ment-***

Table 3. New derivational information in the Lemma Bank: generation of new bases

that connect lemmas to their base, prefix and/or suffix through the dedicated properties `hasBase`, `hasPrefix`, `hasSuffix`. When the input of the relation is a Classical Latin lemma, we often already have information on its base and (possibly) on affixes displayed by it. In such cases, we can simply transfer this information also to the output, that inherits the base and (possibly) affixes of its input, and additionally displays the prefix or suffix involved in the rule that relates the two (except for cases of conversion). For instance, we find a relation between the Classical Latin lemma DEFECTIVUS$_A$ 'missing' and the Medieval Latin lemma INDEFECTIVUS$_A$ 'not missing'. In the Lemma Bank, the former is related to the base of FACIO$_V$ 'make', to the prefix *de-*, and to the suffix *-(t)iu-*. Hence, also the latter will inherit all these connections, and it will be additionally related to the prefix *in-* that is involved in the WFR that relates the two lemmas, as illustrated in Table 2.[17]

On the other hand, when the input is a Medieval Latin lemma, we do not have derivational information on it. Therefore, all we can do is expressing the fact that the input and output lemmas are part of a same family, by generating a new base and linking both of them to it. For instance, the input of the relation between ACUNYDO$_V$ 'be suspicious' and ACUNYDAMENTUM$_N$ 'suspiciousness' is a Medieval Latin lemma itself, hence it is not connected to a base, prefix nor suffix. Therefore, we generate a new base – that we label with the label of the input – and connect both ACUNYDO$_V$ and ACUNYDAMENTUM$_N$ to it. Furthermore, the output is connected to suffix *-mentum* involved in the rule relating it to the input (see Table 3).

[17]In Tables 2 and 3, we distinguish the pieces of information that are already in the Lemma Bank from the ones that we add in this phase by highlighting the latter in bold.

Note that the same treatment is applied when there is no derivational information on the input even if it is a Classical Latin lemma. This happens when the lemma is the only member of its family, and hence it is not derivationally related to any other lemma.

4.4. A quantitative evaluation

In this section, we offer a quantitative evaluation of the different types of derivational information that have been obtained. Table 4 summarizes the rules of WFL that we select as the ones for which we look for new potential derivatives in Medieval Latin lemmas, sorted by descending frequency (i.e., number of relations in WFL). For each rule, we also report how many new potential relations were extracted automatically.[18] For suffixal rules, we report in how many cases more than one input candidate was extracted following the procedure described in Section 4.1, and hence a choice has to be made on which of them is the most appropriate.[19] Lastly, we report how many relations are kept after the manual validation described in Section 4.2. This allows us to evaluate the precision of the automatic procudure to identify candidate pairs. It can be observed that there is a lot of variation across rules regarding the proportion of relations that are considered to be valid after the manual checking among all the relations that are extracted automatically: the percentage ranges from 95.24% for adjectives in *(-t)iu-* to as low as 34.42% for diminutives in *-ul-*, with an average of about 63%. In some cases, the decision to exclude a relation that was extracted automatically is ultimately due to the remarkable amount of noise in the source of Medieval Latin lemmas: we have already seen in Section 4.2 that there are words whose meaning is doubtful, or even entries that refer to forms that are plausibly a copyist mistake, or for which there are philological issues of other kinds. However, the main reason behind this quite low precision is a principled choice in the design of the rules used to extract input and output candidates. These are intended to capture as many potential pairs as possible, including cases where the formal relation between input and output lemmas is less than fully regular, at the cost of a higher number of false positives. Indeed, the phase of manual validation is intended to identify exactly these false positives, that consequently do not affect the final quality of the data. On the other hand, if the rules had been designed to be more restrictive, precision would have been increased at

[18] We do not provide this information for diminutive nouns in *-(us/un)cul-* because in that case we did not extracted candidates automatically. They are included because diminutives that end in *-(us/un)culus/a/um* also end in *-ulus/a/um*, hence they are sometimes identified as the output of the latter rule, and manually corrected when they are actually the output of the former, like in the case of BELLICULUM$_{N}$ 'simulated war', that is a diminutive of *bellum*$_{n}$ 'war' with suffix *-cul-* rather than a diminutive of *bellicum*$_{n}$ 'war signal' with suffix *-ul-*.

[19] For prefixal rules, this never happens because we start from the potential input and look for potential outputs, rather than the reverse: hence, the corresponding cell is left empty.

process	n. relations in WFL 1.0	n. potential new pairs	n. outputs with >1 possible input (%)	n. valid pairs (%)
V-To-V prefixation	4,850	2,194	–	1,129 (51.46%)
V-To-N -(t)io(n)- suffixation	2,555	458	41 (8.95%)	423 (92.36%)
V-To-N -(t)or- suffixation	1,419	382	44 (11.52%)	321 (84.03%)
V-To-N conversion	1,074	210	33 (15.71%)	140 (66.67%)
A-To-N -tas/tat- suffixation	623	225	27 (12.00%)	192 (85.33%)
N-To-A -os- suffixation	563	203	115 (56.37%)	152 (75.00%)
N-To-A -al- suffixation	547	176	100 (56.82%)	126 (71.59%)
A-To-A in- prefixation	508	101	–	64 (63.37%)
N-To-A -ari- suffixation	467	586	315 (53.75%)	396 (67.58%)
N-To-N -ari- suffixation	452	596	355 (59.56%)	387 (64.93%)
N-To-N -ul- suffixation	427	491	299 (60.90%)	169 (34.42%)
V-To-N -(t)ric- suffixation	415	33	6 (18.18%)	25 (75.76%)
N-To-A -at- suffixation	404	342	162 (47.37%)	198 (57.89%)
V-To-A -bil- suffixation	390	145	27 (18.62%)	117 (80.69%)
N-To-N -(us/un)cul- suffixation	370	–	–	21
V-To-V -(i)t- suffixation	343	89	40 (44.94%)	46 (51.69%)
N-To-A -ic- suffixation	339	126	75 (59.52%)	67 (53.17%)
N-To-A -in- suffixation	307	86	51 (59.30%)	43 (50.00%)
V-To-A -(t)iu- suffixation	289	63	18 (28.57%)	60 (95.24%)
V-To-N -ment- suffixation	277	380	95 (25.00%)	304 (80.00%)
N-To-A -e- suffixation	242	56	32 (57.14%)	31 (55.36%)
V-To-I -(at)im- suffixation	203	95	39 (41.05%)	58 (61.05%)
V-To-V -sc- suffixation	199	34	8 (23.53%)	18 (52.94%)
N-To-N -at- suffixation	85	192	135 (70.31%)	71 (36.98%)
TOTAL	–	7,263	–	4,558 (62.76%)

Table 4. Automatically extracted and manually validated relations for Medieval Latin lemmas

number of relations	Classical Latin	34,960
	Medieval Latin	4,558
n. lemmas in Lemlat's database	Classical Latin	43,407
	Medieval Latin	86,745

Table 5. Coverage of Classical and Medieval Latin lemmas in WFL 2.0

the expense of recall, but then there would have been no way to recover the marginal cases that were left out.

Another consequence of this choice is the remarkable proportion of cases for which more than one potential input is found by the automatic procedure: on average, this happens for about 40% of the output candidates, and for some rules this is the case for the majority of them, as shown in the third column of Table 4.

In Table 5, we provide data on the coverage of Classical Latin and Medieval Latin lemmas in WFL 2.0, by giving the number of relations that have a Classical Latin lemma as output (i.e. the ones of WFL 1.0) vs. the ones that have a Medieval Latin lemmas as output (i.e. the ones that are added to WFL 2.0). We also give the number of Classical and Medieval Latin lemmas in Lemlat's database for reference.

Unsurprisingly, the coverage of Medieval Latin is much lower: while all Classical Latin lemmas have been taken into account, for Medieval Latin we have focused only on the most frequent processes reported in Table 4 above. Therefore, the coverage of Medieval Latin can still increase when other relations will be added (see Section 5 below).

For what an evaluation of the derivational information added for Medieval Latin lemmas in the Lemma Bank is concerned, Table 6 gives the number of new triples that connect lemmas to their base, prefix, and/or suffix through the respective dedicated properties. As for the property `hasBase`, we also report how many of the new triples have a Classical Latin lemma as subject. This happens when a relation is established whose input is a Classical Latin lemma that is not related to any other Classical Latin. Since a LiLa's `Base` is nothing but an abstract connector between lemmas of the same family, such lemmas have no base as long as only Classical Latin is considered; if a relation is found with a Medieval Latin lemma, a new `Base` has to be established, as we have seen above in Section 4.3. Table 6 also reports the number of new `Bases` introduced into the Lemma Bank.

5. Conclusions

In this paper, we have described the work that was conducted to extend the derivational information available in the LiLa KB (Section 2) in two directions: structurally, by making also the hierarchical organization of WFL data available into the KB, along-

new triples with `hasBase` property	Classical Latin	96
	Medieval Latin	5,696
new triples with `hasPrefix` property	Classical Latin	–
	Medieval Latin	2,043
new triples with `hasSuffix` property	Classical Latin	–
	Medieval Latin	4,156
new `Bases`		1,143

Table 6. New derivational information in the Lemma Bank

side the flat organization of those same data in the Lemma Bank (Section 3); and diachronically, by extending WFL to cover also Medieval Latin lemmas, and consequently providing information on those lemmas also in the Lemma Bank (Section 4).

In Section 2, we have hinted at the reasons behind the choice of adopting a paradigmatic approach to word formation in the LiLa Lemma Bank – thus yielding a flat structure of related lexemes belonging to the same family. However, there are cases where the more detailed, hierarchical information provided by WFL on the order of application of different word formation processes can prove helpful. For instance, an advantage of the hierarchical structure of WFL is that it allows to focus on smaller, more tightly connected sub-sections of word formation families. This can be helpful especially when dealing with very large and quite heterogeneous families, e.g. the one of the verb FACIO 'to make', which includes 689 lemmas in the Lemma Bank. Since the semantic connection between some members of this family is quite loose, it might be useful to be able to zoom on smaller sub-families with a higher degree of internal semantic cohesion, isolating e.g. only those lexemes that are directly related to the adjective DIFFICILIS 'difficult' (e.g. PERDIFFICILIS and SUBDIFFICILIS 'very/somewhat difficult'), or only the verbs formed by adding a prefix to FACIO itself (e.g. INFICIO 'to put into' and PERFICIO 'to achieve'[20]). Such a focus on sub-families cannot be performed with the representation of word formation in the Lemma Bank, where all lemmas belonging to the same word formation family are simply connected to their common base without any further information about the hierarchy of derivations, whereas in WFL each derived lexeme is directly linked to its source lexeme.

In other cases, however, the flat organization of derivational information in the Lemma Bank can prove helpful. As an example, when considering prefixed and suffixed words, for some purposes it can be useful to focus only on those words that are actually formed by means of a WFR that involves a specific affix, while for other pur-

[20]The different shape of the stem in the base vs. derivative is due to a phonological process of weakening of short vowels in non-initial syllables.

poses it might be better to collect all those words that display that affix somewhere along their word formation history. Consider for instance the structural difference between the adjectives INFRUCTUOSUS 'unfruitful' and INIURIOSUS 'injurious': the former is created by prefixing *in-* (negation) to FRUCTUOSUS 'fruitful' (*INFRUCTUS is not attested as a Latin word), while the latter is formed by suffixing *-os* to INIURIA 'injury' (*IURIOSUS). Therefore, when investigating e.g. *in-* prefixation, it is a matter of choice whether to include also cases like *iniuriosus*. If we want to exclude them, this has to be done using the hierarchical information of WFL. Conversely, however, if we decide to include such cases, then the relevant information can be obtained by exploiting the flat structure of the Lemma Bank, where all lemmas are linked to all the prefixes and suffixes they display, regardless of their order of application in the word formation history. Although, in this specific case, it would be possible to construct a query that goes down one step in the hierarchy of WFL, things would be even more difficult in cases featuring more than two affixes – consider for instance a word like the adverb INADDUCIBILITER 'unobstructively' (lit. 'not in a way that can be pulled back and forth'), with prefixes *in-* (negation) and *ad-* and suffixes *-bil-* and *-ter*.

