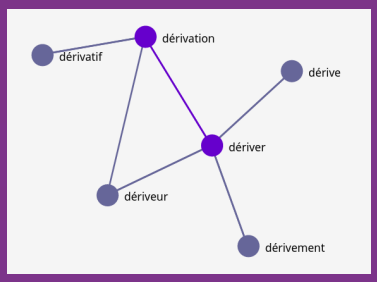
Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)

9–10 September 2021



Edited by:
Fiammetta Namer
Nabil Hathout
Stéphanie Lignon
Magda Ševčíková
Zdeněk Žabokrtský

Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)

Editors:
Fiammetta Namer
Nabil Hathout
Stéphanie Lignon
Magda Ševčíková
Zdeněk Žabokrtský

9–10 September 2021
ATILF & CLLE
Université de Lorraine
(Online conference)
http://nabil.hathout.free.fr/DeriMo2021

Published by:

ATILF (CNRS & UNIVERSITÉ DE LORRAINE)
44, avenue de la Libération
BP 30687
54036 Nancy Cedex
France

ISBN: 978-2-9580006-0-8

Program Committee Chairs

Fiammetta Namer	UMR 7118 ATILF CNRS & Université de Lorraine, Nancy
Stéphanie Lignon	UMR 7118 ATILF CNRS & Université de Lorraine, Nancy
Nabil Hathout	UMR 5263 CLLE CNRS & Université Toulouse Jean Jaurès, Toulouse
Magda Ševčíková	ÚFAL, Charles University, Prague
Zdeněk Žabokrtský	ÚFAL, Charles University, Prague

Program Committee

Olivier Bonami	(France)
Bruno Cartoni	(Switzerland)
Georgette Dal	(France)
Nicola Grandi	(Italy)
Pius ten Hacken	(Austria)
Richard Huyghe	(Switzerland)
Eleonora Litta	(Italy)
Silvia Luraghi	(Italy)
Francesco Mambrini	(Germany)
Rafael Marín	(France)
Claudia Marzi	(Italy)
Fabio Montermini	(France)
Sebastian Padó	(Germany)
Renáta Panocová	(Slovakia)
Marco Passarotti	(Italy)
Vito Pirrelli	(Italy)
Jan Radimský	(Czech Republic)
Franck Sajous	(France)
Andrew Spencer	(UK)
Pavol Štekauer	(Slovakia)
Pavel Štichauer	(Czech Republic)
Marko Tadic	(Croatia)
Delphine Tribout	(France)
Salvador Valera	(Spain)

Preface

This volume contains papers accepted for presentation at DeriMo 2021: The Third International Workshop on Resources and Tools for Derivational Morphology, held online on 9-10 September 2021. DeriMo 2021 follows up on the first and second DeriMo workshops (DeriMo 2017 and DeriMo 2019), which took place in Milan, Italy, in October 2017 and Prague, Czechia, in September 2019.

The submission and reviewing processes have been handled by the EasyChair system. In total, there were 20 submitted contributions, each reviewed by 3 program committee members. The proceedings contains 14 papers selected according to the reviews. In addition, the proceedings include contributions of two invited speakers, Sebastian Padó and Richard Huyghe. We thank the Demonext project (ANR-17-CE23-0005) for financial support for the workshop organization.

Contents

I Program **11**

Schedule **13**
Thursday, 9 September 2021 13
Friday, 10 September 2021 14

Acknowledgements **15**

II Papers **17**

Keynotes **19**
Sebastian Padó: *Building and exploiting resources for derivational morphology: Data-driven and theory-driven approaches* 20
Richard Huyghe: *Building a lexical database to investigate the semantics of French deverbal nouns* 21

Oral presentations **31**
Natalia Bobkova: *Statistical modelling of doublets in denominal adjective formation in Russian* 32
Olivier Bonami, Delphine Tribout: *Echantinom: a hand-annotated morphological lexicon of French nouns* 42
Guillaume Duboisdindien, Georgette Dal: *Critical analysis of clinical resources and tools for derivational morphology used in francophone speech and language therapy* . . . 52
Valeria Generalova: *Describing valence increasing constructions with XMG* 61
Nabil Hathout, Fiammetta Namer: *Adding Glawinette into Démonette: practical consequences and theoretical questions* 70
Mathilde Huguin: *The MoNoPoli database* 76
Marie-Laurence Knittel and Rafael Marín: *Developing a resource for -ance nouns, and related verbs and adjectives* 86
Lior Laks, Fiammetta Namer: *Designing a derivational resource for non-concatenative Morphology: the Hebrewnette database* 95
Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, Giovanni Moretti: *The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources* 105
Daniele Sanacore, Nabil Hathout, Fiammetta Namer: *Scenarios and frames in derivation: the case study of derivational families based on animal names* 114
Vanja Štefanec, Matea Filko, Krešimir Šojat: *Deriving the Graph: Using Affixal Senses for Building Semantic Graphs* 120
Emil Svoboda, Magda Ševčíková: *Splitting and Identifying Czech Compounds: A Pilot Study* 129
Jonáš Vidra, Zdeněk Žabokrtský: *Transferring Word-Formation Networks Between Languages* 139

Part I

Program

Schedule

Thursday, 9 September 2021

09:15 *Opening*

09:30 Sebastian Padó. *Building and exploiting resources for derivational morphology: Data-driven and theory-driven approaches*

10:30 *Break*

11:00 Mathilde Huguin. *The MoNoPoli database*

11:30 Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, Giovanni Moretti. *The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources*

12:00 Lior Laks, Fiammetta Namer. *Designing a derivational resource for non-concatenative Morphology: the Hebrewnette database*

12:30 *Break*

14:00 Natalia Bobkova. *Statistical modelling of doublets in denominal adjective formation in Russian*

14:30 Vanja Štefanec, Matea Filko, Krešimir Šojat. *Deriving the Graph: Using Affixal Senses for Building Semantic Graphs*

15:00 Valeria Generalova. *Describing valence increasing constructions with XMG*

16:00 Guillaume Duboisdindien, Georgette Dal. *Critical analysis of clinical resources and tools for derivational morphology used in francophone speech and language therapy*

16:30 Jonáš Vidra, Zdeněk Žabokrtský. *Transferring Word-Formation Networks Between Languages*

17:00 Daniele Sanacore, Nabil Hathout, Fiammetta Namer. *Scenarios and frames in derivation: the case study of derivational families based on animal names*

Friday, 10 September 2021

09:30 Richard Huyghe. *Building a lexical database to investigate the semantics of French deverbal nouns*

10:30 *Break*

11:00 Olivier Bonami, Delphine Tribout. *Echantinom: a hand-annotated morphological lexicon of French nouns*

11:30 Marie-Laurence Knittel and Rafael Marín. *Developing a resource for -ance nouns, and related verbs and adjectives*

12:00 Nabil Hathout, Fiammetta Namer. *Adding Glawinette into Démonette: practical consequences and theoretical questions*

12:30 *Break*

14:00 Daniil Vodolazsky, Hermann Petrov. *Compound Splitting and Analysis for Russian*

14:30 Emil Svoboda, Magda Ševčíková. *Splitting and Identifying Czech Compounds: A Pilot Study*

15:00 *Break*

15:15 *Business Meeting*

16:15 *Closing session*

Acknowledgements

DeriMo 2021 is organised by the Laboratoire ATILF (Université de Lorraine & CNRS) and the Laboratoire CLLE (Université Toulouse Jean Jaurès & CNRS). DeriMo 2021 is supported by the Demonext project (Agence Nationale de la Recherche – ANR-17-CE23-0005) and Université de Lorraine.

Part II

Papers

Keynotes

Building and exploiting resources for derivational morphology: Data-driven and theory-driven approaches

Sebastian Padó
(University of Stuttgart

A central characteristic of derivational morphology is its (semi)regularity, i.e., the existence of regular patterns which are however subject to exceptions, gaps and subregularities. In the first part of my talk, I will argue that the creation of derivational resources can profit from the combination of theory-driven and data-driven methods, and will present evidence for this claim from the construction of DErivBase, a derivational dictionary for German, which combines hand written rules, distributional data, and graph theoretic methods (Zeller et al., 2013, 2014; Papay et al., 2017). In the second part, I will move to the exploitation of such resources and discuss the tension between how the semantic effects of derivation are captured on the theoretical side (transparency, specificity) and how they are captured on the distributional side (Padó et al., 2016; Lapesa et al., 2017; Varvara et al., 2021).

References

Gabriella Lapesa, Sebastian Padó, Tillmann Pross, and Antje Rossdeutscher. 2017. Are doggies cuter than dogs? emotional valence and concreteness in german derivational morphology. In *Proceedings of IWCS*. Montpellier, France.

Sebastian Padó, Aurélie Herbelot, Max Kisselew, and Jan Šnajder. 2016. Predictability of distributional semantics in derivational word formation. In *Proceedings of COLING*. Osaka, Japan, pages 1285–1296.

Sean Papay, Gabriella Lapesa, and Sebastian Padó. 2017. Evaluating and improving a derivational lexicon with graph-theoretical methods. In *Proceedings of DeriMo*. Milan, Italy.

Rossella Varvara, Gabriella Lapesa, and Sebastian Padó. 2021. Grounding semantic transparency in context: A distributional semantic study on german event nominalizations. *Morphology* .

Britta D. Zeller, Padó Sebastian, and Jan Šnajder. 2014. Towards semantic validation of a derivational lexicon. In *Proceedings of COLING*. Dublin, Ireland, pages 1728–1739.

Britta D. Zeller, Jan Šnajder, and Padó Sebastian. 2013. Derivbase: Inducing and evaluating a derivational morphology resource for german. In *Proceedings of ACL*. Sofia, Bulgaria, pages 1201–1211.

Building a lexical database to investigate the semantics of French deverbal nouns

Richard Huyghe
University of Fribourg
richard.huyghe@unifr.ch

1 Introduction

Because of their grammatical complexity, formal variety and semantic diversity, deverbal nouns have challenged linguistic theory for more than half a century. Since Lees (1960) and Chomsky (1970), many studies have been devoted to the syntactic aspects of nominalization, especially with respect to argument realization (Grimshaw, 1990; Siloni, 1997; Alexiadou, 2001; Borer, 2003; Koptjevskaja-Tamm, 2003; Sleeman and Brito, 2010; Alexiadou et al., 2013; a.o.). Research about the morphosemantic properties of deverbal nouns has developed more recently (Booij, 1986; Gaeta, 2000; Namer and Villoing, 2008; Kawaletz and Plag, 2015; Fradin, 2016; Andreou and Petitjean, 2017; Wauquier et al., 2018; Varvara et al., 2021; a.o.), still leaving many questions unanswered. Extensive analyses of the semantic properties of deverbal nouns require large lexical resources that provide in-depth systemized information, possibly offering an overall picture of their organization in the lexicon. In this paper, I present the design of a database of French deverbal nouns created to answer research questions about derivational semantics, cross-categorial semantic preservation, and affix functionality. I introduce these issues in Section 2, and in Section 3 outline the methodology used to build and annotate a sample of French deverbal nouns. Section 4 provides an example of theoretical exploitation of the data through the examination of the aspectual properties of nouns ending in *-age*, *-ion* and *-ment*.

2 Morphosemantic issues

General issues about the semantics of derivation have been discussed by morphologists in the last decades (Corbin, 1987; Szymanek, 1988; Temple, 1995; Lieber, 2004; Rainer, 2014; Schulte, 2015; Plag et al., 2018; a.o.). Three major topics concerning the semantics of deverbal nouns are presented in this section.

2.1 Semantic diversity

It is well known that deverbal nouns can denote either eventualities or entities, in relation to the base verb meaning, but their semantic diversity needs to be further described. Detailed classifications of deverbal nouns have been proposed in the literature (see e.g. Fradin, 2012 for French nouns, and Lieber, 2016 for English nouns), but with questionable variations and lexical coverage. To ensure a broad application, the semantic analysis of deverbal nouns should be based on a general classification of nouns. It also requires a clear distinction between derivational semantics (i.e. the semantic operations associated to morphological processes) and lexical semantics (i.e. the semantics of conventionalized words in the lexicon). Derivational semantics is often underspecified with respect to lexical semantics, and it may be difficult to precisely determine the semantic outcome of morphological processes. Given that lexicalization idiosyncratically influences the meaning of lexemes, it is uncertain which semantic features of a given complex word result from derivation. For instance, French deverbal nouns *dortoir* 'dormitory' and *tuerie* 'slaughter' include a collective feature that is absent from the meaning of their base verbs—a plurality constraint applies to participants denoted by the external argument of *dormir* 'sleep' and the internal argument of *tuer* 'kill'. Whether the collective feature is implied by derivational semantics or not can only be determined through generalized observations of deverbal nouns suffixed with *-oir* and *-erie*.

2.2 Cross-categorial properties

The transfer of semantic properties through derivational processes raises a number of questions. It can be asked to what extent deverbal nouns inherit the semantic properties of base verbs, which properties are (not) preserved, and why. When deverbal nouns denote entities, the nominalization of verbal arguments can be discussed. For instance, *attaquant* 'attacker' nominalizes the agentive argument of *attaquer* 'attack', whereas *buvette* 'refreshment bar' as a locative nominalization does not correspond to any syntactic or semantic argument of *boire* 'drink'. When deverbal nouns denote eventualities, the preservation of lexical aspect and semantic role assignment can be investigated. It appears that the cross-categorial transfer of semantic properties is not always transparent (Haas et al., 2008; Balvet et al., 2011; Huyghe, 2015a), and it can be asked how frequently cases of non-preservation are observed. Generally speaking, cross-categorial features may depend on lexical class. For instance, the aspectual distinction between occurrential and non-occurrential actions in the nominal domain (e.g. *réunion* 'meeting' vs. *jardinage* 'gardening') does not have any equivalent in the verbal domain (Huyghe, 2011). Such discrepancies attest to differences in the semantic structure of verbs and nouns, which can affect the nature of cross-categorial semantic properties.

2.3 Affix polyfunctionality and rivalry

Many suffixes in French can form deverbal nouns, and their relation with nominal semantics calls for investigation. The general correspondence between affix selection and the meaning of deverbal nouns is known to be a many-to-many relation. For example, French deverbal nouns ending in *-ment* can denote events (*avortement* 'abortion'), states (*énervement* 'irritation'), agents (*gouvernement* 'governement'), instruments (*déguisement* 'costume'), locations (*logement* 'accommodation'), etc., whereas the instrument type can be denoted by nouns ending in *-ail* (*éventail* 'fan'), *-et* (*jouet* 'toy'), *-eur* (*aspirateur* 'vacuum cleaner'), *-oir* (*hachoir* 'mincer'), *-ure* (*couverture* 'blanket'), etc. These many-to-many relations need to be further explored.