One of the main advantages of adopting Linked Data principles and models to represent and publish linguistic information provided by distributed resources is that this makes it possible to represent different approaches within a unified framework, as it is clearly shown in Figure 4. Scholars can choose the approach that is more compatible with their theoretical view, or simply the one that provides the kind of information more appropriate for the case at hand, also allowing to make different approaches interact easily, in case several pieces of information from different sources are needed.

This motivates our structural decision of modelling WFL data into an ontology and linking them to the LiLa KB. As for the diachronic extension documented in this paper, increasing the coverage of our derivational resources to Medieval Latin represents a first step toward filling a gap that is widespread in digital lexical resources for Latin, which tend to focus on the Classical and Late periods. However, it emerges from the data of Table 5 that a full coverage of Medieval Latin lemmas is still far. Hence, it is necessary to explore some possibilities for future work in this direction.

The most obvious thing to do would be to carry on with the same procedure, extracting pairs of lemmas that are potentially related by means of other derivational processes. However, it can be observed that the number of potential pairs is already quite low for some of the more frequent processes already considered and listed in Table 4 above. Therefore, moving on to processes that are even more marginal, the number of additional pairs is likely to progress quite slowly.

Another possibility that might be explored is a machine learning approach to the task, following the hint of recent work such as Lango et al. (2021), Svoboda and Ševčíková (2022) and Kyjánek et al. (2022). In our case, it could be interesting to try using WFL 1.0 as a gold standard to train a machine learning algorithm whose task is to identify derivational relations within a lexicon, and then applying this algorithm to Medieval Latin lemmas for which derivational information is still lacking. How-

ever, it should be highlighted that the data provided in WFL and in the LiLa KB are meant to be used as a reference for philological and linguistic work, rather than for massive NLP on big data. The accuracy of the fully automatic methods proposed in other studies, although variable,[21] do not in any case reach the close-to-perfect value that would be needed for these purposes. Hence, a phase of *ex post* manual check – possibly enhanced with semi-automatic means – would probably be necessary in any case.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

Bibliography

Aronoff, Mark. *Morphology by Itself: Stems and Inflectional Classes*. MIT Press, Cambridge, MA, 1994.

Booij, Geert. Construction morphology. *Language and linguistics compass*, 4(7):543–555, 2010. doi: 10.1017/9781139814720.016.

Booij, Geert and Ans van Kemenade. Preverbs: an introduction. In Booij, Geert and Jaap van Marle, editors, *Yearbook of morphology 2003*, pages 1–11. Kluwer, Dordrecht, 2003. doi: 10.1007/978-1-4020-1513-7_1.

Buitelaar, Paul, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. Ontology Lexicalization: The *lemon* Perspective. In *Proceedings of the Workshops-9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*, pages 33–36, 2011.

Chiarcos, Christian and Christian Fäth. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Gracia, Jorge, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Language, Data, and Knowledge*, pages 74–88, Cham, Switzerland, 2017. Springer. ISBN 978-3-319-59888-8. doi: 10.1007/978-3-319-59888-8_6.

Chiarcos, Christian and Maria Sukhareva. OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386, 2015. doi: 10.3233/SW-140167.

Chiarcos, Christian, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. Computational Morphology with OntoLex-Morph. In *8th Workshop on Linked Data in Linguistics*, pages 78–86. European Language Resources Association (ELRA), 2022.

Cimiano, Philipp, Christian Chiarcos, John McCrae, and Jorge Gracia. *Linguistic Linked Data*. Springer, 2020. doi: 10.1007/978-3-030-30225-2.

du Cange, Charles du Fresne sieur, bénédictins de la congrégation de Saint-Maur, d. Pierre Carpentier, Johann Christoph Adelung, G. A. Louis Henschel, Lorenz Diefenbach, and Léopold Favre. *Glossarium mediae et infimae latinitatis*. Favre, Niort, France, 1883–1887.

[21] The reader is referred to the publications cited above for more details.

Forcellini, Egidio. *Lexicon totius latinitatis*. Arnaldo Forni, Bologna, Italy, 1965.

Georges, Karl Ernst. *Ausführliches lateinisch-deutsches Handwörterbuch*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany, 1998. URL `http://www.zeno.org/Georges-1913`. Reprint of first edition of 1913–1918, Hannover, Germany: Hahnsche Buchhandlung.

Géraud, Hercule. Historique du Glossaire de la basse latinité de Du Cange. *Bibliothèque de l'École des chartes*, pages 498–510, 1839. doi: 10.3406/bec.1840.461649.

Glare, Peter G. W. *Oxford Latin Dictionary*. Oxford Languages. Oxford University Press, Oxford, UK, 2012. ISBN 978-0-19-958031-6.

Gradenwitz, Otto. *Laterculi vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas*. Hirzel, Leipzig, 1904.

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating NLP using Linked Data. In *Proc. 12th International Semantic Web Conference, 21-25 October 2013*, Sydney, Australia, 2013. doi: 10.1007/978-3-642-41338-4_7.

Klimek, Bettina, John McCrae, Julia Bosque-Gil, Maxim Ionov, James K. Tauber, and Christian Chiarcos. Challenges for the Representation of Morphology in Ontology Lexicons. In *Proceedings of eLex*, pages 570–591, 2019.

Kyjánek, Lukáš. Harmonisation of Language Resources for Word-Formation of Multiple Languages. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta, 2020.

Kyjánek, Lukáš, Olga Lyashevskaya, Anna Nedoluzhko, Daniil Vodolazsky, and Zdeněk Žabokrtskỳ. Constructing a Lexical Resource of Russian Derivational Morphology. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 2788–2797, 2022.

Lango, Mateusz, Zdeněk Žabokrtskỳ, and Magda Ševčíková. Semi-automatic construction of word-formation networks. *Language Resources and Evaluation*, 55(1):3–32, 2021. doi: 10.1007/s10579-019-09484-2.

Lassila, Ora and Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specification, 1998.

Lehmann, Christian. Latin preverbs and cases. In Pinkster, Harm, editor, *Latin linguistics and linguistic theory: Proceedings of the 1st International Colloquium on Latin Linguistics*, pages 145–161. John Benjamins, Amsterdam, 1983. doi: 10.1075/slcs.12.15leh.

Litta, Eleonora and Marco Budassi. What we talk about when we talk about paradigms: representing Latin word formation. In *Paradigmatic relations in word formation*, pages 128–163. Brill, 2020.

Litta, Eleonora and Marco Passarotti. (When) inflection needs derivation: a word formation lexicon for Latin. In Holmes, Nigel, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Lemmata Linguistica Latina. Volume 1. Words and Sounds*, pages 224–239. De Gruyter, Berlin, Boston, December 2019. ISBN 978-3-11-064758-7. doi: 10.1515/9783110647587-015.

Litta, Eleonora, Marco Passarotti, and Francesco Mambrini. Derivations and Connections: Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. *The Prague Bulletin of Mathematical Linguistics*, (115):163–186, 2020. doi: 10.14712/00326585.010.

McCrae, John, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex*, pages 587–597, 2017.

Passarotti, Marco, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31, 2017.

Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, LVIII(1):177–212, 2020.

Svoboda, Emil and Magda Ševčíková. Word Formation Analyzer for Czech: Automatic Parent Retrieval and Classification of Word Formation Processes. *The Prague Bulletin of Mathematical Linguistics*, (118):55–73, 2022. doi: 10.14712/00326585.019.

Address for correspondence:
Matteo Pellegrini
matteo.pellegrini@unicatt.it
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano, Italy

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 119 OCTOBER 2022 93–119

The MoNoPoli Database: Extragrammatical and Subverted Processes in French Words Based on Proper Names of Politicians

Mathilde Huguin

UMR 7118 ATILF, Université de Lorraine & CNRS – Nancy, France

Abstract

In this paper we present our method to build a derivational database of French deanthroponyms, which we call MoNoPoli for ***Mo****ts construits sur* ***No****ms propres de personnalités* ***Poli****tiques* ('complex words based on politician proper names'). MoNoPoli contains 6,545 complex words amounting to a total of 55,030 tokens and includes almost only neologistic forms. The Web is the only conceivable resource for collecting them: it alone gives massive access to discourse genres that contain neologisms. To feed the database, a program automatically generates the set of all possible derived words. Generated forms are then used as queries on the Web. Attested forms are kept with their context. This method provides a potential solution to collect data that cannot be found elsewhere. Finally, this article describes some of the remarkable results obtained with the analysis of the deanthroponyms of MoNoPoli. We show that the original nature of our data is reflected both in the use of new extragrammatical patterns (e.g., *Xistan* secretion pattern) and in the subversion of grammatical processes (e.g., foreign suffixation *Xix*).

1. Introduction

We study French deanthroponyms, i.e., words morphologically built on proper names (Schweickard, 1992; Leroy, 2008) that refer to contemporary political figures, henceforth PPN 'politician proper name'. These data have the particularity of being absent from most existing French corpora since they are neologisms and often exhibit the properties of nonce-formations. According to Bauer (1983) and Dal and Namer (2018), nonce-formations are words deemed to be new by their creators and used

 Corresponding author: mathilde.huguin@univ-lorraine.fr
Cite as: Mathilde Huguin. The MoNoPoli Database: Extragrammatical and Subverted Processes in French Words Based on Proper Names of Politicians. The Prague Bulletin of Mathematical Linguistics No. 119, 2022, pp. 93–119.
doi: 10.14712/00326585.024.

intentionally to meet an immediate need in a given context. In (1)[1] the forms *Macronite* ('Macron-itis'), *Balkanycratie* ('Balkany-cracy') and *hollandistan* ('holland-istan') are intentionally used by writers to express their aversion to the referents of the PPNs or their political ideas/actions.

(1) a. *Un nouveau cas de* **Macronite** *aiguë était signalé en France.* (*Emmanuel Macron*)
'A new case of acute Macron-itis was reported in France.'
b. *La* **Balkanycratie** *est très éloignée de la démocratie !* (*Patrick Balkany*)
'Balkany-cracy is very far from democracy!'
c. *Bienvenue en* **hollandistan**, *pays en sous-développement.* (*François Hollande*)
'Welcome to holland-istan, an underdeveloped country.'

As these complex words are most often neologistic nonce-formations, the constitution of the corpus required the elaboration of a specific methodology which we describe in this paper. This methodology consists of two steps. First, we automatically generate hypothetical deanthroponym *candidates* (e.g., *hollandien*, 'holland-ian', *lepen-phobe*, 'lepen-phobic') using the 89 PPNs and 90 suffixes of Huguin (2018, 2021).[2] We then look up these candidates in context, using the Web.

We also present some of the results we obtained focusing on five types of atypical constructions which characterize deanthroponyms. To present these atypical constructions we use the typology of the *Natural Morphology* (Dressler and Barbaresi, 1994; Dressler, 2000; Kilani-Schoch and Dressler, 2005). Our reasoning for choosing Kilani-Schoch and Dressler (2005)'s typology is that it takes into account recreational (i.e., extragrammaticality, marginality) processes which are preponderant in the deanthroponymic lexicon. Indeed, even if the notion of nonce-formation does not overlap with that of extragrammaticality, we will see that extragrammatical deanthroponyms are typically nonce-formations. Deanthroponyms often deviate from prototypical morphology to meet playful needs (see also Zwicky and Pullum, 1987).

Our paper is structured as follows. We present the method used to build MoNoPoli in Section 2. We present some of the results that emerged from the analysis of our data in Section 3. The morphological processes identified in MoNoPoli are spread over the whole scale of *Natural Morphology*: from the most regular French processes (e.g., suffixation) to the most irregular ones (e.g., blending). We show that the original nature of our data is reflected both in the use of new extragrammatical patterns and in the subversion of grammatical processes. Finally, the Section 4 contains the summary of this paper.

[1] Each example of deanthroponym is provided with its context; the base PPN is indicated between brackets (*First Name Last Name*) and the stem/suffix boundary is marked by a hyphen in the English translation.