On the one hand, detailed information should be provided about the possible semantic outputs of each suffix and their frequency. When a suffix is associated with distinct outputs, it should be determined whether these are primary or secondary outputs, because of the possible existence of polysemous nouns derived through metonymy (Ferret and Villoing, 2015). For instance, *-ion* in French seems to form collective agent nouns (*rébellion* 'rebellion', *rédaction* 'editorial board', *administration* 'administration'), but only for nouns which also have an event meaning, so that the existence of a deverbal pattern in *-ion* directly deriving agent nouns is uncertain. In the case of metonymic derivations, the formation of ambiguous nouns could still be seen as an indirect property of the suffix, if a given metonymic extension was only attested for some suffixes. The existence of complex derivational types could thus be hypothesized.

On the other hand, differences of semantic functionality between nominalizing suffixes should be examined to accurately evaluate their rivalry. The extent of suffix rivalry can vary according to (i) the existence of differences between similar semantic functions, (ii) the number of functions shared between polyfunctional suffixes, (iii) the actualization frequency of shared functions. First, it can be questioned whether suffixes with a similar function involve strictly identical constraints on lexical inputs and outputs, or tolerate some variation. Second, rivalry between polyfunctional affixes is usually partial and the number of functions involved in each case of rivalry should be investigated. Third, when suffixes compete for a given function, it can be asked if that function is equally frequently actualized for the different suffixes, both in terms of absolute frequency and of relative frequency (i.e. with respect to the other functions of the suffix). For example, *-ion* (*habitation* 'house') and *-erie* (*distillerie* 'distillery') apparently compete to derive locative nouns, but their degree of rivalry may be low if it appears that *-ion* as opposed to *-erie* rarely forms locative deverbal nouns.

3 Creation of a database

To answer theoretical questions about the semantic aspects of deverbal derivation, extensive data with detailed annotation are needed. Since the existing lexical databases for French do not provide the required

fine-grained semantic information, we intend to describe the semantic properties of a large sample of French deverbal nouns. This section presents the methodology used to build and analyze that sample.

3.1 Data sampling

The sample of deverbal nouns is based on candidates extracted from the FRCOW16A corpus, which is a large French web corpus containing 10.8 billion tokens (Schäfer, 2015; Schäfer and Bildhauer, 2012). Using a large web corpus allows for the inclusion in the sample of non-lexicalized words (nonce words and neologisms). The extracted candidates are verb-noun pairs (tagged with TreeTagger), in which the noun is formally related to the verb, possibly through regular allomorphy, in an apparently suffixed or converted form. Forty suffixes and 4 forms of conversion are considered in the extraction. Candidates are then manually filtered so that there is a semantic relation between at least one meaning of the verb and one meaning of the noun. Cases of double analyzability, in which a noun could be derived from a verb or from another word, are included provided that a deverbal morphological pattern is instantiated by at least two monosemous nouns. For instance, the existence of *causette* 'chat' and *ronflette* 'nap', univocally analyzable as derived from *causer* 'chat' and *ronfler* 'snore', attests to deverbal derivation of event nouns in *-ette*, and therefore ensures the analyzability of *grimpette* 'climb' (which could also be derived from the noun *grimpe* 'climbing') as possibly derived from the verb *grimper* 'climb'. Note that nouns in a relation of conversion with a verb but that do not include any verbal exponent are selected only if they denote eventualities. The sampling of the verb-noun pairs is performed in two stages. First, lists of words corresponding to weakly productive deverbal processes (e.g. suffixation in *-ade*, *-ail*, *-ard*, *-is*, *-ette*, conversion from verb stems in *-at*) are exhaustively filtered to optimize the possibility of quantitative generalization. Random selection across frequency ranges is done for the remaining types (e.g. suffixation in *-age*, *-eur*, *-ion*, *-ment*, *-ure*, conversion from participial verb forms), to finally obtain a sample of 4,000 verb-noun pairs.

3.2 Semantic description

Sampled verb-noun pairs are described with respect to nominal semantic type, verbal and nominal aspectual properties, and verbal and nominal capacity of assigning semantic roles. In order to account for the polysemy of nominalizations, the different meanings of each verb and noun are carefully distinguished and systematically paired. The semantic description is based on controlled manual annotation and explicit definitions of the annotated criteria. The general principles and linguistic tests used to analyze the semantic properties of both verbs and nouns are detailed in Salvadori et al. (2021a).

Ontological and relational properties are separated to appropriately describe nominal semantic types, and each deverbal noun is doubly classified. Ontological types relate to the nature of the referents, whereas relational types depend on the semantic relation with the base (Huyghe, 2015b). Existing semantic classifications of deverbal nouns often assimilate the two kinds of properties, which may lead to some confusion. As shown in examples (1)-(2), ontological and relational types are at least partially independent.

(1) *bâtir* 'build' → *bâtiment* 'building' [ARTIFACT-RESULT]
aspirer 'vacuum up' → *aspirateur* 'vacuum cleaner' [ARTIFACT-INSTRUMENT]
garer 'park' → *garage* 'garage' [ARTIFACT-LOCATION]

(2) *bâtir* 'build' → *bâtiment* 'building' [ARTIFACT-RESULT]
énerver 'irritate' → *énervement* 'irritation' [STATE-RESULT]
traduire 'translate' → *traduction* 'translation' [COGNITIVE OBJECT-RESULT]

Fourteen ontological simple types are distinguished based on distributional properties (Haas et al., submitted). Some of them combine to form complex types, in which case characteristic predicates of different simple types are contextually compatible (Copestake and Briscoe, 1995; Cruse, 1995; Pustejovsky, 1995; Kleiber, 1999; Asher, 2011; Dölling, 2021; a.o.). Relational types are based on the set of semantic roles used to annotate arguments, complemented with a transpositional type for nouns that

denote the same eventualities as their base verb. Seventeen semantic roles are defined and adapted from the sets of roles used in Verbnet (Kipper-Schuler, 2005) and Lirics (Petukhova and Bunt, 2008).

The lexical aspect of verbs and nouns is decomposed into four basic features (dynamicity, durativity, telicity, and post-phase). These properties are analyzed using linguistic tests mentioned in the literature (Vendler, 1967; Dowty, 1979; Rothstein, 2004; Haas et al., 2008; Filip, 2012; a.o.). Telicity is encoded by default with a delimited internal argument, and annotated as variable for degree achievements (Abusch, 1986; Bertinetto and Squartini, 1995; Hay et al., 1999; Rothstein, 2008; a.o.). Post-phase relates to the possibility of denoting a durative result state (Piñón, 1997, 1999; Apothéloz, 2008; Fradin, 2011; Haas and Jugnet, 2013), as in the case of *partir* 'leave' vs. *arriver* 'arrive' in (3).

(3) *Julie {est partie/?est arrivée} pendant deux jours.*
'Julie {left/arrived} for two days'

An important feature is that ambiguous nouns are assigned one entry per meaning in the database. Lexical ambiguity is identified through the variation of at least one annotated semantic property. Verbal and nominal lexemes are paired based on the principle of closest semantic correspondence: if a verb or a noun is ambiguous, the verbal and nominal lexemes that share the more aspectual and role-assigning properties are paired together.

4 A case study: the preservation of lexical aspect through nominalization

To test the methodology presented in the previous section, we analyzed a sample of 300 French deverbal neologisms ending in *-age*, *-ion* and *-ment*. The annotated sample can be used to investigate theoretical issues such as the preservation of verbal aspect in eventuality-denoting nominalizations (Salvadori et al., 2021b). The results of this investigation are discussed in this section.

4.1 The Aspect Preservation Hypothesis

It is often implicitly assumed that eventuality-denoting nominalizations inherit the lexical aspect of their bases. The idea of a cross-categorial transfer of aspect has been explicitly formulated by Fábregas et al. (2012) as the Aspect Preservation Hypothesis (APH), which stipulates that "the lexical aspect of a verb is preserved under nominalization if the resulting nominal denotes an eventuality". However, extended corpus studies have shown that nominalizations could differ from their bases with respect to lexical aspect (Balvet et al., 2011). For instance, *imagination* 'imagination' contrasts with *imaginer* 'imagine' in that in does not denote a dynamic eventuality. The lexical aspect of the base verb is not inherited, unlike what is the case for pairs such as *inventer-invention* 'invent'-'invention', as can be seen in (4)-(6).

(4) *L'auteur a {imaginé/inventé} une nouvelle forme narrative.* [+DYN]
'The author imagined/invented a new narrative form'

(5) *Cette {invention/*imagination} a eu lieu au 20e siècle.* [+DYN]
'This {invention/imagination} occurred in the 20th century'

(6) *Cet enfant a beaucoup d'{imagination/*invention}.* [–DYN]
'This child has a great {imagination/invention}'

Nevertheless, such aspectual shifts could be caused by lexicalization, and not by derivation, which would not violate the APH. To control for the effects of lexicalization, the semantic properties of neologisms can be scrutinized (Corbin, 1987; Plag, 1999). French neologisms suffixed with *-age*, *-ion*, *-ment* are particularly well suited to explore aspect preservation, for these three suffixes are arguably the most productive ones to form eventuality-denoting nouns in French. *-Age*, *-ion* and *-ment* have received a fair amount of attention in recent years (Martin, 2010; Uth, 2010; Dal et al., 2018; Fradin, 2019; Missud and Villoing, 2020; Wauquier, 2020), but no consensus has yet emerged as to what their distinctive semantic (including aspectual) properties could be.

	-age	*-ion*	*-ment*	Average
Aspect preservation	**74.8**	94.8	84.8	85.5

Table 1: Preservation of aspectual properties between verbs and nouns (%)

4.2 Describing the aspectual properties of deverbal neologisms in *-age*, *-ion* and *-ment*

The analysis is based on 300 French deverbal neologisms in *-age*, *-ion* and *-ment* (100 nouns per suffix) randomly extracted from the FRCOW16A corpus, following the methodology described in the previous section. Candidate words were additionally filtered using the Lefff (Sagot, 2010) and Lexique (New et al., 2004) lexicons as exclusion lists, and an ultimate lexicographic control was made to ensure that candidate words were not lexicalized. The aspectual properties of the sampled verbs and nouns were analyzed with respect to the criteria presented in Section 3.2. Nouns and verbs were annotated in a double-blind process and adjudicated with the help of a third annotator. The semantic annotation was based on the occurrences in FRCOW16A, complemented with examples taken from the web. Inter-annotator agreement scores were calculated for the 10 annotated features using Cohen's kappa and prevalence-adjusted PABAK (Byrt et al., 1993). Overall, the scores show a substantial agreement. Observed agreement scores range from 0.78 (verb post-phase) to 0.98 (verb dynamicity) with an average of 0.86. Kappa scores range from 0.56 (verb dynamicity) to 0.85 (noun dynamicity) with an average of 0.72. PABAK scores range from 0.67 (verb post-phase) to 0.96 (verb dynamicity) with an average of 0.81.

4.3 Results

A total of 501 nominal meanings were identified in the dataset, out of which 449 were eventuality meanings. Thirty-five of these meanings were associated with polysemous nouns and equivocally analyzable as resulting from morphological derivation or metonymy. These were excluded by default, considering that metonymic meanings could bias the results. Finally, aspectual shifts were observed for 60 out of 414 nouns. It appears that the lexical aspect of the verb is often, but not always, preserved in eventuality-denoting neologisms ending in *-age*, *-ion*, *-ment*. The preservation rates per suffix are presented in Table 1.

Discrepancies vary with aspectual properties, as shown in Table 2, and some specific aspectual variations can be observed for each suffix. For instance, *-ment* is associated with dynamic eventualities becoming stative (e.g. *jubiler* 'jubilate' denotes an activity whereas *jubilement* 'jubilance' denotes a state), and *-age* is associated with eventualities dropping post-phase or punctual eventualities becoming durative (e.g. *sortir* 'take out' is punctual and includes a post-phase whereas *sortage* 'taking out' is durative and does not include a post-phase).

Neological and lexicalized nominalizations can be compared with respect to aspectual discrepancies. Based on equivalent aspectual categories, the comparison between our dataset and the lexicalized data from the Nomage resource (Balvet et al., 2011) does not show any significant effect of lexicalization on aspect (non-)preservation ($\chi^2(1, N = 1088) = 0.297, p = .585$). It can be concluded that lexical aspect is not necessarily preserved through nominalization as a derivational process. Aspectual properties are not always inherited from the verbal domain, but can develop in nominal semantic structures. Theoretical models of nominalization should therefore account for possible aspectual shifts between base and derivative.

5 Conclusion

The creation of a database containing detailed semantic information about deverbal nouns could help us better understand the semantic aspects of derivation, the structure of the lexicon, and the nature of lexical categories. Analyses combining qualitative and quantitative approaches can make a substantial contribution to the study of deverbal nouns, and of the relations between form and meaning in the lexicon. They should allow us to evaluate both the extent of lexical idiosyncrasies and the content of complex

	-age	*-ion*	*-ment*	Average
Dynamicity pres.	96.7	98.0	**88.4**	94.4
Durativity pres.	**91.0**	99.3	99.2	96.8
Telicity pres.	**98.4**	100.0	99.2	99.3
Post-phase pres.	**80.3**	96.7	95.5	91.4

Table 2: Preservation of aspectual values per feature between verbs and nouns (%)

lexical regularities. The results obtained and the methodology developed will be exploitable for the comparative study of deverbal nouns in different languages. They may also feed research in related fields, such as computational linguistics, psycholinguistics and philosophy of language, by providing elements for computational semantic analysis, investigations of the mental lexicon, and reflection on the ontology of abstract situations.

Acknowledgments

This research was supported by the Swiss National Science Foundation under grant 10001F_188782 ('The semantics of deverbal nouns in French').

References

Dorit Abusch. 1986. *Verbs of change, causation and time*. CSLI report, Stanford.

Artemis Alexiadou. 2001. *Functional Structure in Nominals. Nominalization and Ergativity*. John Benjamins Publishing, Amsterdam.

Artemis Alexiadou, Gianina Iordăchioaia, Mariángeles Cano, Fabienne Martin, and Florian Schäfer. 2013. The realization of external arguments in nominalizations. *The Journal of Comparative Germanic Linguistics* 16:73–95.

Marios Andreou and Simon Petitjean. 2017. Describing derivational polysemy with XMG. In Iris Eshkol-Taravella and Jean-Yves Antoine, editors, *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles, vol. 2*. ATALA, Orléans, France, pages 94–101. https://aclanthology.org/2017.jeptalnrecital-court.12.

Denis Apothéloz. 2008. *Entrer quelques instants* vs *arriver quelques instants*: le problème de la spécification de la durée de l'état résultant. *Verbum* 30:199–219.

Nicholas Asher. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge University Press. https://doi.org/10.1017/CBO9780511793936.

Antonio Balvet, Lucie Barque, Marie-Hélène Condette, Pauline Haas, Richard Huyghe, Rafael Marín, and Aurélie Merlo. 2011. La ressource Nomage: confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus. *TAL* 52(3):129–152.

Pier Marco Bertinetto and Mario Squartini. 1995. An attempt at defining the class of gradual completion verbs. In Pier Marco Bertinetto, Valentina Bianchi, James Higginbotham, and Mario Squartini, editors, *Temporal Reference Aspect and Actionality 1: Semantics and Syntactic Perspectives*, Rosenberg and Sellier, Torino, pages 11–26.

Geert Booij. 1986. Form and meaning in morphology: the case of Dutch agent nouns. *Linguistics* 24:503–517.

Hagit Borer. 2003. Exo-skeletal vs. endo-skeletal explanations: syntactic projections and the lexicon. In John Moore and Maria Polinski, editors, *The Nature of Explanation in Linguistic Theory*, CLSI Publications, Stanford, pages 31–67.

Ted Byrt, Janet Bishop, and John B Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46(5):423–429.

Ann Copestake and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of semantics* 12(1):15–67.

Danielle Corbin. 1987. *Morphologie dérivationnelle et structuration du lexique*. Max Niemeyer Verlag, Tübingen.

D. Alan Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In P. St Dizier and E. Viegas, editors, *Computational Lexical Semantics*, Cambridge University Press, Cambridge, pages 33–49.

Georgette Dal, Nabil Hathout, Stéphanie Lignon, Fiammetta Namer, and Ludovic Tanguy. 2018. Toile *versus* dictionnaires: Les nominalisations du français en *-age* et en *-ment*. In Franck Neveu, Bernard Harmegnies, Linda Hriba, and Sophie Prévost, editors, *Congrès Mondial de Linguistique Française – CMLF 2018*. Institut de Linguistique Française, Paris.

David R. Dowty. 1979. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. D. Reidel Publishing Company, Dordrecht.

Johannes Dölling. 2021. Systematic polysemy. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Zimmermann Thomas Ede, editors, *The Wiley Blackwell Companion to Semantics*, John Wiley Sons, Inc., Hoboken NJ. https://doi.org/https://doi.org/10.1002/9781118788516.sem099.

Karen Ferret and Florence Villoing. 2015. French N-*age* instrumentals: semantic properties of the base verb. *Morphology* 25:473–496.

Hana Filip. 2012. Lexical aspect. In Robert I. Binnick, editor, *The Oxford Handbook of Tense and Aspect*, Oxford University Press, Oxford.

Bernard Fradin. 2011. Remarks on state-denoting nominalizations. *Recherches Linguistiques de Vincennes* 40:73–99.

Bernard Fradin. 2012. Les nominalisations et la lecture 'moyen'. *Lexique* 20:125–152.

Bernard Fradin. 2016. L'interprétation des nominalisations en N-*age* et N-*ment* en français. In Franz Rainer, Michela Russo, and Fernando Sánchez Miret, editors, *Actes du XXVIIe Congrès International de Linguistique et Philologie Romanes*, ATILF, Nancy, pages 53–66.

Bernard Fradin. 2019. Competition in derivation: what can we learn from french doublets in *-age* and *-ment*? In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler, and Hans Christian Luschützky, editors, *Competition in Inflection and Word-Formation*, Springer, Berlin, pages 67–93.

Antonio Fábregas, Rafael Marín, and Louise McNally. 2012. From psych verbs to nouns. In Violeta Demonte and Louise McNally, editors, *Telicity, Change, and State: A Cross-Categorial View of Event Structure*, Oxford University Press, Oxford, pages 162–184.

Livio Gaeta. 2000. On the interaction between morphology and semantics: the Italian suffix *-ata*. *Acta Linguistica Hungarica* 47(1):205–229.

Jane Grimshaw. 1990. *Argument Structure*. MIT Press, Cambridge, Mass.

Pauline Haas, Lucie Barque, Richard Huyghe, and Delphine Tribout. submitted. Pour une classification sémantique des noms en français appuyée sur des tests linguistiques .

Pauline Haas, Richard Huyghe, and Rafael Marín. 2008. Du verbe au nom : calques et décalages aspectuels. In Jacques Durand, Benoît Habert, and Bernard Laks, editors, *Congrès Mondial de Linguistique Française – CMLF 2008*. Institut de Linguistique Française, Paris, pages 2051–2065.

Pauline Haas and Anne Jugnet. 2013. De l'existence des prédicats d'achèvements. *Lingvisticæ Investigationes* 36(1):56–89.

Jennifer Hay, Christopher Kennedy, and Beth Levin. 1999. Scalar structure underlies telicity in 'degree achievements'. In Tanya Matthews and Devon Strolovitch, editors, *Proceedings of SALT 9*, CLC Publications, Ithaca, NY, pages 127–144.

Richard Huyghe. 2011. (A)telicity and the mass-count distinction: the case of French activity nominalizations. *Recherches Linguistiques de Vincennes* 40:101–126.

Richard Huyghe. 2015a. Les nominalisations « d'achèvement graduel » en français. *Le Français Moderne* 83:18–33.

Richard Huyghe. 2015b. Les typologies nominales : présentation. *Langue Française* 185:5–27.

Lea Kawaletz and Ingo Plag. 2015. Predicting the semantics of English nominalizations: A frame-based analysis of *-ment* suffixation. In Laurie Bauer, Lívia Körtvélyessy, and Pavol Štekauer, editors, *Semantics of Complex Words*, Springer, Berlin, pages 289–319.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Georges Kleiber. 1999. *Problèmes de sémantique: la polysémie en question(s)*. Presses Universitaires du Septentrion, Villeneuve d'Ascq.

Maria Koptjevskaja-Tamm. 2003. Action nominal constructions in the languages of Europe. In Frans Plank, editor, *Noun Phrase Structure in the Languages of Europe*, Berlin, pages 723–759.

Rochelle Lieber. 2004. *Morphology and Lexical Semantics*. Cambridge University Press, Cambridge.

Rochelle Lieber. 2016. *English Nouns: The Ecology of Nominalization*. Cambridge University Press, Cambridge.

Fabienne Martin. 2010. The semantics of eventive suffixes in French. In Monika Rathert and Artemis Alexiadou, editors, *The Semantics of Nominalizations across Languages and Frameworks*, De Gruyter Mouton, Berlin, pages 109–141.

Alice Missud and Florence Villoing. 2020. The morphology of rival *-ion*, *-age* and *-ment* selected verbal bases. *Lexique* 26:29–52.

Fiammetta Namer and Florence Villoing. 2008. Interpréter les noms déverbaux : quelle relation avec la structure argumentale du verbe de base ? Le cas des noms en *-oir(e)*. In Jacques Durand, Benoît Habert, and Bernard Laks, editors, *Congrès Mondial de Linguistique Française – CMLF 2008*, Institut de Linguistique Française, Paris, pages 1539–1557.

Boris New, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. *Lexique 2*: a new French lexical database. *Behavior Research Methods, Instruments, & Computers* 36(3):516–524.

Volha Petukhova and Harry Bunt. 2008. LIRICS semantic role annotation: Design and evaluation of a set of data categories. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco.

Christopher Piñón. 1997. Achievements in an event semantics. In Aaron Lawson and Eun Cho, editors, *Proceedings of SALT 7*, CLC Publications, Ithaca, NY, pages 273–296.

Christopher Piñón. 1999. Durative adverbials for result states. In Sonya Bird, Andrew Carnie, Jason D. Haugen, and Peter Norquest, editors, *WCCFL18: Proceedings of the 18th West Coast Conference on Formal Linguistics*. Cascadilla Press.

Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English derivation*. Walter de Gruyter, Berlin.

Ingo Plag, Marios Andreou, and Lea Kawaletz. 2018. A frame-semantic approach to polysemy in affixation. In Olivier Bonami, Gilles Boyé, Georgette Dal, Hélène Giraudo, and Fiammetta Namer, editors, *The Lexeme in Descriptive and Theoretical Morphology*, Language Science Press, Berlin, pages 467–486. https://doi.org/http://doi.org/10.5281/zenodo.1407021.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Franz Rainer. 2014. Polysemy in derivation. In Rochelle Lieber and Pavel Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, Oxford University Press, Oxford.

Susan Rothstein. 2004. *Structuring Events: A Study in the Semantics of Lexical Aspect*. Blackwell Publishing, Oxford.

Susan Rothstein. 2008. Two puzzles for a theory of lexical aspect: semelfactives and degree achievements. In Johannes Dölling, Tatjana Heyde-Zybatow, and Martin Shäfer, editors, *Event Structures in Linguistic Form and Interpretation*, De Gruyter Mouton, Berlin, pages 175–198.

Benoît Sagot. 2010. The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Justine Salvadori, Lucie Barque, Pauline Haas, Richard Huyghe, Alizée Lombard, Sandra Schwab, Delphine Tribout, and Marine Wauquier. 2021a. The semantics of deverbal nouns in French: Annotation guide. https://github.com/semantics-deverbal-nouns/annotation-guide.

Justine Salvadori, Richard Huyghe, Alizée Lombard, and Sandra Schwab. 2021b. The preservation of lexical aspect in nominalization: Insights from competing neologisms in French. Paper presented at *JENom 9: The Ninth Workshop on Nominalizations*.

Marion Schulte. 2015. *The Semantics of Derivational Morphology: A Synchronic and Diachronic Investigation of the Suffixes -age and -ery in English*. Narr Verlag, Tübingen.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, and Andreas Witt, editors, *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*. UCREL, IDS, Lancaster. http://rolandschaefer.net/?p=749.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, pages 486–493. http://rolandschaefer.net/?p=70.

Tal Siloni. 1997. *Noun Phrases and Nominalizations: The Syntax of DPs*. Springer, Berlin.

Petra Sleeman and Ana Maria Brito. 2010. Aspect and argument structure of deverbal nominalizations: A split vP analysis. In Artemis Alexiadou and Monika Rathert, editors, *The Syntax of Nominalizations across Languages and Frameworks*, De Gruyter Mouton, pages 199–218. https://doi.org/doi:10.1515/9783110245875.199.

Bogdan Szymanek. 1988. *Categories and Categorization in Morphology*. Catholic University Press, Lublin.

Martine Temple. 1995. *Pour une sémantique des mots construits*. Presses Universitaires du Septentrion, Villeneuve d'Ascq.

Melanie Uth. 2010. The rivalry of French *-ment* and *-age* from a diachronic perspective. In Monika Rathert and Artemis Alexiadou, editors, *The Semantics of Nominalizations across Languages and Frameworks*, De Gruyter Mouton, Berlin, page 215–244.

Rossella Varvara, Gabriella Lapesa, and Sebastian Padó. 2021. Grounding semantic transparency in context. *Morphology* https://doi.org/10.1007/s11525-021-09382-w.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY.

Marine Wauquier. 2020. *Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels*. Ph.D. thesis, Université Toulouse Jean Jaurès.

Marine Wauquier, Cécile Fabre, and Nabil Hathout. 2018. Différenciation sémantique de dérivés morphologiques à l'aide de critères distributionnels. In Franck Neveu, Bernard Harmegnies, Linda Hriba, and Sophie Prévost, editors, *Congrès Mondial de Linguistique Française – CMLF 2018*. Institut de Linguistique Française, Paris.

Oral presentations

Statistical modelling of doublets in denominal adjective formation in Russian

Natalia Bobkova
CLLE, CNRS
University of Toulouse
natalia.bobkova@univ-tlse2.fr

Abstract

This paper presents a quantitative study of doublets in denominal adjective formation in Russian and aims at identifying the underlying phonological, morphological and semantic properties of base nouns which allow the choice of more than one suffix to form adjectives. First, we extracted doublets from National Corpus of Russian language, then we annotated the properties of base nouns, trained logistic regression models to learn patterns and, finally, analyzed characteristics of nouns which allow the combination with both rival affixes.

1 Introduction

The derivation of adjectives from nouns is a complex process in Russian morphology, as these lexemes display a great deal of variation in the range of suffixes employed. Consequently, they constitute a good testing ground for the study of the competition between rival derivational strategies for the same syntactic and semantic function (Lindsay and Aronoff, 2013; Aronoff, 2016; Bonami and Thuilier, 2018). As various strategies are employed to form adjectives from nouns, doublets (and even triplets) of adjectives formed on the same base with distinct suffixes exist.

The competition between adjectival suffixes is determined by a complex combination of phonological, morphological and semantic factors. In this paper we aim at modeling suffixal rivalry in the construction of denominal adjectives in Russian. In general, three approaches may be applied to address the problems of rivalry. The first one consists in studying non-ambiguous cases for each suffix in the data set and highlighting the emerging properties of base nouns that allow to tease apart the suffixes, making them mutually exclusive. The second approach aims at studying ambiguous cases, e.g. cases where the base noun allows more than one suffix to form adjectives. The third approach is hybrid and consists in combining the previous ones and in investigating the properties of base nouns that allow for multiple adjective-forming affixation as opposed to nouns that do not. This approach would allow to establish a comparison between nouns that do and do not allow for adjectival doublets. In the present paper we focus on the second approach and leave the others for distinct studies. The goal of this paper is thus to shed light on the properties on base nouns that are less restrictive for the choice of the suffixes.

The data on which our study is performed were extracted from the National corpus of Russian language. The data set is composed of doublets: cases where two adjectives are attached to one stem. As various suffixes are employed to form adjectives, we first explore to which extent each of the suffix can be statistically predicted. We use the following statistical tools: correlation coefficients (Cramer's V for categorical variables) and one-to-rest logistic regression. We then focus on the properties of base nouns that allow for the formation of adjectival doublets with a given pair of suffixes as opposed to nouns that allow for doublets with all the other pairs. Binomial logistic regression is used in this case.