[2] These lists are available online at `https://apps.atilf.fr/homepages/mhuguin/these/documents/`.

2. Methodology

This section presents each of the steps taken to build MoNoPoli. First, we explain why the Web is the most methodologically adequate resource to contain deanthroponyms (Section 2.1). Then, we discuss two types of possible strategies to collect them and evaluate their theoretical implications (Section 2.2). Finally, we present the architecture of the program that allowed their generation (Section 2.3) and the results of our collection (Section 2.4).

2.1. Where to find data

In order to know where to look for deanthroponyms, one must ask about their characteristics, including their degree of institutionalization (Bauer, 2000; Hohenhaus, 2005; Lipka, 2007) or their discursive function.

The list of 89 PPNs used to generate candidates contains names of politicians who have held a prominent position in French politics (Presidents, Ministers, leaders of political party, etc.) since 1981 (e.g., Jacques Chirac, Emmanuel Macron). The choice of this date is motivated by the need to limit the data to be studied, on the one hand, and, on the other, the willingness to work on contemporary data. The aim is to study words built on recent PPNs, which have not been impacted by time and whose meaning has not become opaque (Bauer, 2000). As we have selected names of current personalities, we expect that the words based on PPNs are recent creations, i.e., neologisms (Štekauer, 2002; Kerremans, 2015). The list of PPNs is also intended to be representative of the French political landscape at the time the study was launched. The number of personalities selected from each of the French political parties is proportional to the number of seats the parties have in the French National Assembly. As in the National Assembly, women are under-represented in our list (27% women, 73% men). Finally, the choice to limit ourselves to anthroponyms whose referents are French is pragmatic. This facilitates the interpretation of the words built on these names, which can be based on the actions of the personalities.

Given that these politicians make decisions that directly impact the French people, one can assume that deanthroponyms formed on their names will occur in puns, jokes, criticisms or claims. Hence we can except that the complex words we are going to find will occupy argumentative or humorous functions. Therefore they display the characteristics of nonce-formations (Hohenhaus, 2005, 2015) as in (2). In (2a), alongside their morphological creations *royaliste*$_{\text{N}}$ ('follower of Ségolène Royal') and *montebourgeois*$_{\text{N}}$ ('follower of Arnaud Montebourg'), the writer inserts a meta-discursive comment in brackets: "I don't know if that's how you say it". The deanthroponym *hollandophobe*$_{\text{Adj}}$ ('hollando-phobic') (2b) appears in a sequence that contains several terms of the same series (*Xphobe*$_{\text{Adj}}$), which Tanguy (2012, p. 104) calls suffixal outbursts. Comments and outbursts are among the structures that Dal and Namer (2018) have coined (meta)discursive and that often overlap with nonce-formations.

(2) a. *« Perdre la raison », un blog militant. Longtemps* **royaliste**, *maintenant* **montebourgeois** (*je ne sais pas si ça se dit comme ça*). (*Ségolène Royal*, *Arnaud Montebourg*)
"'Lose his mind", a militant blog. Long time royal-ist, now montebourg-ian (I don't know if that's how you say it).'
b. *Il y a de quoi devenir phobe :* **hollandephobe**, **vallsphobe**, **taubiraphobe**, **belkacemophobe**, *gauchophobe, antifaphobe, imamophobe.* (*François Hollande*, *Manuel Valls*, *Christiane Taubira*, *Najat Vallaud-Belkacem*)
'There is enough to come phobic: hollande-phobic, valls-phobic, taubira-phobic, belkacemo-phobic, lefto-phobic, antifa-phobic, imamo-phobic.'

Neologisms are more frequent in opinion genres than in information genres (Gérard, 2018). They indeed tend to be more frequently attested in less formal—or even satirical—contexts. In order to maximize our chances of obtaining them, we should look for resources where speakers/writers will be able to express themselves freely, and where they will be able to reach a wide audience. Social networks, forums and blogs, which are genres specific to the Web (Dal and Namer, 2015), provide such freedom and audience. To build up our corpus, we used the Web as a resource since it alone provides access to these discursive genres in real time.

Plénat et al. (2002), Lüdeling et al. (2007), Fradin et al. (2008) and Dal and Namer (2015) among others, have shown that the Web is useful for collecting contextualized lexical scarcities. Since search engines are constantly performing new indexing, they provide access not only to forms that have been recorded for a long time but also to very recent coinages. To automatically and massively explore the content of the Web, we used a Web scraping program[3] to query the Bing search engine. Our approach can be described as hypothetico-deductive (Tanguy, 2012, p. 101): we first generated a list of deanthroponym candidates and then searched for contexts containing the members of this list on the Web.

2.2. Possible strategies

The hypothetical deanthroponyms used as queries are built from PPNs (3a) by means of suffixes (3b). When generating candidates from such inputs, two strategies can be adopted. Each strategy has its theoretical implications.

(3) a. *Jacques Chirac*, *Christine Lagarde*, *Jean-Marie Le Pen*, *Emmanuel Macron*, *Jean-Luc Mélenchon*, *Nadine Morano*, *Nicolas Sarkozy*, *Christiane Taubira*, [...]
b. *able*, *ade*, *ais*, *al*, *ard*, *erie*, *esque*, *eur*, *ien*, *ification*, *ine*, *iser*, *isme*, *issime*, *iste* *isterie*, *itude*, *ix*, *logue*, *mètre*, *oïa*, *ose*, *ou*, *phage*, *phile*, *phobe*, *thon*, *us* [...]

[3]The program we use is provided by the company Data Observer. Data Observer (`www.data-observer.com`) is a start-up specialized in the collection, processing and analysis of textual data from the Web.

The first strategy, which we call *minimal* strategy, consists in generating only morphologically well-formed candidates. They respect a set of wellformedness morphophonological constraints (Roché and Plénat, 2014) as in *Optimality Theory* (Prince and Smolensky, 1993). For instance, this strategy leads to build exclusively *taubirie* /tobiʁi/ ('taubir-land') from the inputs *Taubira* /tobiʁa/ (from *Christiane Taubira*) and *-ie* /i/ so as to:

(i) avoid the hiatus /ai/ (/tobiʁa/ + /i/ = */tobiʁai/) proscribed by the markedness constraint *HIATUS (Plénat and Roché, 2001),

(ii) tend towards the trisyllabic optimal, required according to the size constraints (Plénat, 2009).

The objective of this method is to model the repair strategies instinctively implemented by speakers—and assumed by the linguist—to obtain an *optimal* derivative. This first strategy therefore assumes that speakers/writers always (unconsciously) apply the phonotactic constraints and/or that we are only interested in well-formed deanthroponyms.

With the *maximal* strategy, all possible forms are generated, regardless of their adequacy to wellformedness principles. This strategy corresponds to the hypothesis that a speaker/writer may ignore morphophonological constraints of wellformedness, especially in a situation of spontaneous written expression. For example, the sequence /ʁaʁi/ from /tobiʁaʁi/, which corresponds to the attested form *Taubirarie* ('taubirar-land') from (4a) entails that the derivative violates the constraints of faithfulness, size as well as the *Obligatory Contour Principle* (Goldsmith, 1976). Faced with such attested examples, we opt for the *maximal* strategy. Moreover, the hierarchy of phonological constraints is not known. We regularly observe several output variants of a construction process, as the derivatives of (4) attest. The variants in the output of a morphological construction process are due to the idiosyncratic ordering of constraints as shown by Roché (2010).

(4) a. *Mais où sommes-nous ? En France ? Ou* **Taubirarie** *?* (*Christiane Taubira*)
'But where are we? In France? Or Taubirar-land?'

b. *Vous vous foutez de qui en* **Taubirie** *?* (*Christiane Taubira*)
'You do not care who in Taubir-land?'

c. *Il risque très peu en* **Taubirasie***... no problemo.* (*Christiane Taubira*)
'He risks very little in Taubiras-land ... no problemo.'

In sum, we choose to generate as many forms as possible using PPN stems or variants thereof and a list of suffixes. (5) is an excerpt from the set of graphical forms obtained from the PPN *François Bayrou* and the French suffix *-able*.

(5) *françoisbayrouable*, *françoisbayroussable*, *françoisbayrousable*, *françoisbayroutable*, *françoisbayroulable*, *françoisbayroudable*, *françoisbayrouzable*, *françoisbayrounable*, *françoisbayrourable françoisbayrolable*, *bayrouable*, *bayroussable*, *bayrousable*,

bayroutable, *bayroulable*, *bayroudable*, *bayrouzable*, *bayrounable*, *bayrourable*, *bayrolable*, *françoisable*, *françoissable* [...]

In terms of costs and benefits, the maximal strategy produces much more noise than the minimal strategy. The higher the number of queries, the higher the noise. Nevertheless, the noise is an inconvenience that has a lesser impact than the dearth of results from the minimal strategy. Noisy results can be filtered out, whereas the lack of data cannot be compensated. In addition, this strategy allows us to collect unexpected formal creations, and, consequently, nonce-formations and extravagant formations that the minimal strategy does not allow us to obtain because it obeys morphological standards.

2.3. Generating derived forms

We run our candidate generation program on all the graphical forms that realize each of the 89 PPNs in our list and all 90 suffixes from our set (i.e., derivational suffixes or a verbal inflectional endings). PPNs are indeed realized in different forms, at least 3 (the *first name*, the *last name*, the *full name*), and up to 6, which we call sub-names and present in Table 1.

Sub-name	**Example**	**Derived form**	**Gloss**
Last name	Strauss-Kahn	*strausskahnité*	'strausskahn-ity'
First name	Dominique	*dominiqueur*	'dominiqu-er'
Full name	Dominique Strauss-Kahn	*dominiquestrauss--kahnien*	'dominiquestrauss kahn-ian'
Last name 1st part	Strauss	*straussophile*	'strausso-phile'
Last name 2nd part	Kahn	*kahnisation*	'kahn-ization'
Acronym	DSK	*dskie*	'dsk-land'

Table 1. Sub-names from the PPN Dominique Strauss-Kahn

The sub-names of a PPN are coreferential names that are used both autonomously in syntax and as bases in derivation. Unlike what happens with lexeme stems (Bonami et al., 2009), derivation rules do not impose constraints on sub-names, which is an additional argument for choosing the maximal strategy. We have shown (Huguin, 2018; Lignon et al., 2019) that the selection of a sub-name depends on sociolinguistic or extralinguistic conditionings such as the gender of the referent: e.g., the *firstname* is more often used if the referent is a woman (6).

(6) a. *Le Figaro, merci de défendre la langue française. On a déjà assez à faire pour lutter contre la* **najatisation** *de l'enseignement !* (*Najat Vallaud-Belkacem*)

'Le Figaro, thank you for defending the French language. We are already busy enough fighting against the najat-ization of education!'

b. *Le clientélisme et le* **clémentinisme** *se rejoignent.* (*Clémentine Autain*)
'Clientelism and clémentin-ism come together.'

The program inputs and outputs are sequences of characters. These graphical forms encode morphophonological phenomena as well as purely orthographic variations. The program generates all possible tuples formed by a ***stem*** of sub-name and a ***suffix***. For each tuple, the outputs of the program correspond to a set of possible derived words that we name ***Candidates***. Each ***Candidate*** is obtained by concatenating ($\oplus$) the form of a sub-name $\boldsymbol{Stem}_i^{n\prime}$ and a suffix $\boldsymbol{Suff}_j$ (7a) ($0 < \boldsymbol{j} \leq 90$). For a given PPN_i ($0 < \boldsymbol{i} \leq 89$), the symbol $\boldsymbol{Stem}_i^{n\prime}$ corresponds to the ***stem*** of one of its sub-names $\boldsymbol{n}$ ($0 < \boldsymbol{n} \leq 6$), or consists of a variation of this ***stem*** (7b) produced by one of the 36 $\mathfrak{R}$ rules of the program.