2 Rivalry in denominal adjectival formations

There are various strategies to derive adjectives from nouns in Russian. Classical grammars such as Townsend (1975) or Švedova (1980), for instance, enumerate more than 25 suffixes, which have different

degrees of productivity. Three suffixes are identified as being productive in synchrony (Zemskaya, 2015; Hénault and Sakhno, 2015; Kustova, 2018): *-n-*, *-sk-* and *-Ov-* (capital *O* in both cases represents a vowel that may correspond, phonologically, to different surface forms, and orthographically to <o> or <e>). The suffixes in question can be considered as the three main adjectival suffixes (abstract entities, denoted in capital letters), while others may be interpreted as their extended variants, denoted in small letters (Bobkova and Montermini, 2019):

- -N-: *-n-*, *-Ovn-*, *-ičn-*, *-ivn-*, *-on(n)-*, *-en(n)-*, *-(e)stven(n)-*, *-ozn-*, *-al'n-*, *-onal'n-*, *-arn-*, *-in-*;
- -SK-: *-sk-*, *-esk-*, *-česk-*, *-ičesk-*, *-ističesk-*, *-ijsk-*, *-ansk-*, *-ensk-*, *-insk-*, *-istsk-*, *-Ovsk-*;
- -OV-: *-Ov-*.

In this paper we are interested in the rivalry between the following suffixes: *-n-*, *-sk-*, *-Ov-*, *-Ovsk-*, *-ičesk-*, *-ičn-*, *-esk-*, and, in particularly, in cases where two different suffixes can result in the coexistence of two adjectives. The choice of this particular set of suffixes and their extended variants is motivated by the fact that they constitute the most frequent cases of rivalry in our data set (cf. Section 3).

Recent developments in derivational morphology, cf. Hathout (2011); Plénat (2011); Roché (2011) among others, consider that various types of constraints (phonological, morphological, semantic, pragmatic, etc.) display a complex interaction, resulting in the choice of one of the rival suffixes, or in the emergence of doublets. As far as the Russian language is concerned, the doublets are commonly encountered in denominal adjective formation, along with triplets, however less numerous, as shown in Table 1.

Base noun	**Adj_1**	**Adj_2**	**Adj_3**	**Suffixes**
ZIMA 'winter'	ZIMOVOJ	ZIMNIJ		*-Ov-/-n-*
MUZEJ 'museum'	MUZEJNYJ	MUZEJSKIJ		*-n-/-sk-*
LONDON 'London'	LONDONOVSKIJ	LONDONSKIJ		*-Ovsk-/sk-*
ANEMIJA 'anemia'	ANEMIČESKIJ	ANEMIČNYJ		*-ičesk-/-ičn-*
DRUID 'druid'	DRUIDIČESKIJ	DRUIDSKIJ		*-ičesk-/-sk-*
LOGIKA 'logic'	LOGIČESKIJ	LOGIČNYJ		*-esk-/-n-*
BOEC 'fighter'	BOJCOVYJ	BOJCOVSKIJ		*-Ov-/-Ovsk-*
KON' 'horse'	KONEVOJ	KONNYJ	KONSKIJ	*-n-/-sk-/-Ov-*

Table 1: Doublets and triplets in Russian adjectival formation

The choice of one or the other of the suffixes is accounted for by scholars (Townsend, 1975; Švedova, 1980; Hénault and Sakhno, 2015) by purely phonological factors, semantic or lexico-morphological ones:

- *-n-* tends to form more qualitative adjectives, whereas *-sk-* is used to form more relational ones;
- *-Ov-* appears with inanimate base nouns, *-Ovsk-* choses to combine with animate ones;
- *-esk-* privileges nouns with stems ending with velars;
- *-ičesk-* appears in particular in lexemes of foreign origin, and consequently also with lexemes containing specific suffixes / combining forms (e.g. *-ija*, *-izm*, *-ik*, etc.).

However, little studies are devoted to the existence of doublets or triplets (Antipina, 2012), namely to the properties of base nouns which do not restrict the choice of one affix. The goal of this paper is to use statistical approaches to reveal the main properties of base nouns (constraints) which may allow the choice of more than one affix (for instance, exactly two suffixes).

3 Data

3.1 Corpus

To perform our analysis, we extracted adjectives from the National corpus of Russian language (https://ruscorpora.ru/), a corpus of modern Russian containing over 600 million words. This corpus is divided in several subcorpora:

- Main subcorpus: texts representing standard Russian. It can be subdivided into 3 parts, each of which has its distinguishing features: modern written texts (from the 1950s to the present day), a subcorpus of real-life Russian speech (recordings of oral speech from the same period), and early texts (from the middle of the 18th to the middle of the 20th centuries);
- Media subcorpus: articles from mass media between 1990 and the 2000s;
- Multimedia subcorpus: Russian movies between 1930 and 2000;
- Corpus of Spoken Russian: recordings of public and spontaneous spoken Russian and the transcripts of the Russian movies;
- Poetry subcorpus: covers the time frame between 1750 and the 1890s, but also includes some poets of the 20th century;
- Dialectal subcorpus: recordings of dialectal speech (presented in loosely standardized orthography) from different regions of Russia;
- Educational subcorpus: small disambiguated corpus adapted for the Russian educational program;
- Parallel text subcorpus: texts in Russian are complemented by their translations into different languages, and vice versa.

For the purpose of this study we are interested in standard Russian, written or spoken. Dialectal, as well as educational and parallel subcorpora were therefore ruled out. The adjectives thus come from five subcorpora: main, media, multimedia, oral and poetic.

3.2 Data collection

Having established the types of subcorpora we are interested in, we automatically extracted adjectives based on their derivational suffixes. The raw extraction resulted in more than 75 thousands of adjectives. We then automatically grouped adjectives derived from the same base noun; base nouns were automatically reconstructed as well. This operation generated a list of 1968 raw base nouns with at least two adjectives derived from each noun.

Manual verification followed and concerned the verification of the exact shape of the base noun and the correct assignment of all the adjectives which might be potentially formed on it. Manual cleaning resulted in suppression of false positives as well. Among false positives we encounter:

- proper nouns formed mainly with *-Ovsk-*, *-Ov* and *-sk-* suffixes: STANISLAVSKIJ, MENDELEEV, AJVAZOVSKIJ;
- forms of nouns corresponding to genitive plural with *-ov* as an inflectional suffix: dvor$_{\text{NOM}}$ 'yard' - dvorov$_{\text{GEN}}$;
- possessive adjectives with the suffix *-ov-*: DED 'grandpa' - DEDOV. Despite of the fact that they are denominal, these adjectives were also excluded from this study due to their morphological and semantic peculiarities.

Manual verification led us to a data set composed of 773 base nouns with 1593 derived adjectives (729 cases of doublets, 41 cases of triplets, 3 cases of quartets). The individual suffixes distribution is presented in Table 2, showing the most frequent suffixes used to form adjectives in this data set.

	-n-	***-Ov-***	***-sk-***	***-ičesk-***	***-Ovsk-***	***-ičn-***	***-esk-***
Frequencies	450	322	320	180	169	81	71

Table 2: Distribution of individual suffixes

The data in Table 2 suggest that the three main suffixes are the most frequent among doublets, while their extended variants are less numerous. Table 2 also shows some trends concerning suffixes that form doublets: extended variants of *-sk-* are more frequent than those of *-n-* (*-ičn-* is the only extended variant of *-n-* encountered here).

Since doublets represent 95% of the data, we kept only them for the present study. Triplets and quartets were excluded as statistical models can perform poorly due to the little data. Our final data set is thus composed of 773 base nouns and 1458 adjectives. The most common couples of suffixes that can combine with the same base and form doublets are displayed in Table 3.

Suffixes	**N of bases**
-Ov-/-n-	226
-n-/-sk-	100
-Ovsk-/sk-	75
-ičesk-/-ičn-	71
-ičesk-/-sk-	63
-esk-/-n-	51
-Ov-/-Ovsk-	41

Table 3: Distribution of doublets

As shown in Table 3, the three main suffixes enter in competition not only with other suffixes but with each other as well - these are the most frequent cases of rivalry. The exception is the rivalry between *-n-* and *-Ov-* which seem to privilege distinct nominal bases. Similarly, *-Ovsk-* competes with both *-Ov-* and *-sk-* but not with *-n-*: the doublets with *-Ovsk-/-n-* are not frequent.

3.3 Annotation

Since the competition between affixes is driven by a complex combination of factors, base nouns were annotated according to some of their properties.

Phonological properties include information about the following features:

- `LastP`: the last phoneme of the stem (`Lab`: labial, `Den`: dental, `Alv`: alveolar, `Vel`: velar or `Vow`: vowel);
- `SyllB`: the length of the base noun in syllables - the only continuous property in the dataset;
- Stress position is also taken into consideration:
  - `AccSyl`: from the phonological point of view: which syllable is stressed – `D`: ultimate, `Ad`: penultimate, etc (\zim'a \'winter', \'viʃnʲa \'cherry', \'raduga \'rainbow');
  - `AccPos`: from the morphological point of view: if the stress is positioned on `R`: the root of the base noun, or – if any – `S`: derivational or `F`: inflectional suffix (\'son \'dream', \mark'sizm \'marxism', \galav'a \'head').

Both the last phoneme of the stem and the length of base noun in syllables are highlighted as important in prediction of the suffix by Lignon (2010) and Bonami and Thuilier (2018) in French, by Lindsay and Aronoff (2013) in English. We complete the list of phonological properties with information on stress position since it is not fixed in Russian and may influence the choice of the suffix.

Morpho-phonological allomorphies typical of Russian inflection and derivation were annotated as well. They include such properties as:

- `VowAlt`: vowel / Ø alternation, binary property (DVOREC 'palace' - DVORCOVYJ);
- `ConsM`: consonant mutation, binary property (TVOROG 'cottage cheese' - TVOROŽNYJ).

Both vowel alternation and consonant mutation reflect diachronic processes in Russian and do not correspond to a synchronically productive phonological phenomenon (Kapatsinski, 2010; Sims, 2017; Timberlake, 2004).

Morphological properties include only one predictor :

- `InflCl`: the inflectional class of base nouns.We follow a canonical distinction between 3 inflectional classes ($\text{PAPA}_{\text{I.M}}$ 'dad', $\text{PESNJA}_{\text{I.F}}$ 'song'; $\text{STOL}_{\text{II.M}}$ 'table', $\text{DELO}_{\text{II.N}}$ 'business'; $\text{TEN'}_{\text{III.F}}$ 'shadow').

Semantic properties include the following features:

- Binary distinct properties of [±proper], [±human], [±animate], [±concrete], [±countable];
- `Anim`: animacy, or the combination of the properties listed above into five groups (Thuilier, 2012):
  - `AnimA`: proper human (PIFAGOR'Pithagoras');
  - `AnimB`: common human/animate (SOBAKA 'dog');
  - `AnimC`: common concrete (DOM 'house');
  - `AnimD`: proper non-human (AL'PY 'Alps');
  - `AnimE`: common abstract (SOJUZ 'alliance').

The choice of these properties was motivated by their presence in the literature as potential factors to distinguish between two rival affixes. We hypothesize that the same properties could be less restrictive for some couples of rival affixes and allow the combination of the base noun with both of them.

4 Results

4.1 Exploration

Before diving into the statistical analysis we investigate if the data contain strongly correlated features among the predictors. A multicollinearity problem arises when there are two or more features heavily correlated to each other. Multicollinearity does not really affect the quality of the logistic regression but can have an impact on the reliability of effects of individual predictors in the model. If some of the predictors overlap in their measures, their effects become indistinguishable.

To detect if the predictors in the data set are affected by multicolinearity we create dummy variables for non binary categorical data and use a Pearson correlation test. The results are shown in Table 4.

Table 4 keeps the features whose correlation to other features is ≥ 0.3. As it could have been anticipated, all the classes of animacy are highly correlated to their constituents ([±proper], [±human], [±animate], [±concrete], [±countable]). The multicolinearity analysis revealed other correlations. It is possible, to a certain extent, to derive some stress position values from the constituents of amimacy. Similarly, the shift in semantic properties of the base noun can be associated with changes in values of inflectional class.

To address the multicolinearity issue we proceed with a straightforward method of dropping highly correlated features: we only used animacy for further investigations, and remove its constituents. Phonological and morphological stress positions and inflectional class are kept as well. The final set contains thus quite independent features.

	Propre	**Concr**	**Compt**	**Anim**	**Human**
AnimA	**0.66**	0.13	-0.24	0.24	0.25
AnimB	-0.17	**0.51**	**0.46**	**0.92**	**0.87**
AnimC	-0.16	**0.46**	0.05	**-0.44**	**-0.41**
AnimD	**0.73**	0.13	-0.26	-0.11	-0.13
AnimE	-0.19	**-0.99**	**-0.31**	**-0.54**	**-0.51**
AccSylAad	-0.05	**-0.38**	**-0.30**	-0.26	-0.25
AccPosR	0.12	-0.13	-0.17	**-0.30**	**-0.33**
AccPosS	-0.10	0.21	0.19	**0.41**	**0.44**
InflCl1	-0.07	**-0.36**	-0.28	**-0.35**	**-0.33**
InflCl2	0.05	**0.37**	**0.31**	**0.37**	**0.35**

Table 4: Correlation coefficient for predictors, where $\rho \geq 0.3$

Next, we examine how strong the correlation between each predictor (independently) and the suffix is. We use Cramer's V test, the measure of correlation between two categorical variables based on Pearson's Chi squared statistics. Feature importance for every suffix choice is displayed in Table 5.