(7) a. $\boldsymbol{Candidate} = \boldsymbol{Stem}_i^{n\prime} \oplus \boldsymbol{Suff}_j$

b. $\begin{cases} \boldsymbol{Stem}_i^{n\prime} = \boldsymbol{Stem}_i^{n} \\ \boldsymbol{Stem}_i^{n\prime} = \mathfrak{R}(\boldsymbol{Stem}_i^{n\prime}) \end{cases}$

Each of them selects two arguments: $\boldsymbol{Stem}_i^{n\prime}$ and $\boldsymbol{Suff}_j$. Rules are organized in four blocks, cf. Figure 1. The rules in the block $\mathcal{B}1$ remove a graphical sequence from $\boldsymbol{Stem}_i^{n\prime}$. The rules of $\mathcal{B}2$ add a graphical sequence to $\boldsymbol{Stem}_i^{n\prime}$. The rules of $\mathcal{B}3$ operate graphical substitutions. When relevant, the rules are the graphical transcriptions of morphophonological rules: truncation for $\mathcal{B}1$, epenthesis for $\mathcal{B}2$ and allomorphy for $\mathcal{B}3$. Finally, the $\mathfrak{R}36$ rule of $\mathcal{B}4$ concatenates the inputs $\boldsymbol{Stem}_i^{n\prime}$ and $\boldsymbol{Suff}_j$.

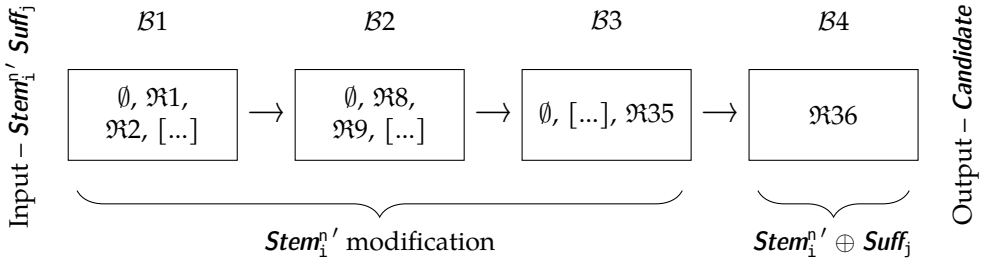


Figure 1. Rule combinations

The program is oriented and acyclic. In each block, the rules are in complementary distribution. The values of $\boldsymbol{Stem}_i^{n\prime}$ et $\boldsymbol{Suff}_j$ constrain which blocks ans which rules can be activated. This organization leads to 136 possible rule combinations. Each ***stem***/***suffix*** input explores the 136 combinations, but a ***Candidate*** is only produced when

the conditions of application of all the rules of the combination are met. Otherwise, the program tries the next combination.

Let us take the example of the input ***taubira*/*ique***.

- If we follow the first possible combination, we apply the null rule ($\emptyset$) in each of the blocks $\mathcal{B}1$, $\mathcal{B}2$ and $\mathcal{B}3$. Then rule $\mathfrak{R}36$ of concatenation in $\mathcal{B}4$ gives the ***Candidate*** (8).
- In contrast, the second combination using $\mathfrak{R}1$ in $\mathcal{B}1$ will be discarded by the program. $\mathfrak{R}1$ corresponds to deleting the final ***e*** of a ***stem***, hence it cannot be applied to ***taubira***. The next rule in $\mathcal{B}1$, i.e., $\mathfrak{R}2$, can be applied, since the input ***taubira*/*ique*** satisfies the conditions for application of the $\mathfrak{R}2$ truncation rule: ***taubira*** ends with a vowel and ***ique*** begins with a vowel. $\mathfrak{R}2$ deletes the vowel ***a*** at the end of ***stem***, to produce ***taubir*** (9a). Then, the null rules apply in $\mathcal{B}2$ and $\mathcal{B}3$. The output of $\mathfrak{R}2$ is given as input to $\mathfrak{R}36$ in $\mathcal{B}4$ to generate the ***Candidate*** (9b).
- Testing all rule combinations exhaustively will eventually produce all other candidates: e.g., with epenthesis (10) and (11).

(8) $\mathfrak{R}36$(***taubira, ique***) = ***taubiraique***

(9) a. $\mathfrak{R}2$(***taubira, ique***) = (***taubir, ique***)
b. $\mathfrak{R}36$(***taubir, ique***) = ***taubirique***

(10) $\mathfrak{R}8$(***taubira, ique***) = (***taubirat, ique***)
$\mathfrak{R}36$(***taubirat, ique***) = ***taubiratique***

(11) $\mathfrak{R}9$(***taubira, ique***) = (***taubiras, ique***)
$\mathfrak{R}36$(***taubiras, ique***) = ***taubirasique***

2.4. Data collection and annotation

The program produces 110,658 candidate forms, and each is used as a query, i.e., submitted to Bing. The set of attested deanthroponyms, their contexts, the URLs, and the number of pages associated with each query are saved in a tabulated file. A manual post-processing is then applied. It consists, for example, in deleting the entries of candidates homographs to attested lexemes with another meaning. For instance, *hollandais* ('holland-ese') is derived from *François Hollande* (12) but more often refers to the inhabitants of Holland.

(12) *Dans le cercle des* **hollandais**, *certains émettent l'hypothèse d'une absence du président sortant dans la course présidentielle.* (*François Hollande*)
'In the circle of holland-ese, some speculate that the outgoing president will not be in the presidential race.'

The database we obtain contains 6,545 different deanthroponyms, for a total of 55,030 tokens. This corpus contains 3,830 complex words whose formation mode were expected as they were explicitly generated by the program. But Bing's indexing process accidentally brought back a significant amount of unexpected forms: 40% of the deanthroponyms harvested are not part of our candidate list. For instance, we

obtained occurrences of the prefixed noun *anti alliot-marisme*$_{\text{N}}$ ('anti-alliot-marism') (13a) and of the compound adjective *chiraco-raffarinesque*$_{\text{Adj}}$ ('chiraco-raffarinian') (13b), looking for attestations of the candidates ***marisme*** and ***raffarinesque***.

(13) a. *Est-ce qu'une vague d'***anti alliot-marisme** *peut déferler sur la circonscription* [...] *?* (*Michèle Alliot-Marie*)
'Can a wave of anti-alliot-marism sweep through the riding [...]?'
b. *La majorité parlementaire* **chiraco-raffarinesque** *supprime une bonne partie des moyens financiers* [...]. (*Jacques Chirac*, *Jean-Pierre Raffarin*)
'The chiraco-raffarinian parliamentary majority suppresses a good part of the financial means [...].'

Each entry in the database describes an occurrence of one of the 6,545 deanthroponyms collected. This description is decomposed into a hundred or so features which describe, for each deanthroponym: its morphological properties such as its pattern (*Xade* for *peillonnade* in Table 2), its category and its morphophonological characteristics. Other properties do not result directly from the observation of the deanthroponyms but from our own analysis and are absent from Table 2. These include, for example, the semantic category of derivatives. The reader will be able to find the database and an explanation of each feature online via the Open Resources and TOols for LANGuage (ORTOLANG) platform.[4]

PPN	Derivative in context	Pattern	POS	Phonology
Vincent Peillon	*une nouvelle* **peillonnade** *: la rentrée en août !* 'a new peillon-ade: back to school in August!'	*Xade*	Nc	/pejonad/
Rama Yade	*Matignon redoute la* **ramayadisation** *de Vallaud-Belkacem* 'Matignon fears the ramayadization of Vallaud-Belkacem'	*Xisation*	Nc	/ʁamajadizasjɔ̃/
François Hollande	*le chimpanzé à cul rose* **homo hollandus** 'the pink ass chimpanzee homo holland-us'	*homoXus*	Nc	/omoolãdys/

Table 2. Exerpt of MoNoPoli

[4]The database is available at this address: `https://www.ortolang.fr/market/corpora/monopoli`. MoNoPoli is distributed under CC-by 4.0 license.

3. Analysis of deanthroponyms: some remarkable results

Unsurprisingly, the PPNs most often used as bases in MoNoPoli (14) correspond to the most prominent public figures. They have held a more important position (President *vs* Member of Parliament) or have been involved in high-profile events (laws, judicial scandals). The 13 PPNs in (14) are the bases of 50% of the deanthroponyms in our corpus. Jacques Chirac (15) was President of the French Republic. Dominique Strauss-Kahn (16) was implicated in scandals (sexual assault and rape). The PPNs of these referents are typically used to create nonce-formations since the referents are subject to controversy.

(14) *Dominique Strauss-Kahn, Marine Le Pen, Emmanuel Macron, Manuel Valls, Jean-Luc Mélenchon, François Mitterrand, Christiane Taubira, Ségolène Royal, François Bayrou, Lionel Jospin, François Hollande, Jacques Chirac, Nicolas Sarkozy*

(15) a. *On dit "arrête de* **chiraquer***" pour dire arrête de faire des bétises.* (*Jacques Chirac*)
'We say "stop chiraqu-ing" to say stop doing stupid things.'
b. *Heureux français, pensez à votre chance : vous n'avez rien à craindre du* **sarkozium***, du* **ségolénium***, du* **chiraquium***, mais méfiez-vous quand même du* **lepénium**. (*Nicolas Sarkozy*, *Ségolène Royal*, *Jacques Chirac*, *Jean-Marie Le Pen*)
'Happy French people, think about your luck: you have nothing to fear from sarkoz-ium, segolen-ium, chiraqu-ium, but beware of lepen-ium.'

(16) a. *Enfumages sans feux : après l'éruption mentale de* **viol-kahnisme** *sulfureux présumé, retour au volcanisme réel.* (*Dominique Strauss-Kahn*)
'Smoke and mirrors without fire: after the mental eruption of presumed sulphurous rape-kahnism, back to real volcanism.'
b. *Les socialos vont lancer le* **dskthon** *car faut aider cet homme qui aurait voulu diriger la France dans la plus grande hypocrisie.* (*Dominique Strauss-Kahn*)
'The socialos are going to launch the dsk-thon because it is necessary to help this man who would have wanted to lead France in the biggest hypocrisy.'

The nonce-formations are identifiable thanks to the meta-discursive signals and the discursive processes (see Section 2.1) but also sometimes thanks to the morphological processes used. 10% of the deanthroponyms of MoNoPoli are produced by extragrammatical processes (Kilani-Schoch and Dressler, 2005; Fradin et al., 2009). As a reminder, according to Fradin et al. (2009) a process belongs to the extragrammatical domain if:

(17) a. it is conscious;
b. it is located at the bottom of a scale of typological prototypicity of the processes (see (18) below);
c. it is not productive (see Dal and Namer, 2016);
d. it interacts with the different modules of grammar (e.g., syntax).

As Fradin (2003) points out, the reason these criteria are not precise is that extragrammatical morphology is defined in the negative: it is broadly what grammatical

morphology is not. Kilani-Schoch and Dressler (2005, p. 94) give seven criteria that define grammatical morphology:

(18) a. the rules of grammatical morphology manipulate form and meaning simultaneously;
b. they are applied in a regular and predictable way;
c. a morphological rule applies to a basic distinct class of bases (notably distinct by its part of speech);
d. the meaning change in rule operations is additional;
e. morphological rules produce new lexemes, distinct from their base;
f. the grammatical morphology is the object of an unconscious knowledge;
g. the grammatical morphology is productive.

We provide an overview of the morphological diversity of the content of MoNoPoli in Table 3.

Process & Frequency		**Example & Gloss**
Grammatical processes	90%	*hollandifier*$_{\text{V}}$ 'holland-ify'
Derivation	58%	*jospinerie*$_{\text{N}}$ 'jospin-ery'
Suffixation	51%	*chiraquiste*$_{\text{N}}$ 'chiraqu-ist'
Prefixation	5%	*ex-bovétiste*$_{\text{Adj}}$ 'ex-bovét-ist'
Conversion	2%	*bayrouer*$_{\text{V}}$ 'to-bayrou'
Compounding	32%	*lepénisme-mégretisme*$_{\text{N}}$'lepenism-megretism'
Extragrammatical processes	10%	*aubrython*$_{\text{N}}$'aubry-thon'

Table 3. Overview of the processes of MoNoPoli *according to Kilani-Schoch and Dressler (2005)'s typology*

The Table 3 leads to some comments concerning the methodology and the processes attested in MoNoPoli.