-Ov-/-n-	***-n-/-sk-***	***-Ovsk-/-sk-***	***-ičesk-/-ičn-***	***-ičesk-/-sk-***	***-esk-/-n-***	***-Ov-/-Ovsk-***
Anim (0.59)	Anim (0.24)	Anim (0.51)	SyllB (0.36)	AccPos (0.41)	AccSyl (0.49)	SyllB (0.19)
SyllB (0.35)	ConsM (0.24)	LastP (0.16)	LastP (0.28)	Anim (0.38)	LastP (0.49)	InflCl (0.13)
AccPos (0.26)	AccSyl (0.19)	InflCl (0.15)	Anim (0.27)	LastP (0.25)	InflCl (0.35)	LastP (0.10)
VowAlt (0.23)	LastP (0.16)	AccPos (0.14)	AccSyl (0.25)	AccSyl (0.14)	SyllB (0.24)	Anim (0.09)
InflCl (0.17)	AccPos (0.15)	SyllB (0.13)	InflCl (0.23)	InflCl (0.13)	Anim (0.22)	ConsM (0.08)
AccSyl (0.16)	SyllB (0.14)	ConsM (0.07)	ConsM (0.21)	SyllB (0.12)	AccPos (0.11)	AccSyl (0.08)
LastP (0.006)	InflCl (0.12)	AccSyl (0.05)	AccPos (0.20)	ConsM (0.10)	ConsM (0.11)	AccPos (0.06)
ConsM (0.04)	VowAlt (0.08)	VowAlt (0.04)	VowAlt (0.06)	VowAlt (0.05)	VowAlt (0.00)	VowAlt (0.00)

Table 5: Cramer's V for each suffix, from ρ max to ρ min

Animacy appears to be one of the features that are most correlated to the affix choice. It seems to be strongly correlated to the choice of both *-Ov-* and *-n-*, as well as of *-Ovsk-* and *-sk-*; to a lesser extent - of *-n-* and *-sk-*. The last phoneme of the stem appears to be related to the emergence of *-esk-* and *-n-*, *-ičesk-* and *-ičn-*, *-Ovsk-* and *-sk-*. Vowel-zero alternations, as well as consonant mutations, seem to be the least correlated to the choice of suffix. The strongest correlations between the properties of base noun and suffixes are observed for *-Ov-/-n-*, *-Ovsk-/-sk-* and *-esk-/-n-* rivalries. For both *-n-/-sk-* and *-Ov-/-Ovsk-* the correlations seem to be week. Lastly, a significant gap in correlation coefficient values is observed for the *-Ovsk-/-sk-* rivalry: after $\rho = 0.51$ for animacy it drops to 0.16 for the last phoneme of the stem.

While Cramer's V can provide some insights about the data, it only indicates to which extent each predictor correlates independently to the suffix. Cramer's V does not allow the visualisation of correlation coefficients when all the predictors act simultaneously. Moreover, the predictors in our data set are categorical (all but the length of base noun in syllables), and some of the predictors are non binary. For instance, animacy appears to be the most highly correlated feature to the suffix choice. However, at this stage it is not clear whether all the five constituents of animacy are equally relevant.

To address these issues and to go deeper into the investigation we proceed with a logistic regression. The choice of logistic regression is driven by an easy interpretability and visualization of its results. It also provides a fine grained analysis of predictors (entering into the constituents of categorical variables), its coefficients allow to establish a ranking of the most important predictors.

In what follows we use two types of logistic regression: first we assess to which extent the individual suffixes may be predicted, given that we face a multilabel classification problem. We use one-to-all

logistic regression for this purpose. Next, we transform multilabel classification problem into a binary classification and analyse to which extent a pair of competing suffixes may be predicted, as opposed to all other suffixes. Binomial logistic regression is used for this task.

4.2 Multilabel classification

To perform a statistical modelling of suffixal rivalry we use a multi-label logistic regression, in particular one-to-rest approach. This heuristic method allows the decomposition of one multilabel classification problem into multiple binary classification tasks. A set of binary classifiers is thus leveraged for a multiclass classification. Our data set suggests building 7 binary classifiers, since we take 7 suffixes for this study. Instead of being mutually inclusive the labels become mutually exclusive, since each classifier solves such problems as "- *n*- vs. all the rest", "-*sk*- vs. all the rest", etc.

To evaluate the performance of these models we used the exact match ratio. This strict metric indicates the percentage of samples that have all their labels classified correctly. The data set was randomly split into train and test sets, the results of our models for test set are shown in Table 6.

	-esk-	***-ičn-***	***-Ovsk-***	***-Ov-***	***-sk-***	***-ičn***	***-n-***
Exact match ratio	90	86	82	81	75	70	53

Table 6: Results of logistic regression model (Multilabel approach)

These results show that the suffix *-esk-* is highly predictable, which allows us to think that there might be distinct properties of base nouns that allow the combination with this suffix. As far as other suffixes are concerned, the accuracies of the models are also quite high, except for *-n-*, for which the prediction is just slightly better than pure chance. This may mean that suffix *-n-* can potentially combine with all types of base nouns regardless of their properties, and the restrictions here can be less specific. Another explanation is that the data set contains exactly two labels for the outcome of the classification, it therefore may be confusing for the classifier to correctly predict *-n-* suffix. To assess fully the classification of *-n-* and to verify the numbers given by one-to-rest approach, a data set with only one dependent variable is necessary. This study however lies beyond the scope of this paper.

4.3 Binary classification

The logistic regression model allows us to access the parameters of the model and to visualize their weights. We are particularly interested in properties of base nouns allowing the derivation of two adjectives with distinct suffixes. Binary classification of the pair of competing suffixes as opposed to all other suffixes provides some insights, in particular due to the p-value, a measure of statistical significance of independent variables (typically $\leq$0.05).

As far as the doublets with *-Ov-*/*-n-* suffixes are concerned, the model shows a number of properties that can allow both of these rival suffixes:

- nouns designating common concrete entities, p<0.001 (OGUREC 'cucumber' - OGURCOVYJ / OGUREČNYJ);
- nouns with stressed radical, p<0.027, not a derivational or inflectional affix (/smet'ana/ 'cream' - SMETANNYJ / SMETANOVYJ);
- nouns affected by vowel alternation seem to prefer these two suffixes to the others as well, p<0.049 (PEPEL 'ashes' - PEPLOVYJ / PEPEL'NYJ).

Both *-ičesk-* and *-ičn-* combine with base nouns possessing certain morphological and phonological properties:

- inflectional class II, p<0.034 (SINONIM 'synonym' - SINONIMIČNYJ / SINONIMIČESKIJ);

- stress on derivational affix of the base noun, p<0.050, namely *-izm* or *-ist*, as in CINIZM 'cynicism' - CINIČESKIJ / CINIČNYJ.

The acceptance of both *-esk-* and *-n-* suffixes seem to rely on the semantic and phonological properties of base nouns:

- these nouns are mostly common abstract, p<0.003 (MISTIKA 'mysticism' - MISTIČESKIJ / MISTIČNYJ);
- their stems end with velars, p<0.010, which, in turn, provokes a consonant mutation in derived adjectives, though this property is not considered to be a statistically significant factor (p<0.483); cf. ISTERIKA 'hysterics' - ISTERIČESKIJ / ISTERIČNYJ.

Morphological and phonological properties determine the choice of both *-n-* and *-sk-*:

- inflectional class II, p<0.028 (INVALID 'disabled person' - INVALIDNYJ / INVALIDSKIJ);
- the length of the stem in syllables, p<0.012 (ZRITEL' 'viewer' - ZRITEL'NYJ / ZRITEL'SKIJ)

The same properties are also significant for the choice of both *-ičesk-/-sk-*:

- inflectional class II, p<0.050 (VAMPIR 'vampire' - VAMPIRIČESKIJ / VAMPIRSKIJ);
- the length of the stem in syllables, p<0.043 (monosyllabic nouns and nouns with two syllables).

The combination of a base noun with both *-Ov-* and *-Ovsk-* seems to be driven by phonological and semantic properties:

- the length of the stem in syllables, p<0.000. Both of these affixes privilege short stems (of 1-2 syllables);
- two classes of animacy - common concrete nouns, p<0.019 (BOJEC 'fighter' - BOJCOVYJ / BOJCOVSKIJ) and common abstract nouns, p<0.001 (ONLAJN 'online' - ONLAJNOVYJ / ONLAJNOVSKIJ).

Lastly, morphological and semantic properties seem to allow both *-Ovsk-* and *-sk-*:

- inflectional class II, p<0.018 (BANKIR 'banker' - BANKIROVSKIJ / BANKIRSKIJ);
- two subsets of animacy: common human or common animate, p<0.001 (SULTAN 'sultan' - SULTANOVSKIJ / SULTANSKIJ), common abstract, p<0.000 (INTERNET 'internet' - INTERNETOVSKIJ / INTERNETSKIJ).

To evaluate the predictive power of all the binary classifiers we use accuracy metric, the results are shown in Table 7.

	-esk-/-n-	***-ičesk-/-sk-***	***-Ovsk-/-sk-***	***-Ov-/-Ovsk-***	***-ičesk-/-ičn-***	***-n-/-sk-***	***-Ov-/-n-***
Accuracy	95.79	94.52	93.84	92.47	91.78	87.67	81.50

Table 7: Results of logistic regression model (Binary classification)

The results of binary classification are globally superior to the results of one-to-rest classification, given in Table 6. Despite the fact that *-esk-* and *-n-* are individually predicted with the best and the worst results respectively, their combination has the best accuracy. As for other combinations with *-n-* (*-n-/-sk-* and *-Ov-/-n-*), they can be found in the end of the ranking, however, the accuracy of predictions remains high. The results of logistic regression are not always congruent with the observations based of Cramer's V analysis: each property of the base noun does not have the same weight for the suffix choice when taken separately from other parameters, or jointly - when all the predictors are taken into consideration.

5 Discussion

Our study based on statistical models gave us some insights in order to identify the properties of base nouns which may allow the choice of two rival affixes. The most recurrent are the semantic factor of animacy and the morphological factor of inflectional class. Animacy is relevant for *-Ov-/-n-*, *-esk-/-n-*, *-Ov-/-Ovsk-* and *-Ovsk-/-sk-* doublets. Inflectional class plays a role in the acceptance of both *-ičesk-/-ičn-*, *-n-/-sk-*, *-ičesk-/-sk-* and *-Ovsk-/-sk-* suffixes. Phonological factors are determinant to a lesser extent: the length of stem in syllables (namely for monosyllabic nouns) allows both *-n-/-sk-*, *-ičesk-/-sk-* and *-Ov-/-Ovsk-* suffixes.

The results, however, are based only on properties of base nouns, the discussion on doublets would be incomplete without a deeper investigation on the nature of these doublets. The data extracted from National Corpus of Russian Language allow us to include for further studies such properties of adjectives as their frequency and the type of subcorpus they appear in.

The frequency of doublets needs further investigation because of two factors. First, one of the doublets may have undergone phenomena of lexicalization and be formally or semantically opaque, whereas another one is more transparent, as in TRUDNYJ / TRUDOVOJ, both derived from TRUD 'labor', however the first adjective means 'difficult', and the second one - 'labor, or work related'. Moreover, different adjectivizing affixes can be used to derive adjectives which correspond to two distinct senses of the underlying noun; the semantic of the whole adjective in a couple formed with two rival suffixes needs to be assessed. Second, even if both doublets are semantically and formally transparent, one may be frequently and commonly used, whereas another one may be an hapax, reflecting the result of the creative use of morphological constructions by speakers (Dal and Namer, 2012), as in PRIZRAČNYJ / PRIZRAKOVYJ, both derived from PRIZRAK 'ghost' and both transparent, however the first one is attested with frequency 2724, the second one appears in the corpus only once.

The type of subcorpus the doublets appear in can shed a light on their linguistic specialization. For instance, there might be a difference between suffixes chosen in general and newspaper subcorpora and the poetic and oral subcorpora: ALMAZNYJ 'dimond' and NOVOSTNOJ 'news' are both attested in main subcorpus, whereas ALMAZOVYJ and NOVOSTEVOJ - in poetic and oral subcorpora respectively. Furthermore, the two adjectives attested in the general subcorpus are very frequent ones, their doublets in oral and poetic subcorpora are hapaxes. The correlation between the frequency and the type of subcorpora the adjectives appear in also needs further investigation.

References

Ol'ga Pavlovna Antipina. 2012. *Sopostavitel'nyj analiz paronimov russkogo i anglijskogo jazykov*. Ph.D. thesis, Bashkir State University.

Mark Aronoff. 2016. Competition and the lexicon. In Elia, Annibale, Iacobini, Claudio, and Voghera, Miriam, editors, *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della Società Linguistica Italiana*. Bulzoni, pages 39–52.

Natalia Bobkova and Fabio Montermini. 2019. Suffix rivalry in russian: what low frequency words tell us. In *Mediterranean Morphology Meetings*. volume 12, pages 1–17.

Olivier Bonami and Juliette Thuilier. 2018. A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*. *Word Structure* 11(2).

Georgette Dal and Fiammetta Namer. 2012. Faut-il brûler les dictionnaires? ou comment les ressources numériques ont révolutionné les recherches en morphologie. In *SHS Web of Conferences*. EDP Sciences, volume 1, pages 1261–1276.

Nabil Hathout. 2011. Une approche topologique de la construction des mots: propositions théoriques et application à la préfixation en anti. *Des unités morphologiques au lexique* pages 251–318.

Christine Hénault and Sergueï Sakhno. 2015. Çem supermarket-n-yj luçşe supermarket-sk-ogo? slovoobrazovatel'naja sinonimija v russkix ad"ektivnyj neologizmax po dannym interneta. *B. Tošovic, A. Wonisch. Wortbildung und Internet* .

Vsevolod Kapatsinski. 2010. Velar palatalization in russian and artificial grammar: Constraints on models of morphophonology. *Laboratory phonology* 1(2):361–393.

Galina Ivanovna Kustova. 2018. Prilagatel'nye. *Materialy k korpusnoj grammatike russkogo jazyka. Vyp.3. Časti reči i leksiko-grammatičeskie klassy* pages 40–107.

Stéphanie Lignon. 2010. –iser and –ifier suffixations in French: Verify data to verize hypotheses? In *Décembrettes 7*.

Mark Lindsay and Mark Aronoff. 2013. Natural selection in self-organizing morphological systems. In *Morphology in Toulouse. Selected Proceedings of Décembrettes 7 (Toulouse 2-3 December 2010)*. Lincom Europa, pages 133–153.

Marc Plénat. 2011. Enquête sur divers effets des contraintes dissimilatives en français. *M. Roché, G. Boyé, N. Hathout, S. Lignon & M. Plénat, Des unités morphologiques au lexique. Paris: Hermès-Lavoisier* pages 145–190.

Michel Roché. 2011. Quel traitement unifié pour les dérivations en-isme et en-iste?

Andrea D Sims. 2017. Slavic morphology: Recent approaches to classic problems, illustrated with Russian. *Journal of Slavic Linguistics* 25(2):489–524.

Juliette Thuilier. 2012. *Contraintes préférentielles et ordre des mots en français*. Ph.D. thesis, Université Paris-Diderot-Paris VII.

Alan Timberlake. 2004. *A reference grammar of Russian*. Cambridge University Press.

Charles Edward Townsend. 1975. *Russian word-formation*. Slavica Publishers.

Natal'ja Švedova. 1980. *Russkaja grammatika*, volume 1. Moskva: Nauka.

Elena Andreevna Zemskaya. 2015. *Jazyk kak dejatel'nost'. Morfema, slovo, reč*. Moskva: Flinta.

Échantinom: a hand-annotated morphological lexicon of French nouns

Olivier Bonami
Université de Paris, LLF, CNRS
olivier.bonami@u-paris.fr

Delphine Tribout
Université de Lille, STL, CNRS
delphine.tribout@univ-lille.fr

Abstract

We present *Échantinom*, a new morphological resource for French nouns based on random sampling of frequent lexemes. The resource documents $5,000$ items in terms of their morphological type at two levels of granularity, as well as, for suffixed nouns, the exact identity of the base and process, and the formal and semantic transparency of the relationship between base and derivative. We outline the motivations for the development of such a resource, the sampling method, main annotation decisions, and provide some preliminary descriptive statistics.

1 Motivation

The very existence of the DeriMo workshop series testifies to a renewed interest in the development of large scale resources for derivational morphology. Table 1 lists most of the resources available for French, focussing on freely available machine-readable ressources developed in the last 15 years. This collection of resources provides a very rich view of the French word formation system; and an integration of those resources into a coherent unified database is the main goal of the ongoing Démonext project (Namer et al., 2019).

Resource	Publication	Processes
Démonette	Hathout and Namer (2014)	Agent/Instrument deverbal nouns, Event nominalizations, *-if* adjectives, ...
Lexeur	Wauquier et al. (2020)	Agent/Instrument deverbal nouns, Event nominalizations
Dénom	Strnadová (2014)	All derived adjectives
Mordan	Koehl (2012)	Deadjectival nouns
Converts	Tribout (2010)	Verb<>Noun conversions

Table 1: Existing resources documenting French word formation

One defining characteristic of that collection of resources is that they were all constructed with a focus on *breadth* rather than *width*. Each resource was designed with the goal of documenting one or more specific word formation processes, and attempted to retrieve as many types as was possible given the practical constraints of the project. As a consequence, the sample of the French lexicon that is documented has strange characteristics. Some vanishingly rare derived lexemes are included in the sample, while very frequent ones are not, because the process they implement happens not to have been the focus of attention. Even for those processes that are documented, samples for different processes have different characteristics. For instance, *Lexeur* or *Dénom* contain many items not documented in dictionaries, because it was relatively straightforward to collect instances from corpus data; by contrast, *Converts* focuses on items documented in a dictionary, in the absence of a good method for extracting conversions semi-automatically from a corpus. An unwelcome consequence of this set of affairs is that there is no obvious way to make meaningful statistical comparisons of different processes by combining resources.

Another characteristic of this collection of resources is the variability of annotation both in terms of quantity and quality. For instance, *Démonette* contrasts with all the other sources in that pairings of derivationally-related words have not systematically been curated manually, leading to an undocumented quantity of false positives.

From these observations it follows that currently available resources do not provide us with a holistic view of the distribution of word formation processes in the lexicon. The goal of the present research is the development of a new resource that fills that gap: we provide a coherent and relatively detailed set of morphological annotations for a carefully sampled set of French nouns.

2 Sampling

The sampling procedure was as follows. We started from the *Lexique* (New et al., 2007) and *flexique* (Bonami et al., 2014) databases: *Lexique* provides various types of annotations for words attested in either a French literary corpus or a corpus or subtitles, and *flexique* tabulates all nouns, verbs and adjectives of *Lexique* in inflectional paradigms, and provides manually corrected phonemic transcriptions and grammatical gender information for all forms of the corresponding lexemes. We limited attention to the 13,046 nouns with a summed relative frequency in the two reference corpora higher than 0.3 per million, ensuring that we were focusing on nouns that are relatively frequent, but may be more prevalent either in formal or in informal French.[1]

Sampling was done in two steps. In an initial annotation campaign, we excluded all nouns homophonous with another noun, either with the same orthography but the other gender (there can be inanimates, e.g. LIVRE$_F$ 'pound' vs. LIVRE$_M$ 'book', or animates, e.g. JOURNALISTE$_{F/M}$ 'female/male journalist'), or with different orthographies (e.g. SERRE 'greenhouse' vs. CERF 'deer'). These constitute 8% of the 13,046 nouns we sampled from. This particular sampling strategy was motivated by the needs of a separate study on the phonological and morphological predictability of gender (Bonami et al., 2019). In a second annotation campaign, we first sampled 318 nouns with homophones so as to rebalance the sample; we then sampled more nouns until we reached a total of 5,000 nouns after exclusion of tagging errors. Note that, for purposes of sampling, masculine and feminine variants of common gender nouns such as JOURNALISTE were counted as two separate items; hence for some nouns (e.g. HUMORISTE 'comedian') both the masculine and the feminine variants are present in our sample, while for others either the feminine (e.g. HUMANISTE 'humanist') or the masculine (e.g. EXISTENTIALISTE 'existentialist') is. This is of course a disputable choice (Bonami and Boyé, 2019), but there was no way of avoiding taking a stance on the status of human common gender nouns.

3 Manual morphological annotation

The annotation of the dataset was made by two annotators, both authors of the paper. In a first step, each one annotated about 850 nouns that were checked by the other annotator afterwards. All difficulties were discussed and decisions were made collectively. After guidelines for the annotation were drawn up,[2] the remaining nouns were distributed between the authors. All problems and questions were discussed and solved collectively.

Each noun was annotated for different properties. First, we annotated the broad morphological status of the noun as being either simplex or not; nonsimplex nouns were then classified on the basis of the outermost word formation process involved: prefixation, suffixation, conversion, any nonconcatenative process (`nonconcat` in the tables) or formation from more than one word (`polylexical` in the tables). When there was uncertainty as to what the last process was, we relied on frequency for arbitration. For example, SOUS-ALIMENTATION 'undernourishment' is ambiguous between an outermost prefixation (from ALIMENTATION 'feeding') or suffixation (from SOUS-ALIMENTER 'undernourish'). Because ALIMENTATION has a higher frequency than SOUS-ALIMENTER in *Lexique*'s reference corpora, we considered it to be the base of SOUS-ALIMENTATION and thus coded the last process as prefixation.

[1] The particular threshold of 0.3 per million was motivated by backward compatibility with the previous study of Tribout et al. (2014), although nothing crucial hinges on that choice.

[2] The guidelines are distributed with the resource in the following OSF repository: `https://osf.io/rdxqk/`.

All broad morphological categories except prefixation and suffixation were divided into fine grained sub-categories. Simplex nouns can be native underived nouns (e.g. CAHIER 'notebook'), borrowings (e.g. JAZZ), antonomasia (e.g. POUBELLE 'bin') or onomatopeic nouns (e.g. CLIC 'click'). The nonconcatenative processes found in the database are reduplication (e.g. BABALLE, from BALLE 'ball'), back formations (e.g. NUMISMATE 'numismatist', from NUMISMATIQUE 'numismatics'), slang processes such as verlan (e.g. KEUF from FLIC 'cop') or louchébem (e.g. LARFEUIL, from PORTE-FEUILLE 'wallet') and different types of truncation: mere apocope (e.g. IMPRO, from IMPROVISATION 'improvisation'), apocope with addition of an ending (e.g. VALOCHE, from VALISE 'suitcase') and apheresis (e.g. SCOPE, from MICROSCOPE 'microscope'). Among polylexical processes, we distinguished native compounds (e.g. SÈCHE-CHEVEUX, 'hairdryer', from SÉCHER 'dry' and CHEVEUX 'hair'), neoclassical compounds (e.g. BARYTON, 'baritone'), blends (e.g. FADETTE, from FACTURE 'bill' and DÉTAILLÉE 'detailed'), acronyms (e.g. SIMA, from SILICIUM 'silicon' and MAGNÉSIUM 'magnesium') and frozen word sequences, which we call `agglomerate`s (e.g. ARC-EN-CIEL 'rainbow', litterally 'bow in sky'). The difference between native compounds and agglomerates lies in the nature of elements: a lexeme was classified as an agglomerate if and only if one of the combined expressions is a grammatical word (e.g. *en* in ARC-EN-CIEL) or an inflected form (e.g. *dira* in QU'EN-DIRA-T-ON 'word of mouth', litt. 'what will one say'). Conversions were classified by base part of speech; Table 2 gives an example for each of the documented situations. Note that, following (Tribout, 2012), we distinguish four subcases of conversion from verbs depending on

POS	Stem type	Base	Translation	DERIVATIVE	Translatiin
Adjective	—	PEUREUX	'fearful'	PEUREUX	'fearful person'
Verb	basic stem	RECHERCHER	'research'	RECHERCHE	'research'
	infinitive	SOUVENIR	'remember'	SOUVENIR	'memory'
	past participle	ENTRER	'enter'	ENTRÉE	'entrance'
	learned	CONCEVOIR	'conceive'	CONCEPT	'concept'
	indeterminate	FAILLIR	'fail'	FAILLITE	'bankruptcy'
Noun	—	RAVIN	'ravine'	RAVINE	'small ravine'
Proper name	—	SUISSE	'Switzerland'	SUISSE	'Swiss'
Adverb	—	DEHORS	'outside'	DEHORS	'outside'
Pronoun	—	MOI	'me'	MOI	'ego'
Numeral	—	ONZE	'eleven'	ONZE	'the number eleven'

Table 2: Examples illustrating the diversity of base part of speech in converted nouns.

which stem allomorph is used. Note also that all deverbal nouns that can be analyzed as conversions from past participles are so analyzed, irrespective of whether or not a suffix is involved in the formation of that participle: hence examples such as ENTRÉE 'entrance', ACCALMIE 'lull', VENUE 'arrival', and ENCEINTE 'enclosure' are all treated on a par.

While we neither provide a full account of a lexeme's derivational history nor its relationship to all members of its derivational family, we did use the tabular structure of the database to document more word formation processes involved in a lexeme's formation. For instance, in addition to the outermost process, we also noted in dedicated columns whether other word formation processes (conversion, compounding, prefixation, or suffixation) are involved in the formation of the noun. For instance, the entry for EMBARQUEMENT 'boarding' documents it as formed by suffixation from EMBARQUER 'board', but also notes that it contains the verb-forming prefix *en*.

This annotation is particularly useful in situations where determination of the base-derivative relationship is nonobvious. As a case in point, consider the situation with conversion between nouns and verbs. Following Tribout (2020), we distinguish three situations: where the verb is clearly morphologically complex, it has to be the base of the noun (e.g. RECHERCHER 'research' has to be the base of RECHERCHE 'research' because of the presence of the verb-forming prefix *re-*); where the noun is clearly morphologically complex, it has to be the base (e.g. PARLEMENT 'parliament' is clearly based on PARLER 'speak' by

-ment suffixation and hence is the base of PARLEMENTER 'negociate'); in all other cases (e.g. with CLOU 'nail' vs. CLOUER 'to nail'), directionality cannot be established—in particular Tribout (2020) shows that neither etymological information nor semantic intuitions are reliable indicators or directionality. We account for the commonality of all three types of cases by noting in the `conversion` column the existence of a relationship with a verb, but differentiate them by coding the last process as `conversion` for RECHERCHE, `suffixation` for PARLEMENT, and `simplex` for CLOU.

In the same spirit, for compounding, we noted the compound type even if compounding is not the last morphological process. For example, THALASSO is the apocope of THALASSOTHÉRAPIE 'thalassotherapy' which is a neoclassical compound. In this case, the `compound` column indicates `neoclassical` while the last morphological process is noted as `apocope`. When there was hesitation between prefixation or compounding, particularly when the first element corresponds to a preposition such as *sur*, *sous*, *arrière*. . . , we arbitrated in favor of prefixation. Finally, when suffixation is involved in the formation of the noun, we noted the suffix in a specific suffix column, be it the last process (e.g. MINCEUR 'slimness' from MINCE 'slim') or not (e.g. PORTE-CIGARETTES 'cigarette case' from PORTER 'to bear' and CIGARETTE 'cigarette' that itself comes from CIGARE 'cigar'), like we did for the other processes.

Because suffixes are the most frequent derivational processes in our data, in addition to the mention of the suffix we also annotated different kinds of information linked to the suffixation process. First, we annotated suffix identity at two levels of granularity: the `sfx` column indicates the surface orthography of the precise allomorph, while `sfx_broad` lumps together allomorphs and gendered variants. For instance, *-oir* as found in RASOIR 'razor' and *-oire* as found in PASSOIRE 'colander' are distinguished at the fine grained level but grouped together under *-oir* at the coarse grained level. Similarly, the *-able* and *-ible* suffixes found in NOTABLE 'noteworthy' (from NOTER 'to note') and NUISIBLE 'harmful' (from NUIRE 'to harm') are distinguished at the fine grained level and noted as two allomorphs of the suffix *-able* at the coarse grained level. It is important to note that, except for gender variation and basic allomorphy, the identification of the suffixes is only based on the form of the suffixes and the gender they assigned to the nouns: no semantic or syntactic information is taken into account. For example, we distinguished two *-ure* suffixes, one feminine (e.g. BRÛLURE$_{\text{F}}$ 'burn' from BRÛLER 'to burn') and one masculine (e.g. SULFURE$_{\text{M}}$ 'sulphide' from SOUFRE 'sulphur'), but only one suffix *-ier*, be it used to form the name of a tree (e.g. AMANDIER 'almond tree' from AMANDE 'almond'), a person (e.g. BANQUIER 'banker' from BANQUE 'bank') or an artifact (e.g. SUCRIER 'sugar bowl' from SUCRE 'sugar'). As a consequence, we identified only one suffix in cases of homonymous suffixes used in distinct derivational processes if they assign the same gender to the outputs. Therefore the resource contains one suffix *-age*, even if we usually differentiate two suffixation processes: one deverbal *-age* suffixation that forms action nouns (e.g. JARDINAGE 'gardening' from JARDINER 'to garden') and one denominal that forms collective nouns (e.g. OMBRAGE 'shade' from OMBRE 'shadow'). However, the difference between the two *-age* suffixations can still be retrieved through the part of speech of the base that is noted in a dedicated column. In addition to the fine grained and coarse grained suffixes, we noted in dedicated columns the base of suffixation, its part of speech and whether it is autonomous (e.g. MINCEUR 'slimness' from MINCE 'slim') or not (e.g. LACTOSE 'lactose' from LACT- 'milk'). In some cases the identification of the base is tricky. Below, we describe these cases and the decisions we made.

i) There could be a mismatch between the formal and the semantic base of suffixation, as in ROYALISTE 'royalist': it formally derives from the adjective ROYAL by addition of the suffix *-iste*, but it is semantically related to the noun ROI 'king' rather than the adjective ROYAL. In such cases we arbitrated in favor of the formal base.

ii) For all demonym formation processes as well as *-iste* suffixation, which form parallel adjectives and nouns, following (Roché, 2008) we considered a direct suffixation from the base to the inhabitant or supporter noun, without an intermediate adjectival step. For example, the noun PARISIEN 'parisian' is treated as directly derived from PARIS (1a), not from an adjective itself deriving from the city name (1b). However, in order to capture the relation between the noun and the homonymous adjective, the existence of an adjectival counterpart is noted in the `conversion` column.