- The most frequent types of constructs are formed by suffixation which is expected since candidates are generated thanks to 90 suffixes (cf. Section 1 and Section 2.3).
- For the same reason, the compounds systematically contain at least one suffixed deanthroponym and MoNoPoli does not contain any deanthroponym with only a prefix, i.e., without the simultaneous presence of a derivational suffix (e.g, *-isme* in *anti-aubrisme* 'anti-aubr-ism') or a verbal inflectional ending (e.g., *-er* in *re-macroner* 'to re-macron').

- Derivation by conversion is the least represented,[5] which is certainly an effect of our methodology. Indeed, we only looked for forms corresponding to infinitive verbs (no noun or adjective).
- The rather small proportion of extragrammatical formations is also the result of the methodology used to form the deanthroponym candidates. Indeed, non-concatenative processes have little chance of being collected with our strategy.

In the following, we show that PPNs are privileged bases for original constructions on several levels. On the one hand, they serve as a bases for processes classified as extragrammatical in French (Section 3.1). On the other hand, they also sometimes lead to the subversion of more classical, i.e., more regular processes (Section 3.2).

3.1. Extragrammatical processes

MoNoPoli contains 662 deanthroponyms constructed by extragrammatical processes (see Table 3). More than half (54%) of the extragrammatical forms are hapaxic. This number confirms indirectly that they are original creations. The extragrammatical formations are indeed guided by a will to stand out from the crowd. These are often constructions whose *form* is atypical and rare, which indicates the occasional character of the deanthroponym. For instance, (19) contains a hapax legomenon: *taubirier*$_V$. It is a blend (Bat-El, 1996; Fradin, 2000; Fradin et al., 2009) built on the bases *Christiane Taubira* and *marier* ('to marry'). It means 'get married thanks to Christiane Taubira' (Christiane Taubira is the Minister of Justice who legalized gay marriage in France). This occasional character is often confirmed by the analysis of the context which presents a remarkable discursive structure (e.g., rhymes, chiasms or meta-discursive comments, see Section 2.1). For example, the two contexts of the blends in (20) are built on the model of a dictionary definition with a generic concept (*genus proximum*) and its specific features (*differentia specifica*).

(19) *Il a eu son heure de gloire en organisant un faux mariage d'homos il y a quelques années. Maintenant qu'Hollande et Taubira ont fait passer leur loi en force il peut en* **Taubirier** *à la pelle chaque semaine.* (*Christiane Taubira*)
'He had his moment of glory by organizing a fake gay wedding a few years ago. Now that Hollande and Taubira have rammed through their law, he can Taubirier a lot every week.'

(20) a. **Moranogastrique***: (n.c. masc) Dispositif médical visant à couper l'appétit. Dispositif lourd, le moranogastrique est proposé en dernier recours et ne saurait être utilisé comme simple coupe faim. Parmi ses nombreux effets secondaires, on note qu'il a souvent tendance à provoquer nausée et diarrhée.* (*Nadine Morano*)
'Moranogastric: (n.c. masc) Medical device designed to suppress appetite. A heavy device, the moranogastric is proposed as a last resort and should

[5]The database contains 128 converted verbs (1,045 tokens). 55 PPNs out of the 89 bases used give a converted verb.

not be used as a simple appetite suppressant. Among its many side effects, it often tends to cause nausea and diarrhea.'

b. **Duflotaïne***: Psychotrope hallucinogène entraînant une propension à renier ses convictions et des troubles du comportement.* (*Cécile Duflot*)
'Duflotaine: Hallucinogenic psychotropic drug leading to a propensity to deny one's convictions and behavioral disorders.'

In the following, we focus on two extragrammatical patterns present in MoNoPoli which have not been described in French. Section 3.1.1 is devoted to the pattern X*o*X*suff* exemplified in (21a) and Section 3.1.2 to the pattern X*istan* exemplified in (21b).

(21) a. *Imaginez que Lionel Jospin ait au second tour le projet de fondre toute la gauche plurielle dans un seul mouvement* **jospino-jospinien***. Si tel était son intention, je peux vous assurer d'entrée de jeu qu'il a perdu les élections.* (*Lionel Jospin*)
'Imagine that Lionel Jospin had a plan in the second round to merge the entire plural left into a single jospino-jospinian movement. If this was his intention, I can assure you from the outset that he lost the election.'

b. Nous regrettons aussi qu'il n'y ait pas eu de rubriques sur le **Mitterrandistan** ou le **Jospinistan**. (*François Mitterrand*, *Lionel Jospin*)
' We also regret that there were no sections on Mitterrand-istan or Jospinistan.'

3.1.1. Reduplication

MoNoPoli contains 23 deanthroponyms that instantiate the pattern X*o*X*suff* as complex words of (22) where X is the stem of a sub-name of PPN. In (22a) *suff* is *-iste* ('-ist'), in (22b) and (21a) it is *-ien* ('-ian').

(22) a. *Il s'explique dans un entretien à paraître ce mardi dans les éditions mayennaises d'Ouest-France : sur le plan départemental, c'est la motion la droite forte qui est arrivée largement en tête. Les militants ont choisi la ligne dure,* **sarkozo-sarkozyste**. (*Nicolas Sarkozy*)
'He explains himself in an interview to appear this Tuesday in the Mayenne editions of Ouest-France: on the departmental level, it is the motion of the hard right which arrived largely in head. The militants chose the hard line, sarkozo-sarkozist.'

b. *En témoigne le remaniement ministériel, qui fait la part plus que belle aux* **chiraco-chiraquiens** *: il s'apparente à la formation de la tortue, notent les bons observateurs de la vie politique.* (*Jacques Chirac*)
'This is evidenced by the ministerial reshuffle, which gives the lion's share to the chiraco-chiraquians: it is similar to the formation of the turtle, remark astute observers of the political sphere.'

At first sight, the process used to obtain these derivatives is compounding. One could see in *sarkozo-sarkozyste*$_{\text{Adj}}$ and *chiraco-chiraquien*$_{\text{N}}$ a particular case of the compound X*o*Y*suff*. For example, the demonym *franco-canadien*$_{\text{Adj/Nc}}$ ('French and Canadian') is compounded from the bases X *français*$_{\text{Adj/Nc}}$, which appears truncated, et Y *canadien*$_{\text{Adj/Nc}}$ (Dal and Amiot, 2008). As in *franco-canadien*$_{\text{Adj/Nc}}$, we find in *sarkozo-sarkyste* the intercalary vowel /o/ typically associated with compounds, and more specifically *learned* compounds, i.e., including the stem of a lexeme inherited from Greek or Latin. In *franco-canadien*$_{\text{Adj/Nc}}$, the vowel can be thought of as subverted, in that it does not mark the learned character of a stem but constitutes an iconic marker of compounding (Dal and Amiot, 2008).

Contrary to the compounds which are constituted of two distinct bases, in the deanthroponyms of (22), it is always the same (truncated) stem of X which is used in each element. This observation constitutes a first obstacle, of a formal nature, to the analysis of *sarkozo-sarkozyste*$_{\text{Adj}}$ and *chiraco-chiraquien*$_{\text{N}}$ as compounds. The second obstacle is semantic. The meanings of deanthroponyms of (22) do not correspond to coordination, subordination or apposition if we follow the tripartite classification of compounds of Scalise et al. (2005), for example. Semantically, *sarkozo-sarkozyste* (22a) means 'very/typically/exclusively sarkozist' and *chiraco-chiraquiens* (22b) are 'very/typically/exclusively chiraquian people'. This meaning consists of an amplification or an exaggeration of the property denoted by the base, and this semantic value is attested in cases of reduplication. The deanthroponyms of (22) are therefore not compounds but exhibit characteristics of reduplication.

This intensifying and restrictive polarization of reduplication is notably attested in syntax. In some syntactic reduplications, called *Identical Constituent Compounding* (Hohenhaus, 2004) or *Contrastive Reduplication* (Ghomeshi et al., 2004), the reduplicant allows a meta-discursive comment on the other term. In (23) the reduplication *mad mad* means 'very/completely mad'. When these are reduplicated nouns the associated paraphrase may be 'exclusively NOUN': e.g., *I want salad salad* means 'only/exclusively salad' (not *tuna salad* or *mixed salad*).

(23) She's mad [...] Not **mad mad**, but, you know. Out of control.
(Hohenhaus, 2007, p. 26)

These paraphrases have semantic values equivalent to those we have determined for the complex words of (22). In syntax, as in morphology, it is a matter of either intensifying a property or restricting the designated referential class to referents which possess prototypical properties of the class (Kleiber, 1990). Intensification and restriction are well known semantic values for reduplication (Moravcsik, 1978) and attested in many languages (e.g., English, Italian, French, Turkish). In French, reduplication is used exclusively for evaluative purposes. We therefore analyze the deanthroponyms X*o*X*suff* as resulting from reduplication; more specifically, this is partial pre-reduplication since the reduplicant X*o* is on the left and does not use the entire

phonological material of the base. Finally, we should add that the identified process is not specific to anthroponymic bases since it also applies to ethnic adjectives such as *français*$_{\text{Adj}}$ which gives *franco-français*$_{\text{Adj}}$ 'very/typically/exclusively French' in (24).

(24) *« Nuit debout » n'est plus un mouvement* **franco-français**.
"'Nuit debout" is no longer a Franco-French movement.'

3.1.2. Secretion

MoNoPoli contains 15 deanthroponyms that instantiate the pattern X*istan* as in (25). All of them are proper names of places where PPN referents hold power.

(25) a. *Pourquoi ne pas coller le nom de chacun des 12 départements pendant qu'ils y sont ?* **Wauquiezistan** *ça sonnait mieux non ?* (*Laurent Wauquiez*)
'Why not stick the name of each of the 12 departments while they are at it? Wauquiez-istan sounded better?'
b. *Le petit train-train de la honte dans la catégorie humeurs et gueule de bois en* **sarkozistan**. (*Nicolas Sarkozy*)
'The little train of shame in the category moods and hangovers in sarkozistan.'

We have not found any analysis of X*istan* in morphology works. The final *-istan* is also not listed in dictionaries of exponents like Cottez (1982). Some Internet users, however, consider *-istan* to be a suffix and have dedicated an entry to it in the *Wiktionary*,[6] which provides two types of information.

- According to the collaborative platform, the form comes from the suffix *-stan*, which designates a 'place' in Persian (*dari*). It is found in Asian toponyms like *Turkmenistan*$_{\text{Npr}}$ or *Kyrgyzstan*$_{\text{Npr}}$.
- The form *-istan*, with the vowel /i/, would be a meta-analysis of *Afghanistan* which the speakers interpret as 'the country of the Afghans' and thus split into /afgã/ + /istã/.

PPN derivatives in *-istan* are imaginary countries where despotism or corruption embodied by the PPN reigns. *Wauquiezistan*$_{\text{Npr}}$ (25a) is the 'place ruled by Laurent Wauquiez in an authoritarian and dishonest way' (i.e., the French region *Auvergne-Rhône-Alpes*). *Sarkozistan*$_{\text{Npr}}$ (25b) is the 'place ruled by Nicolas Sarkozy in an authoritarian and dishonest way' (i.e., France when it was led by Nicolas Sarkozy). If we examine the toponyms in *-istan* (*Afghanistan*$_{\text{Npr}}$, *Kurdistan*$_{\text{Npr}}$, *Pakistan*$_{\text{Npr}}$), we notice that the associated countries are often at war, non-democratic or considered by the Western press as corrupt. We can therefore establish a semantic correlation between the properties of the toponym referents and the (imaginary) referents X*istan* of our corpus. The context of the neologism *Zemmouristan*$_{\text{Npr}}$ based on Éric Zemmour,[7] cre-

[6] https://fr.wiktionary.org/wiki/-istan

[7] Éric Zemmour is an editorialist and politician of the extreme right.

ated by Jean-Luc Mélenchon[8] during a televised debate, illustrates our point. It is transcribed in (26). It contains the properties that Jean-Luc Mélenchon attaches to the country called *Zemmouristan*$_{\text{Npr}}$: "A country where women are demeaned, where there is the death penalty that you like, a country where homosexuals are punished and a country where we do not adhere to international conventions on refugees".