(1) a. PARIS → PARISIEN$_N$
→ PARISIEN$_A$
b. PARIS → PARISIEN$_A$ → PARISIEN$_N$

The same method was applied to *-isme* and *-iste* nouns: when a shared base exists, both nouns were analyzed as derived from that base. For example, ARRIVISTE 'social climber' and ARRIVISME 'ambition' are annotated as both derived from ARRIVER 'to arrive'.

iii) Sometimes suffixation applies to a bound stem. If this stem appears in at least one other word, it was considered to be a non autonomous base (Corbin, 1987). For example, in DÉLATRICE 'informer' the *-rice* suffix applies to the string *délat-* that is also found in DÉLATION 'informing', so that *délat-* was annotated as the non autonomous base of DÉLATRICE.

iv) When the string the suffix attaches to is not found elsewhere in the lexicon, but the noun belongs to a derivational series—it has the shape and expected meaning of a derivative (Hathout, 2009), the noun was considered to be a suffixed noun having no base. Therefore, we noted `0` in the `base` column. For instance, MAQUETTE 'model' ends with *-ette* while the stem *maqu* does not appear in other words, so that it cannot be a non autonomous base. However, MAQUETTE has the same ending and the same diminutive meaning as suffixed nouns in *-ette* like FILLETTE 'small girl' (from FILLE 'girl') so that it belongs to the derivational series of diminutive nouns suffixed with *-ette*. Therefore MAQUETTE was annotated as a suffixed noun having no base.

4 Descriptive statistics

	Count	Proportion
Simplex	2064	41%
Suffix	1865	37%
Conversion	564	11%
Polylexical	298	6%
Nonconcat	125	2%
Prefix	84	2%

Table 3: Type frequency by broad morphological type

	Count	Proportion
Verb	887	48%
Noun	603	32%
Adjective	179	10%
No POS	101	5%
Name	83	4%
Numeral	11	1%
Adverb	1	0%

Table 4: Type frequency of suffixed nouns by base part of speech[3]

We briefly comment on some descriptive statistics. Table 3 reports the breakdown of the dataset in terms of broad morphological types. The striking results are the low prevalence of polylexical units, and the high prevalence of simplex nouns. The latter result is partly due to the presence of 431 borrowings, as well as many items which were morphologically analyzable at some point in the history of French (e.g. MANIÈRE 'manner', historically derived by conversion from a now disappeared adjective MANIER 'to be used with the hand', itself from MAIN 'hand') or were analyzable in Latin (e.g. VICTOIRE 'victory', from Latin VICTORIA which itself was derived from Latin VICTOR 'victor').

Since suffixes make up the bulk of nonsimplex nouns and have been annotated in more detail, we focus on them in the remainder of this paper. 82 broad suffixes are attested in the dataset, with a high diversity of type frequencies, as shown in Figure 1: 8 suffixes account for more than half of the data, and two thirds of the suffixes have a type frequency lower than 10. Table 4 shows that deverbal formations make up almost half of the data, dominating nouns and then adjectives.

Finally, we report in Figure 2 the median token frequency of derivatives by suffix, for those suffixes with 10 or more instances in the dataset. It is striking that suffixes forming abstract feminine nouns

[3]The 'No POS' label corresponds to situations where either there is no identifiable base (while there is an identifiable suffix) or the base is a bound stem.

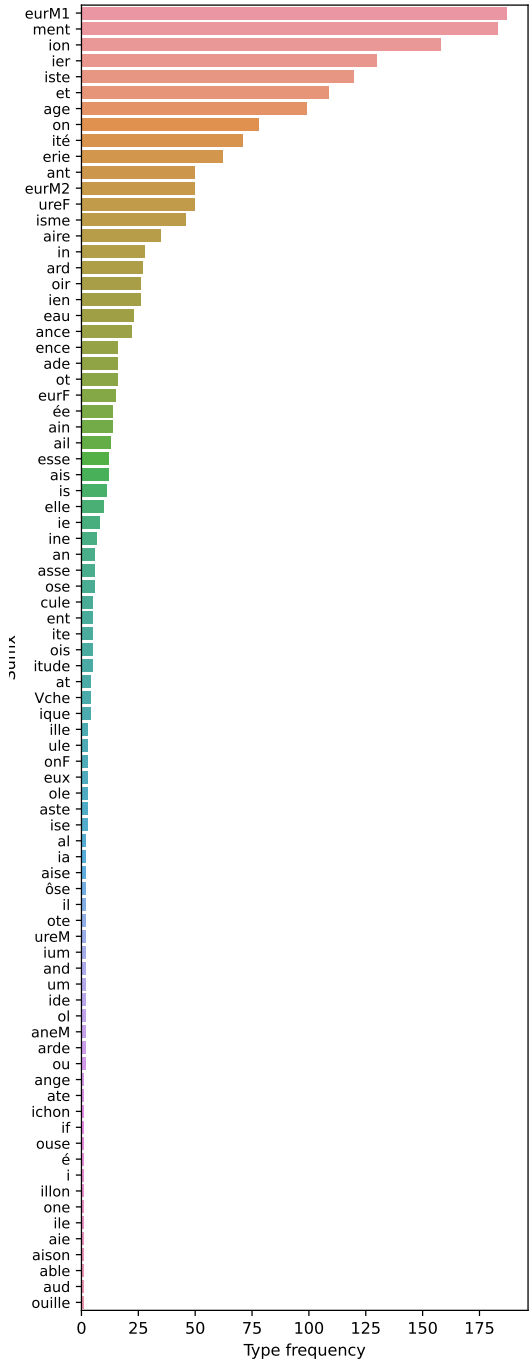


Figure 1: Type frequency of the 82 suffixes

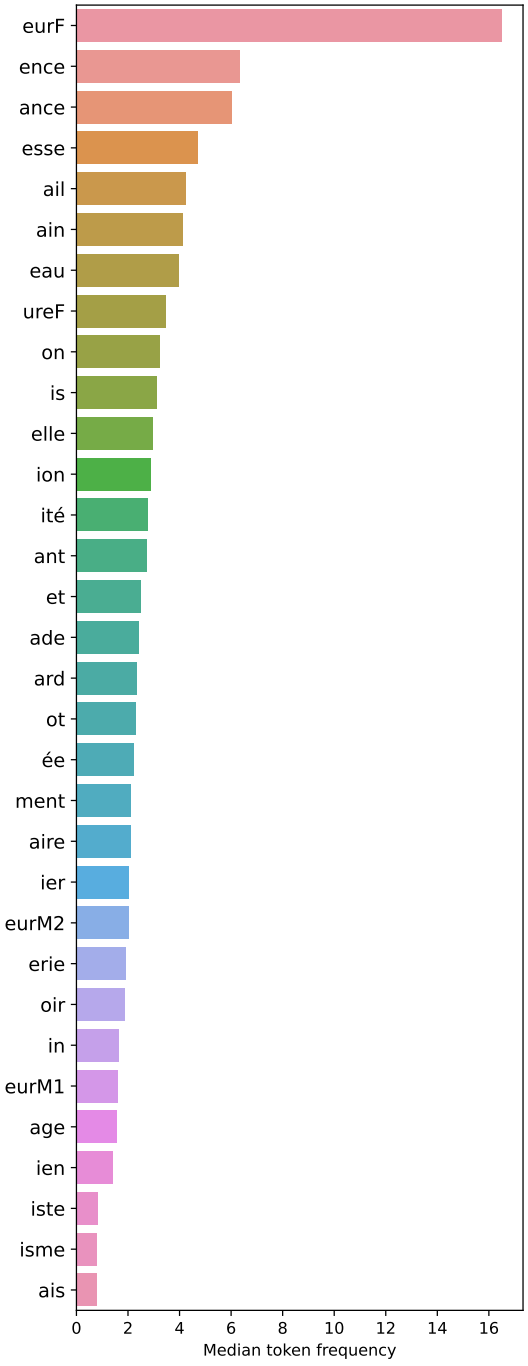


Figure 2: Median token frequency of the 32 most type-frequent suffixes

occupy the bulk of the high frequency range, above those forming individual or event-denoting nouns; and that *-isme* on the other hand has very low median frequency. This paper is not the place to attempt an explanation of these tendencies, but they illustrate the type of study allowed by a balanced annotated sample of a word formation system such as *Échantinom*.

5 Transparency

One of our goals with Échantinom was to document the formal and semantic transparency of suffixed derivatives so as to be able to use that information in future modelling efforts. After experimenting with using the raw intuitions of the authors, we concluded that these were unreliable, and that conducting a serious norming experiment over a multi-thousand item lexicon was out of the picture. Hence we report quantitative measures computed from the data.

5.1 Formal transparency

We report two measures of formal transparency: edit distance between base stem and derivational stem, and type frequency of patterns of alternation.

To compute the first measure, we first collected from *flexique* phonemic transcriptions for the citation forms of all nouns in the database. The derivational stem of suffixed derivatives was then deduced by simply stripping out the phonology of the appropriate suffix allomorph. Deducing the appropriate base stem was more challenging. First, we collected from *flexique* a reference stem for each lexeme: the singular form of nouns, the feminine singular form of adjectives, and the imperfect indicative 3SG form of verbs, stripped of the final /ɛ/. These are arguably the basic stem for each part of speech (Bonami and Boyé, 2003, 2005), and are definitely the stem allomorph most often relied on by suffixal derivation. Second, we computed an adapted Levenshtein distance between the derivational stem and the candidate base stem, which ignores differences between tense and lax mid-vowels, and between nasal vowels and matching vowel-/n/ sequences. Third, we examined by hand all cases where the resulting edit distance was larger than zero: in some cases this corresponds to genuine lack of formal transparency, in others it was found to be due to regular morphophonology, or the choice of a distinct stem allomorph. In addition, there were a few dozen of cases where the lexeme documented as the base contains a suffix absent from the derivative; e.g. the base for INSOUCIANCE 'carelessness', suffixed in *-ance*, is INSOUCIANT 'careless', itself suffixed in *-ant*. While this is a sensible decision, it leads to an artificially inflated formal distance between the base and derivational stem. In all such cases, the derivational stem was corrected by hand. We report in the resource the edit distance between the derivational stem and this manually corrected base stem. For instance the distance between INFORMATION 'information' and its base INFORMER 'inform' is `0`, that between INTERDICTION 'prohibition' and INTERDIRE 'forbid' is `2`, and that between DESTRUCTION 'destruction' and DÉTRUIRE 'destroy' is `4`.

Whether edit distance is a good measure of formal transparency in derivational morphology is disputable; Strnadová (2014, chap. 4) argues that it is not, and that the type frequency of patterns of alternation between surface forms is a better indication. The idea is that alternations that are judged as opaque are not those that are formally complex but those that are unexpected, and that unexpectedness is a consequence of low type frequency. To provide a rough operationalization of Strnadova's idea, as follows, we used the difflib Python library's SequenceMatcher algorithm to identify patterns relating the citation forms of the base and the derivative, [4] and then report the relative frequency of a pattern among derivatives formed with the same suffix. The higher the frequency of a pattern is, the more transparent the noun is. For example, the alternation pattern between INFORMATION and INFORMER (_∼_asjɔ̃) has the highest relative frequency (around 0.591) among *-ion* derivatives, which indicates that INFORMATION is very transparent. Conversely, the alternation pattern between CONTRADICTION 'contradiction' and CONTREDIRE 'contradict' (_ə_z∼_a_ksjɔ̃) has the lowest frequency (around 0.007), which indicates that the noun is not transparent. Note that the accuracy of our estimation of relative frequencies is highly dependent of the overall frequency of the affix; while we report relative frequencies for all suffixal formations,

[4] The use of SequenceMatcher as a rough but efficient way to classify surface alternations is inspired by Hathout et al. (2020). See Beniamine (2017) for a much more principled approach to the topic.

we advise against using them for suffixes with fewer than 10 types.

5.2 Semantic transparency

To operationalize semantic transparency, we rely on distributional semantics (see Boleda 2020 for a recent overview). We rely on a distributional vector space computed from frcow (Schäfer and Bildhauer, 2012) corpus for the purposes of Guzmán Naranjo and Bonami (2021), which provides lexeme-based rather than word-based distributional vectors for French.[5]

Using these vectors, we computed two separate measures of semantic transparency. First, we provide the cosine similarity between the vector for a suffixed noun and the vector for its base. This captures the idea that words that are transparently related occupy adjacent regions of semantic space, which may not be the case if the relationship is not transparent (see e.g. Varvara et al. 2021 for recent discussion). For instance, the cosine similarity between the vector of INITIALISATION 'initialisation' and that of INITIALISER 'initialize' is around 0.785, which indicates a high semantic similarity between the noun and the verb. Conversely, the cosine between the vector of PRESSION 'pressure, stress' and that of PRESSER 'press' is very low (around 0.011), which correlates with the semantic difference between the two words, as the noun is usually used with a psychological meaning while the verb has almost always a physical denotation.