(26) *Pour le candidat LFI à la présidentielle, « le* **Zemmouristan**, *ça existe » : « Un pays où les femmes sont rabaissées, où il y a la peine de mort qui vous plaît, un pays où les homosexuels sont punis et un pays où on n'adhère pas aux conventions internationales sur les réfugiés, ça s'appelle l'Arabie Saoudite! ».* (*Éric Zemmour*)
'For the LFI presidential candidate, "the **Zemmouristan**, it exists": "A country where women are demeaned, where there is the death penalty that you like, a country where homosexuals are punished and a country where we do not adhere to international conventions on refugees, it is called Saudi Arabia!"'

Our hypothesis is that this is a new pattern of secretive suffixation. As a reminder (Fradin, 2000; Mattiello, 2007; Fradin, 2015; Mattiello, 2018), secretive suffixation is a process of segmentation change where a non-affixed part of a *model word* is reanalyzed as an affix: in this case *-istan*. The meaning related to the construction of complex words with the secreted affix contains only part of the meaning of the model word: in this case, the stereotype 'place ruled in an authoritarian and dishonest way'.

The originality of the X*istan* pattern is that it is impossible to determine a single model word. It can be hypothesized that the series of institutionalized toponyms is pressing to generate the form /istã/. *-istan* is thus a secreted element, but, unlike the secreted forms analyzed in the literature,[9] it is generated from a morphological series.

Finally, this secretion pattern is not specific to anthroponymic bases, nor to French. It applies to any participant that speakers/writers associate with one or more decisions that they consider hegemonic or totalitarian. It can be the person or company (27) that makes these decisions or the purpose to which it relates (28). For example, the documentary *Facebookistan*$_{\text{Npr}}$ (27) deals with the limits to freedom of expression imposed by the social network *Facebook*. More recently, the sanitary measures taken during the Coronavirus epidemic gave birth to *Hygiénistan*$_{\text{Npr}}$ ('hygien-istan') (28a) or *Vaccinistan*$_{\text{Npr}}$ (28b), particularly used in conspiracy discourses.

(27) **Facebookistan** *is a new documentary that takes a close look at Facebook, its laws, power and its influence on privacy and freedom of expression.*[10]

[8]Jean-Luc Mélenchon is a far left politician.

[9]For instance, Mattiello (2018) gives a list of related model words for each secreted form she identifies in English.

[10]`https://facebookistan.org/`

(28) a. *Yes, un nouveau néologisme de Jérôme Blanchet-Gravel. « Nous ne sommes plus au Québec mais en* **Hygiénistan***. ».*[11]
'Yes, a new neologism by Jérôme Blanchet-Gravel. "We are no longer in Quebec but in hygien-istan."'
b. **Vaccinistan***. Since the start of the coronavirus pandemic, anti-vax conspiracy theories have become dominant in the disinformation landscape.*[12]

3.2. Subverted grammatical processes

We have just seen that MoNoPoli contains original complex words and analyzed two extragrammatical patterns not described in French morphology. However, the originality of our data also lies in the subversion of processes that are traditionally considered as grammatical, or more regular, in French, such as suffixation and composition. Among these, we propose to examine three types of constructions exemplified in (29). In Section 3.2.1, we analyze what we call *foreign suffixation* (29a); in Section 3.2.2 we focus on disease nouns such as *macronite*$_{\text{N}}$ 'macron-itis' contextualized in (29b) and in Section 3.2.3 we discuss compounds such as *macrono-vallso-hollando-montebouresque*$_{\text{Adj}}$ contextualized in (29c).

(29) a. *N'oubliez pas* **Macronix***, sûrement plus dangereux pour notre pays que les 3 précédemment cités réunis.* (*Emmanuel Macron*)
'Don't forget Macron-ix, surely more dangerous for our country than the 3 previously mentioned together.'
b. *La* **macronite** *aigüe médiatique est lourde. Ce type est nul, un jeune déjà vieux, sans carrure, arrogant, méprisant le peuple !* (*Emmanuel Macron*)
'The media suffer of a severe case of acute macron-itis. This guy is garbage, a young man already old, without stature, arrogant, despises the people!'
c. *Les insoumis, eux, ont finalisé leur programme, alors que les autres, à droite, font semblant d'être en désaccord quand ils sont tous pareils, et à « gauche », fausse gauche* **macrono-vallso-hollando-montebouresque***, c'est le pataquès.* (*Emmanuel Macron*, *Manuel Valls*, *François Hollande*, *Arnaud Montebourg*)
'The insoumis [far left party] have finalized their program, while the others, on the right, pretend to disagree when they are all the same, and on the "left", false left macrono-vallso-hollando-montebour-esque, it's a jumble.'

[11] https://www.facebook.com/9936886137/posts/10158771088016138/

[12] https://eooh.eu/articles/plandemic

3.2.1. Foreign suffixation

MoNoPoli contains 339 deanthroponyms constructed from a process we call *foreign suffixation*. The deanthroponyms in (30) and (31) exhibit one or more inflectional or derivational exponent(s) inherited or borrowed from another language.

- The forms of (30) are derivatives in *-ix*, *-(i)us*, *-(i)um*. These are Latin exponents. According to Blancher (2015), they are used to evoke in the mind of the speakers/writers fantasy languages, by analogy with the lexicon of *The adventures of Asterix*.
- Deanthroponyms of (31) result from the addition of a suffix on both the *first name* and the *last name* of the same PPN. The suffixes used on the two sub-names may be identical as in *Jérômo Cahuzaco*$_{\text{Npr}}$ (31a) or not, as in *Nicolaï Sarkozine*$_{\text{Npr}}$ (31b). The anthroponym endings in *-o* recall anthroponyms of Italian or Spanish origin (*Livio*, *Diego*, *Pedro*). *-aï* and *-ine* are French transpositions of the final sequences of Russian anthroponyms (*Lénine* 'Lenin', *Staline* 'Stalin').

(30) a. **Alain juppix** *dénonce la « nullité du débat ». Celui de « nos ancêtres les gaulois », lancé par ce* **nicolas sarkozix**. (*Alain Juppé*, *Nicolas Sarkozy*)
'Alain jupp-ix denounces the "pointlessness of the debate". That of "our ancestors the Gauls", launched by this nicolas sarkoz-ix.'

b. **Jospinus** *est en exil,* **hollandus** *est chassé de chez lui.* (*Lionel Jospin*, *François Hollande*)
'Jospin-us is in exile, holland-us is driven from his home.'

c. *Olivier Desbordes a truffé le livret d'allusions superfines à notre politique actuelle "*__Balladurium__, __Mitterrandium__, __Chiracium__*, mysterium".* (*Édouard Balladur*, *François Mitterrand*, *Jacques Chirac*)
'Olivier Desbordes has filled the booklet with super-subtle allusions to our current politics "Balladur-ium, Mitterrand-ium, Chirac-ium, mysterium".'

(31) a. **Jérômo Cahuzaco** *mis en examen.* (*Jérôme Cahuzac*)
'Jérôm-o Cahuzac-o is indicted.'

b. *Je ne parlerai pas de monsieur* **Nicolaï Sarkozine** [...]. (*Nicolas Sarkozy*)
'I will not speak about Mr. Nicol-aï Sarkoz-ine [...].'

The derivatives of (30) are created by analogy with names of Gauls and Romans, as in the French comic strip *The adventures of Asterix* by A. Uderzo and R. Goscinny. Moreover, formally, these derivatives often mobilize long sub-names, i.e., *full names* as in (30a), following the example of what is done in the comic strip where the forms used as bases are often syntagmatic, thus also long (32).

(32) a. *ordre alphabétique* ('alphabetical order') > *Ordralphabetix*$_{\text{Npr}}$ (named *Unhygienix*$_{\text{Npr}}$ in English, according to Delesse, 2006)

b. *âge canonique* ('canonical age') > *Agecanonix*$_{\text{Npr}}$ (named *Geriatrix*$_{\text{Npr}}$ in English)

Blancher (2015), who analyzes the ludic mechanisms of *The adventures of Asterix*, calls these wordplays on anthroponyms *patronimi drôlati sistematici* (see also Kabatek, 2015). According to the author, R. Goscinny developed a model of anthroponymic game from the Gaulish *Vercingétorix*$_{\text{Npr}}$ from which he extracted the suffix form *-ix* /iks/. The writer then proceeded to systematize the use of the suffix for characters of Gaulish origin. He operated in the same way, but on other models, to generate the forms in *-us*, specific to Roman referents, in *-en* /ɛn/ (e.g., *Zøøvinsen*$_{\text{Npr}}$) for Vikings or *-is* /is/ (e.g., *Amonbofis*$_{\text{Npr}}$) for Egyptians. The choice of a suffix sequence thus evokes the ethnocultural belonging of the referent.

In our corpus, all PPN referents are French and contemporary. The suffix differences are therefore not related to the ethnicity of the referent. Rather, it is a matter of attributing stereotypes, or ethnic typing of the PPN referents, attributing to them properties (often assumed) emblematic of a people. For example, deanthroponyms in *-(i)us* and *-(i)um*, formed on the Latin model, sometimes have a warlike or conquering connotation as in (33a). The forms in *-ix* can accentuate the French side of the referent. This is the effect of meaning obtained in *montebourgix*$_{\text{Npr}}$ (33b) where the context refers to the fact that Arnaud Montebourg, former Minister of Economy, is known to have defended the relocalizing of French industry.

(33) a. *Mais très rapidement les coriaces de l'opposition abattent les champions de* **sarkosius**, *en dénichant des malversations qui virent au scandale et causent leur mort politique :* **blancus**, **joyandetus**, **estrosius** *sont ainsi rapidement mis hors combat.* (*Nicolas Sarkozy*, *Christian Blanc*, *Alain Joyandet*, *Christian Estrosi*)
'But very quickly fierce of the opposition slaughter the champions of sarkosius, by unearthing malpractices which turn to the scandal and cause their political death : blanc-us, joyandet-us, estros-ius are thus quickly put out of combat.'

b. *Il nous reste notre chef et président, le bien nommé Abraracourcix,* **airaultix** (*erotix*) *qui nous prépare des potions magiques aux effets aléatoires, et bien sûr* **montebourgix** *qui confond souvent les chefs d'entreprises gaulois avec des légionnaires romains.* (*Jean-Marc Ayrault*, *Arnaud Montebourg*)
'We still have our chief and president, the well named Abraracourcix, airaultix (erotix?) who prepares magic potions with random effects, and of course montebourg-ix who often confuses Gaulish company managers with Roman legionnaires.'

The method adopted by the authors of *The adventures of Asterix* to create anthroponymic wordplays is thus subverted by the speakers/writers to create nonce-formations based on PPN. These games indirectly allow speakers/writers to express their abjection towards the referents of the PPN or to ridicule them by attaching stereotypes to them. This process is not extragrammatical since it is indeed suffixation. However, the affixes are borrowings and the derivatives have an exclusively playful function.

This justifies the analysis of these derivatives as the result of subverted grammatical process.

3.2.2. Disease nouns

Another way for speakers/writers to express their views about PPN referents is to create medical-like names for diseases. MoNoPoli contains a total of 193 disease nouns that instantiate 6 different patterns shown in Table 4.

Pattern	**Type**	**Example**	**Gloss**
X*ine*	1	*Nicolas Sarkozy > sarkozine*	'sarkoz-ina'
X*pathie*	4	*Claude Guéant > guéantopathie*	'guéanto-pathy'
X*ide*	5	*Nicolas Sarkozy > sarkozide*	'sarkoz-id'
X*ose*	17	*Dominique De Villepin > villepinose*	'villepin-osis'
X*ite*	81	*Rama Yade > ramanite*	'raman-itis'
X*manie/mania*	85	*Dominique Strauss-Kahn > dskomanie*	'dsko-mania'

Table 4. Disease nouns of MoNoPoli

These nouns question the sanity of the PPN referent or its supporters: they are 'disease caused by the PPN referent' as in (34), or 'diseased passion for the PPN referent' as in (35).

(34) a. *C'est Nathalie Kosciusko-Morizet qui prendra donc le numérique où elle pourra soigner sa* **borlose** *allergique.* (*Jean-Louis Borloo*)
'Therefore Nathalie Kosciusko-Morizet will put in charge of the digital where she will be able to get her allergic borl-osis treated.'
b. *C'est pire que la grippe cette* **macronite** *aiguë, elle se répand dans toutes les rédactions.* (*Emmanuel Macron*)
'This acute macron-itis is worse than the flu, it's spreading to every newsroom.'
c. *Marrant cette nouvelle maladie qu'est la* **sarkozine**. (*Nicolas Sarkozy*)
'Funny this new disease that is the sarkoz-ina.'

(35) a. *L'ancien Premier ministre garde l'espoir : « Je ne peux pas remonter mon handicap en pleine* **ségolènomania***.». (Ségolène Royal*)
'The former Prime Minister remains hopeful: "I can't raise my handicap in the midst of ségolèno-mania."'.

The patterns in Table 4 are attested in the medical lexicon. In other words, the processes that manipulate these patterns are not limited to anthroponymic bases: X can be a common noun.

- For instance, *Xite* nouns are derivatives where the base specifically designates the site of an inflammation: e.g., a *bronchite*$_{\text{N}}$ ('bronchitis') is an 'inflammation of the bronchi'.
- The *Xose* derivatives also designate names of pathological processes where the base is the diseased organ: e.g., *dermatose* ('dermatosis') (Chebouti, 2014), but the base sometimes also designates the manifestation of the phenomenon (*furoncle*$_{\text{N}}$ 'furuncle' > *furonculose*$_{\text{N}}$ 'furunculosis'), or the origin of the disease (*bacille*$_{\text{N}}$ 'bacillus' > *bacillose*$_{\text{N}}$ 'bacillosis').

In anthroponym based games, these distinctions are irrelevant. While these exponents are used to distinguish between pathology names in medical terminology, they behave uniformly on PPNs. The learned exponents (mostly from Latin or Greek) are thus systematically subverted from their initial terminologizing function. A disease noun allows, metaphorically, to express an aversion to the PPN. This phenomenon of terminological despecialization has already been reported for other patterns handling learned exponents. Namer and Villoing (2015) and Lasserre (2016) have shown that the neoclassical exponents *-logue* and *-logie* are now added to any base X to designate—often ironically—the one who talks about X and about his so-called specialty (X*logie*).

3.2.3. Subverted compounding

MoNoPoli contains 1,925 deanthroponyms instantiating the general pattern X*o*-(X'(*o*)-)*Y*suff* where the brackets indicate the optionality of an element and the asterisk notes that the number of components at that position is 0, 1 or more, cf. (36).

(36) a. *Ainsi, selon une méthode éprouvée, le « camp du bien », pensant pouvoir l'achever, se livre en vain à une exégèse sémantique de sa critique du totem* **aubryo -strausskhanien**. (*Martine Aubry*, *Dominique Strauss-Kahn*)
'Thus, according to a tried and fruitlessly tested method, the "camp of the good", thinking to be able to finish it, engages in vain in a semantic exegesis of its criticism of the aubryo-strausskhanian totem.'

b. *Voilà le fruit de quinze années de* **pasquaïo-sarkozo-bessonisme**. (*Charles Pasqua*, *Nicolas Sarkozy*, *Éric Besson*)
'Here is the fruit of fifteen years of pasquaïo-sarkozo-bessonism.'

c. *C'est triste que le seul pendant à votre pensée unique* **bobo-marxo-stalino-taubiro-hollando-demissiono-comploto-lgbt-communiste**, *soit juste un propos « anti-système » d'extrême droite.* (*Christiane Taubira*, *François Hollande*)
'It's sad that the only counterpart to your boho-marxo-stalino-taubiro-hollando-resigno-conspiratoro-lgbt-communist unique thought, is just an extreme right-wing "anti-establishment" statement.'

The compound *aubryo-strausskhanien*$_{\text{Adj}}$ from (36a) has the minimal format of the pattern: X*o*-Y*suff*. It includes the adjective *strausskahnien*$_{\text{Adj}}$ ('strausskahn-ian') and the sub-names *Aubry* suffixed by /o/. It is the same /o/ that we have already ob-

served in Section 3.1. It is the typical vowel of learned compounding, subverted from its usual function since, here again, the stem is not inherited. This compound adjective is interpreted as a coordination: 'aubry-ist and strausskahn-ian'. We can see that the first component is a suffixed truncated adjectival form. The meaning of the compound guides our analysis: since the compound is coordinative and the coordinated elements are, by definition, of the same morpho-semantic type, we conclude that if Y*suff* is a denominative adjective, X is a denominative adjective like Y*suff*. So X is probably the truncated form of the relational adjective *aubryiste*$_{\text{Adj}}$ ('aubry-ist') of the sub-name *Aubry* (*aubryien*$_{\text{Adj}}$ 'aubry-ian' is attested with a lower frequency).

In (36b), we see that the minimal pattern can be extended to X*o*-X'*o*-Y*suff*. We can thus accumulate constituents in /o/. The compound is coordinative as in (36a). So, X*o* constituents are truncated forms of deanthroponyms of the same nature as Y*suff* *bessonisme*$_{\text{N}}$ ('besson-ism') which refers to the ideology of Eric Besson. They are truncated forms of the common nouns *pasquaïsme*$_{\text{N}}$ ('pasqua-ism') and *sarkozysme*$_{\text{N}}$ ('sarkoz-ism').

The examination of (36c), finally shows that what counts for the writer is above all the rhyme in /o/, since *bobo* 'boho' is not suffixed. Moreover, in *bobo-marxo-stalino-taubiro-hollando-demissiono-comploto-lgbt-communiste*$_{\text{Adj}}$, the accumulation of components to the left of the final suffixed component Y*suff* (*communiste*$_{\text{Adj}}$ 'communist') is not limited to a suffixed form in /o/. Indeed, one of the constituents is the acronym LGBT. In any case, all these forms always refer to adjectival properties, as does the last constituent *communiste*$_{\text{Adj}}$. The writer's goal is to accumulate as many constituents as possible, like an outburst, to distinguish himself.

These compounds are exclusively coordinative. Moreover, the more constituents the writer adds, the more the effect of meaning obtained is that of a cacophony. The longer the deanthroponym, the more original and remarkable it is. In conclusion, even if the compounding process is grammatical, compound deanthroponyms are not always grammatical (especially when they involve more than two bases). In our corpus, compounding is sometimes subverted in the benefit of the writer's argumentation or humor.

4. Conclusion

The method used to create the MoNoPoli database is replicable and adaptable to other languages or other inputs (bases or affixes). The database created is accessible online. It provides a large corpus of contextualized deanthroponyms, which to our knowledge does not exist in French.

MoNoPoli also displays a diverse and large panel of morphological processes. Its contents are relevant for further research on words based on anthroponyms or French nonce-formations. Its analysis reveals that deanthroponyms are often nonce-formations sometimes constructed by extragrammatical processes. We have also shown that grammatical processes can be subverted to satisfy the enunciative needs of the

writer. These needs demonstrate, at the same time, the writer's epilinguistic capacity to play with language.

To our knowledge, intensifying reduplication (*sarkozo-sarkoziste*), X*istan* pattern secretion, foreign suffixation (*sarkozix*), disease names subversion (*macronite*), and iterative composition (*bolcho-bayrouo-trotsko-villepiniste*) are all processes that have never been described in French. We have shown that these processes were not systematically specific to anthroponymic bases, nor to French (e.g., X*istan*). However, these new patterns seem specific to nonce-formations.

In the future, the content of MoNoPoli will be used as a basis for answering questions that we have not discussed in this paper.

- Do anthroponyms form a homogeneous category? For instance, do the names of authors, the names of journalists or collective anthroponyms have the same morphological functioning as proper names of politicians (e.g., are they the basis of the same processes)?
- Does the collection of deanthroponyms from other languages lead to the same results? The French are known to be dissenting, and are often portrayed as such in the international press. They express via morphological creation an appreciative dimension towards the referents of PPNs. Will another political culture translate into the language of its citizens by less expressive or less varied results?
- Do proper names form a homogeneous category? We could extend our study to other categories of proper names (e.g, toponyms, ergonyms). One of the objectives of this analysis would be to clarify whether the notion of sub-name is a specific feature of anthroponyms or whether other units possess similar dimensions (i.e., group together a set of syntactically autonomous and co-referential lexical units). This last question will feed the discussions about the definition of the nominal category and the units manipulated as bases in morphology.

Acknowledgements

We would like to thank the reviewers of the paper for their helpful comments. This research has been realized as part of the project Demonext (`https://www.demonext.xyz/`), supported by the Agence Nationale de la Recherche, grant number: ANR-17-CE23-0005.

Bibliography

Bat-El, Outi. Selecting the best of the worst : The grammar of Hebrew blends. *Phonology*, 13: 283–328, December 1996. doi: 10.1017/S0952675700002657.

Bauer, Laurie. *English Word-Formation*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, 1983. ISBN 978-0-521-28492-9. doi: 10.1017/CBO9781139165846.

Bauer, Laurie. System vs. norm: coinage and institutionalization. In Booij, Geert, Christian Lehmann, and Joachim Mugdan, editors, *Morphology: an International Handbook of*

Inflection and Word Formation, pages 832–840. De Gruyter, Berlin, New York, 2000. doi: 10.1515/9783110111286.1.11.832.

Blancher, Marc. « Ça est un bon mot ! » ou l'humour (icono-)textuel à la Goscinny. In Winter-Wroemel, Esme and Angelika Zirker, editors, *Enjeux du jeu de mots : Perspectives linguistiques et littéraires*, pages 273–290. De Gruyter, Berlin, München, Boston, October 2015. ISBN 978-3-11-040834-8.

Bonami, Olivier, Gilles Boyé, and Françoise Kerleroux. L'allomorphie radicale et la relation flexion-construction. In Fradin, Bernard, Françoise Kerleroux, and Marc Plénat, editors, *Aperçus de morphologie*, pages 103–126. Presses Universitaires de Vincennes, Saint-Denis, 2009.

Chebouti, Karim. *Le vocabulaire médical du point de vue des trois fonctions primaires.* Phd thesis, Université Paris 13, 2014.

Cottez, Henri. *Dictionnaire des structures du vocabulaire savant : éléments et modèles de formation*. Les Usuels du Robert. Le Robert, Paris, 1982. ISBN 978-2-85036-090-9.

Dal, Georgette and Dany Amiot. La composition néoclassique en français et l'ordre des constituants. In Amiot, Dany, editor, *La composition dans une perspective typologique*, pages 89–113. Artois Presses Université, Arras, 2008.

Dal, Georgette and Fiammetta Namer. Internet. In Müller, Peter O., Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation:An International Handbook of the Languages of Europe*, pages 2372–2386. De Gruyter Mouton, Berlin, New York, 2015.

Dal, Georgette and Fiammetta Namer. Productivity. In Stump, Gregory and Andrew Hippisley, editors, *The Cambridge Handbook of Morphology*. Cambridge University Press, Cambridge, 2016. doi: 10.1017/9781139814720.

Dal, Georgette and Fiammetta Namer. Playful nonce-formations in French: Creativity and productivity. In Arndt-Lappe, Sabine, Angelika Braun, Claudine Moulin, and Esme Winter-Froemel, editors, *Expanding the Lexicon Linguistic Innovation, Morphological Productivity, and Ludicity*, number 5 in The Dynamics of Wordplay, pages 203–228. De Gruyter, Berlin, Boston, 2018. doi: 10.1515/9783110501933-205.

Delesse, Catherine. Les noms propres dans la série Astérix et leur traduction anglaise. *Palimpsestes. Revue de traduction*, Hors série:297–315, September 2006. ISSN 1148-8158. doi: 10.4000/palimpsestes.1067.

Dressler, Wolfgang U. Extragrammatical vs marginal morphology. In Doleschal, Ursula and Anna Thornton, editors, *Marginal and Extragrammatical Morphology*, pages 1–10. Lincom Europa, Müchen, 2000.

Dressler, Wolfgang U. and Lavinia M. Barbaresi. *Morphopragmatics*. Trends in Linguistics. Studies and Monographs, 76. De Gruyter Mouton, reprint 2011 ed. edition, 1994. ISBN 978-3-11-014041-5. doi: 10.1075/hop.m2.mor1.

Fradin, Bernard. Combining forms, blends and related phenomena. In Doleschal, Ursula and Anna M. Thornton, editors, *Extragrammatical and Marginal Morphology*, pages 11–59. Lincom Europa, München, 2000.

Fradin, Bernard. *Nouvelles approches en morphologie*. Linguistique Nouvelle. Presses Universitaires de France, Paris, 2003. doi: 10.3917/puf.fradi.2003.01.

Fradin, Bernard. Blending. In Müller, Peter O., Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation – An International Handbook of the Languages of Europe*, volume Volume 1, page 386. De Gruyter Mouton, Berlin, New York, 2015.

Fradin, Bernard, Georgette Dal, Natalia Grabar, Stéphanie Lignon, Fiammetta Namer, Delphine Tribout, and Pierre Zweigenbaum. Remarques sur l'usage des corpus en morphologie. *Langages*, 171(3):34–59, 2008. ISSN 0458-726X. doi: 10.3917/lang.171.0034.

Fradin, Bernard, Fabio Montermini, and Marc Plénat. Morphologie grammaticale et extragrammaticale. In *Aperçus de morphologie du français*, Sciences du langage, pages 22–45. Presses Universitaires de Vincennes, Saint-Denis, 2009.

Ghomeshi, Jila, Ray Jackendoff, Nicole Rosen, and Kevin Russell. Contrastive Focus Reduplication in English (The Salad-Salad Paper). *Natural Language & Linguistic Theory*, 22(2): 307–357, 2004. ISSN 0167-806X. doi: 10.1023/B:NALA.0000015789.98638.f9.

Goldsmith, John. *Autosegmental phonology*. Phd thesis, MIT, 1976.

Gérard, Christophe. Variabilité du langage et productivité lexicale: Problèmes et propositions méthodologiques. *Neologica*, 12:23–45, 2018. doi: 10.15122/ISBN.978-2-406-08196-8.P.0023.

Hohenhaus, Peter. Identical Constituent Compounding – a Corpus-based Study. *Folia Linguistica*, 38(3-4):297–332, 2004. ISSN 1614-7308. doi: 10.1515/flin.2004.38.3-4.297.

Hohenhaus, Peter. Lexicalization and institutionalization. In Štekauer, Pavol and Rochelle Lieber, editors, *Handbook of Word-Formation*, Studies in Natural Language and Linguistic Theory, pages 353–370. Springer, Dordrecht, 2005. doi: 10.1007/1-4020-3596-9_15.

Hohenhaus, Peter. How to do (even more) things with nonce words (other than naming). In Munat, Judith, editor, *Lexical Creativity, Texts and Contexts*, Studies in Functional and Structural Linguistics 58, pages 15–38. John Benjamins Publishing Company, Amsterdam, Philadelphia, 2007. ISBN 978-90-272-1567-3. doi: 10.1075/sfsl.58.08hoh.

Hohenhaus, Peter. Anti-naming through non-word-formation. *SKASE Journal of Theorical Linguistics*, 12(3):272–291, 2015.

Huguin, Mathilde. Anthroponyms and paradigmatic derivation in French. *Lingue e Linguaggio. Defining paradigms in word formation*, XVII(2):217–232, 2018. doi: 10.1418/91866.

Huguin, Mathilde. *Analyse morphologique des mots construits sur base de noms de personnalités politiques*. Phd thesis, Université de Lorraine, December 2021.

Kabatek, Johannes. Wordplay and Discourse Traditions. In Zirker, Angelika and Esme Winter-Wroemel, editors, *Wordplay and Metalinguistic / Metadiscursive Reflection: Authors, Contexts, Techniques, and Meta-Reflection*, pages 213–228. De Gruyter, Berlin, München, Boston, October 2015. ISBN 978-3-11-040671-9. doi: 10.1515/9783110406719-010.

Kerremans, Daphné. *A Web of New Words: A Corpus-Based Study of the Conventionalization Process of English Neologisms*. English Corpus Linguistics. Peter Lang GmbH, Internationaler Verlag der Wissenschaften, 2015. ISBN 978-3-631-65578-8. doi: 10.1515/east-2016-0007.

Kilani-Schoch, Marianne and Wolfgang U. Dressler. *Morphologie naturelle et flexion du verbe français*. Gunter Narr Verlag, Tübingen, 2005.

Kleiber, Georges. *La Sémantique du prototype. Catégories et sens lexical*. Presses Universitaires de France, Paris, 1990.

Lasserre, Marine. *De l'intrusion d'un lexique allogène. L'exemple des éléments néoclassiques*. PhD thesis, Université Jean Jaurès, Toulouse, 2016.

Leroy, Sarah. Les noms propres et la dérivation suffixale. *Neuphilologische Mitteilungen*, 109: 55–71, 2008.

Lignon, Stéphanie, Fiammetta Namer, Nabil Hathout, and Mathilde Huguin. When sarkozysation leads to the hollandade, or the rejection of phonological well-formedness constraints by anthroponym-based derived words. In *International Symposium of Morphology (ISMo) 2019*, Paris, France, September 2019.

Lipka, Leonhard. Lexical creativity, textuality and problems of metalanguage. In Munat, Judith, editor, *Lexical Creativity, Texts and Contexts*, Studies in Functional and Structural Linguistics 58. John Benjamins Publishing Company, 1st edition, 2007. ISBN 978-90-272-1567-3. doi: 10.1075/sfsl.58.06lip.

Lüdeling, Anke, Stefan Evert, and Marco Baroni. Using web data for linguistic purposes. In Hundt, Marianne, Nadja Nesselhauf, and Carolin Biewer, editors, *Corpus Linguistics and the Web*, number 59 in Language and Computers: Studies in Practical Linguistics, pages 7–24. Rodopi, Amsterdam, New York, 2007.

Mattiello, Elisa. Combining forms and blends: The case of scape. In Jottini, Laura, Gabriella Del Lungo, and John Douthwaite, editors, *Cityscapes: Islands of the Self. Language Studies*, volume 2, pages 115–130, Cagliari, 2007. CUEC.

Mattiello, Elisa. Paradigmatic morphology splinters, combining forms, and secreted affixes. *SKASE Journal of Theoretical Linguistics*, 15(1):2–22, January 2018.

Moravcsik, Edith. Reduplicative Constructions. In Greenberg, Joseph H., editor, *Universals of Human Language*, number 3 in Word Structure, pages 297–334. Stanford University Press, Standford, 1978.

Namer, Fiammetta and Florence Villoing. Composition néoclassique en –logue et en –logiste : les noms en –logue sont ils encore des noms de spécialistes ? *Verbum*, 34(2):213–231, 2015.

Plénat, Marc. Les contraintes de taille. In Plenat, Marc, Bernard Fradin, and Françoise Kerleroux, editors, *Aperçus de morphologie du français*, Sciences du langage, pages 47–64. Presses Universitaires de Vincennes, Saint-Denis, 2009.

Plénat, Marc and Michel Roché. Prosodic constraints on suffixation in French. In Booij, Geert, Janet DeCesaris, Angela Ralli, and Sergio Scalise, editors, *Topics in Morphology. Selected Papers from the Third Mediterranean Morphology Meeting*, pages 285–299, Barcelona, September 2001. IULA-Universitat Pompeu Fabra (Barcelona).

Plénat, Marc, Stéphanie Lignon, Nicole Serna, and Ludovic Tanguy. La conjecture de Pichon. *Corpus et recherches linguistiques*, 1:105–150, 2002. doi: 10.4000/corpus.15.

Prince, Alan and Paul Smolensky. Optimality theory: constraint interaction in generative grammar. *Technical Report, Rutgers University Center for Cognitive Science and Computer Science Department*, 1993. doi: 10.1002/9780470756171.ch1.

Roché, Michel. Base, thème, radical. *Recherches linguistiques de Vincennes*, 39:95–134, December 2010. ISSN 0986-6124. doi: 10.4000/rlv.1850.

Roché, Michel and Marc Plénat. Le jeu des contraintes dans la sélection du thème présuffixal. In *SHS Web Conferences*, volume 8, pages 1863–1878. EDP Sciences, 2014. doi: 10.1051/shsconf/20140801143.

Scalise, Sergio, Antonietta Bisetto, and Emiliano Guevara. Selection in Compounding and Derivation. In Dressler, Wolfgang Ulrich, Dieter Kastovsky, Oskar E. Pfeiffer, and Franz Rainer, editors, *Morphology and its demarcations: Selected papers from the 11th Morphology meeting, Vienna, February 2004*, number 264 in Current Issues in Linguistic Theory, pages 133–150. John Benjamins Publishing Company, Amsterdam, Philadelphia, 2005. doi: 10.1075/cilt.264.09sca.

Schweickard, Wolfgang. *"Deonomastik". Ableitungen auf der Basis von Eigennamen im Französischen*. De Gruyter, Tübingen, 1992. doi: 10.1515/9783110933901.

Tanguy, Ludovic. *Complexification des données et des techniques en linguistique : contribution du TAL aux solutions et aux problèmes*. Habilitation à diriger des recherches, Université Toulouse-Le Mirail, Toulouse, 2012.

Zwicky, Arnold M. and Geoffrey K. Pullum. Plain morphology and expressive morphology. In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, pages 330–340, 1987. doi: 10.3765/bls.v13i0.1817.

Štekauer, Pavol. On the Theory of Neologisms and Nonce-formations. *Australian Journal of Linguistics*, 22(1):97–112, April 2002. ISSN 0726-8602. doi: 10.1080/07268600120122571.

Address for correspondence:
Mathilde Huguin
mathilde.huguin@univ-lorraine.fr
ATILF (CNRS & Université de Lorraine)
44, avenue de la Libération
B.P. 30687 54063 Nancy Cedex, France

The Prague Bulletin of Mathematical Linguistics
NUMBER 119 OCTOBER 2022

INSTRUCTIONS FOR AUTHORS

Manuscripts are welcome provided that they have not yet been published elsewhere and that they bring some interesting and new insights contributing to the broad field of computational linguistics in any of its aspects, or of linguistic theory. The submitted articles may be:

- long articles with completed, wide-impact research results both theoretical and practical, and/or new formalisms for linguistic analysis and their implementation and application on linguistic data sets, or
- short or long articles that are abstracts or extracts of Master's and PhD thesis, with the most interesting and/or promising results described. Also
- short or long articles looking forward that base their views on proper and deep analysis of the current situation in various subjects within the field are invited, as well as
- short articles about current advanced research of both theoretical and applied nature, with very specific (and perhaps narrow, but well-defined) target goal in all areas of language and speech processing, to give the opportunity to junior researchers to publish as soon as possible;
- short articles that contain contraversing, polemic or otherwise unusual views, supported by some experimental evidence but not necessarily evaluated in the usual sense are also welcome.

The recommended length of long article is 12–30 pages and of short paper is 6–15 pages.

The copyright of papers accepted for publication remains with the author. The editors reserve the right to make editorial revisions but these revisions and changes have to be approved by the author(s). Book reviews and short book notices are also appreciated.

The manuscripts are reviewed by 2 independent reviewers, at least one of them being a member of the international Editorial Board.

Authors receive a printed copy of the relevant issue of the PBML together with the original pdf files.

The guidelines for the technical shape of the contributions are found on the web site `https://ufal.mff.cuni.cz/pbml`. If there are any technical problems, please contact the editorial staff at `pbml@ufal.mff.cuni.cz`.