In addition to this, we provide a measure of the predictability of the relationship between base and derivative. To this effect, for all suffixed nouns, we compute the difference (or offset) between the vector for the derivative and the vector for the base: this represents the shift in semantic space from the base semantics to the derived semantics. Such offset vectors tend to be similar for instances of the same derivational processes or even for rival processes (Guzmán Naranjo and Bonami, 2021). However, within a set of pairs of words related by the same process, we expect to find some variation, with transparent formations having very similar vectors while opaque ones will diverge (Bonami and Paperno, 2018). Hence we compute the average of offset vectors for all derivatives formed using the same suffix, and then the cosine similarity between that average and each individual offset vector. This similarity measure, which we call *offset vector typicality*, tells us the extent to which one instance of a derivational process implements a semantic relation that is similar to what happens on average for other instances of that process. For instance, the offset vector typicality of DESTRUCTION is high (around 0.8), in contrast to that of MUNITION 'ammunition' (from MUNIR 'to provide'), which is about 0.4. Just as with formal transparency assessed through the type frequency of patterns, the quality of our evaluation of offset vector typicality is heavily dependent on the number of datapoints going into the average vector. Hence, while we provide numbers for all derivatives, we urge users to proceed with caution, and definitely advaise against using them for suffixes with fewer than 10 types.

5.3 Discussion

In both the formal and semantic dimension, we provided two operationalizations of transparency: one based on bare comparison of base and derivative, the other based on an assessment of the typicality of their relationship. In the case of formal transparency, we observe a strong although far from perfect correlation between the two measures (Pearson's $r = -0.62$). The violin plot in Figure 3 confirms that there are very few cases where a nonzero edit distance does not coincide with a very low pattern frequency. This suggests that, despite Strnadová's (2014) principled reservations, in practice, edit distance is not such a bad indicator of the formal regularity of a derivative. In the case of semantic transparency on the other hand, our two measures are not correlated at all (Pearson's $r = -0.02$). The density plot in Figure 4, also suggests no interesting nonlinear relationship between the two variables: this suggests that the two measures indeed capture very different aspects of similarity. This is unsurprising though, for the following reason. While distributional vectors do capture some lexical semantic contrasts, they are also heavily influenced by aspects of distribution that have little to do with the two words corresponding to related concepts. As a case in point, consider the fact that the nouns PATINAGE 'skating' and PATINEUR

[5]The vector space is based on a curated version of the lemmatization provided by the corpus, and was obtained using the gensim (Řehůřek, 2010) implementation of the word2vec algorithm (Mikolov et al., 2013). The vector space is lexeme-based in the sense that each word was replaced by its lemma: thus the vector space documents the distribution of lexemes among other lexemes, rather than inflected forms among other inflected forms.

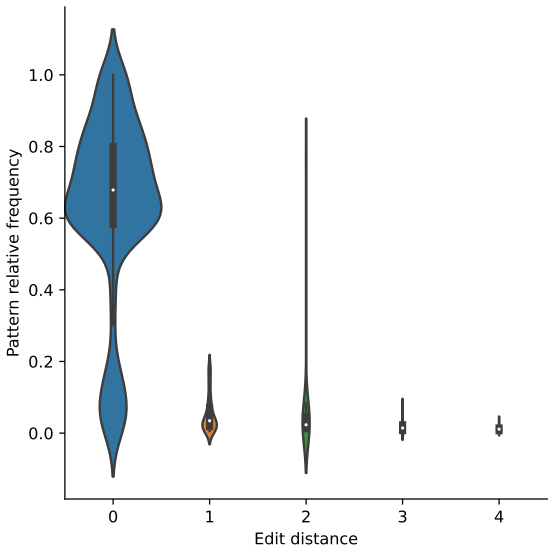


Figure 3: Relative distribution of the two measures of form transparency

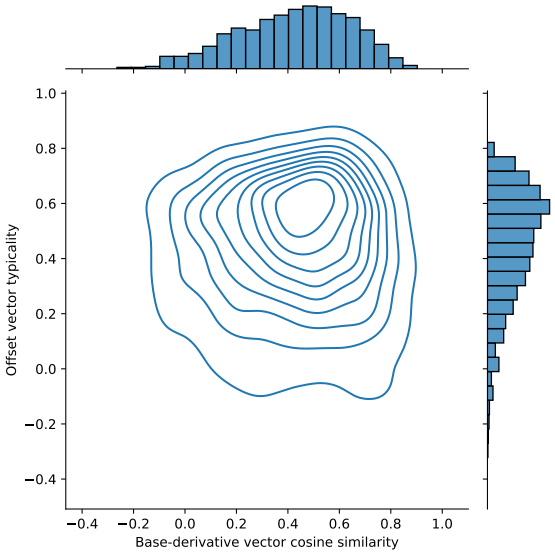


Figure 4: Relative distribution of the two measures of semantic transparency

'skater' are much more similar to one another (cosine similarity 0.72) than either is to the verb PATINER 'skate' (respective cosine similarities 0.20 and 0.28), whereas intuitively the even noun is semantically closer to the verb: clearly what is at play here is the general distributional similarities among nouns and differences between nouns and verbs. On the other hand, offset vector similarity should not be influenced by such factors; as a matter of fact, the similarity between the base and the derivative plays no role here: we do not care how distant they are from one another, but only about the direction in which the difference vector points.

We leave it to future research, or to the attention of future users of the resource, to study how formal and semantic transparency correlate with one another and with other variables of interest.

Acknowledgments

Part of this work was done in the context of a project on gender assignment in French, in collaboration with Matías Guzmán Naranjo, and which benefited from in internship by Nadège Demanée. We thank them both for their input on relevant aspects of the research. This work was partially supported by the Démonext project (ANR 17-CE23-0005) as well as a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

References

Sacha Beniamine. 2017. Une approche universelle pour l'abstraction automatique d'alternances morphophonologiques. In *Actes de TALN 2017*. pages 77–85.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6(1):213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303.

Olivier Bonami and Gilles Boyé. 2003. Supplétion et classes flexionnelles dans la conjugaison du français. *Langages* 152:102–126.

Olivier Bonami and Gilles Boyé. 2005. Construire le paradigme d'un adjectif. *Recherches Linguistiques de Vincennes* 34:77–98.

Olivier Bonami and Gilles Boyé. 2019. Paradigm uniformity and the French gender system. In Matthew Baerman, Oliver Bond, and Andrew Hippisley, editors, *Perspectives on morphology: Papers in honour of Greville G. Corbett*, Edinburgh University Press, Edinburgh, pages 171–192.

Olivier Bonami, Gauthier Caron, and Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer, and Sophie Prévost, editors, *Actes du quatrième Congrès Mondial de Linguistique Française*. pages 2583–2596.

Olivier Bonami, Matías Guzman Naranjo, and Delphine Tribout. 2019. The role of morphology in gender assignment in French. Presented at the Second International Symposium on Morphology (ISMo 2019).

Olivier Bonami and Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio* 17(2):173–195.

Danielle Corbin. 1987. *Morphologie dérivationnelle et structuration du lexique*. Max Niemeyer Verlag, Tübingen.

Matías Guzmán Naranjo and Olivier Bonami. 2021. Comparing derivational processes with distributional semantics. Presented at the second Paradigmo workshop.

Nabil Hathout. 2009. *Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie*. Habilitation à diriger des recherches. Toulouse 2 - Le Mirail.

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5):125–168.

Nabil Hathout, Franck Sajous, Basilio Calderone, and Fiammetta Namer. 2020. Glawinette: a linguistically motivated derivational description of French acquired from GLAWI. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, pages 3870–3878.

Aurore Koehl. 2012. *La construction morphologique des noms désadjectivaux suffixés en français*. Ph.D. thesis, Université de Lorraine.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Fiammetta Namer, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout, and Delphine Tribout. 2019. Demonette2 — Une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de TALN*. Toulouse, France. https://halshs.archives-ouvertes.fr/halshs-02275652/document.

Boris New, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28:661–677.

Radim Řehůřek. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. pages 45–50.

Michel Roché. 2008. Structuration du lexique et principe d'économie : le cas des ethniques. In *Actes du Congrès Mondial de Linguistique Française 2008*. pages 1571–1585.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. pages 486–493.

Jana Strnadová. 2014. *Les réseaux adjectivaux: Sur la grammaire des adjectifs dénominaux en français*. Ph.D. thesis, Université Paris Diderot et Univerzita Karlova V Praze.

Delphine Tribout. 2010. *Les conversions de nom à verbe et de verbe à nom en français*. Ph.D. thesis, Université Paris Diderot.

Delphine Tribout. 2012. Verbal stem space and verb to noun conversion in french. *Word Structure* 5(1):109–128.

Delphine Tribout. 2020. Nominalization, verbalization or both? Insights from the directionality of noun-verb conversion in French. *Zeitschrift für Wortbildung / Journal of Word Formation* 2/2020:187–207.

Delphine Tribout, Lucie Barque, Pauline Haas, and Richard Huyghe. 2014. De la simplicité en morphologie. In *Actes du 4^e Congrès Mondial de Linguistique Française (CMLF 2014)*. volume 8 of *SHS Web of Conferences*, pages 1879–1890.

Rossella Varvara, Gabriella Lapesa, and Sebastian Padó. 2021. Grounding semantic transparency in context. *Morphology* https://doi.org/https://doi.org/10.1007/s11525-021-09382-w.

Marine Wauquier, Cécile Fabre, and Nabil Hathout. 2020. Semantic discrimination of technicality in french nominalizations. *Zeitschrift für Wortbildung / Journal of Word Formation* 2/2020:100–121.

Critical analysis of clinical resources and tools for derivational morphology used in francophone speech and language therapy

Guillaume Duboisdindien **Georgette Dal**

Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France

✉ guillaume.duboisdindien@univ-lille.fr ✉ georgette.dal@univ-lille.fr

Abstract

Context: An increasing number of speech and language pathologists (SLP) tools are being marketed by specialized French-language publishers. Given the clinical focus of these tools, a critical approach to their evaluation is required. The aim of this preliminary study is to identify the main characteristics of the clinical resources designed for derivational morphology used by French-speaking SLPs. **Method:** A french criterion-referenced and critical analysis grid was developed to collect and analyze data from 15 resources for morphological remediation and/or learning. **Results:** The corpus of occurrences compiled from the 15 clinical tools is a collection of 8251 entries. The collected structures were automatically filtered and revealed 5134 occurrences of (presumed) complex lexemes. We present in this paper the 10 most frequent lexemes in such tools. The preliminary results of this study indicate that the francophone remedial materials used by SLPs for working on morphology and derivational morphology present weaknesses in their general characteristics, in the typology of the morphological tasks provided, and in the efficacy of the choice of derivational lexemes targeted for remedial treatment.

1. Introduction & Context

Over the past fifteen years, several studies have collected developmental and clinical data on the developmental role played by derivational morphology in word reading, literacy and vocabulary (Carlisle, 1995; Nation and Snowling, 2004; Reed, 2008). In adults, studies suggest that the mechanisms devoted to word formation can be altered and generate combinations that deviate from the derivational rules in the case of neurological disorders such as post-stroke aphasia or the semantic variant of Primary Progressive Aphasia (Badecker and Caramazza, 2001; Auclair-Ouellet et al., 2017). Some promising avenues of clinical research in children, with more modest outcomes in adults, explore treatment models that include derivational morphology among their active components (Goodwin and Ahn, 2010; Galuschka and Schulte-Körne, 2016). Alongside these studies, increasing numbers of Speech and Language Pathology (SLP) and educational tools are being marketed by specialized francophone publishers. However, their efficacy remains to be proven. Given the clinical focus of these tools, a critical approach to their evaluation is needed (Profetto-McGrath, 2005; Lof, 2011).

Empirical evaluations of the efficacy of remediation tools and resources is rather tricky to find. For several years, studies in the theory of decision-making behaviour (Bettman et al., 1991; Alba and Hutchinson, 1987) have shown that the marketing community - including the publishing industry - has consistently used the influence of cognitive bias to guide users' choices. As an example, some studies show that the adaptations (e.g. modifying the text font, creating contrast through use of a coloured background, adjusting line spacing) that publishers design for children's literature for learners, people with dyslexia and/or poor readers[1], are not equally or even demonstrably beneficial (Bachmann and

[1] During reading, **poor readers** may have decoding difficulty, thus difficulty reading words in context accurately (see among others Stanovich, 1988).

Mengheri, 2018; Hakvoort et al., 2017; Kuster et al., 2018). Thanks to cognitive bias, it is possible to play on a certain number of beliefs about these adaptations in order to convince the public of their efficacy as therapeutic or educational tools. Critical thinking is a good springboard for questioning user behaviour and tendencies regarding their choice of tools or practices (Law et al., 2008).

In order to determine the uses and needs of Speech and Language Pathologists (SLPs) with respect to derivational morphology, a survey was conducted via the French Demonext Project[2] (Namer and Hathout, 2019) starting in November 2020. It was addressed to SLPs in a number of French-speaking areas (France, Belgium, Switzerland, Canada, Monaco, Luxembourg, Niger). The initial results of this survey, which is still in progress, indicate that 35% of SLPs design treatment targeting derivational morphology. Ten percent of them estimate their level of knowledge on this topic at *very little* to *little awareness* and 23% claim to have *general awareness* of derivational morphology. However, the qualitative analysis of the corpora of spontaneous responses regarding the types of activities they propose and their terminological and theoretical knowledge of derivational morphology suggests that the gaps are 10 to 20% greater than the gaps revealed by the self- assessment among 387 respondents.

The choice of clinical and learning materials specific to derivational morphology is thus ripe for a deeper investigation. **The three research questions developed for this preliminary study are:**

1) What are the qualitative characteristics of French speech and language materials for clinical activities in morphology?

2) Are there any preferred (presumed) complex lexemes found in therapeutic tools targeting derivational morphology? If so, which ones and what are their properties?

3) What tasks are used most frequently to stimulate the mechanisms of derivational morphology in remedial materials?

The aim of this preliminary study is to identify the main characteristics of the therapeutic materials designed for derivational morphology used by French-speaking SLPs. In this context, we will present the preliminary results of our study providing a critical analysis of francophone SLP tools and resources for derivational morphology.

2. Methodology for the evaluation of SLP tools and resources for derivational morphology

2.1 Elaboration of a two-level criterion-referenced grid for data collection

A French criterion-referenced and critical analysis grid was designed to collect and analyze data from SLP and educational resources oriented toward morphology, and specifically derivational morphology. The methodological framework combines the principles of criterion-referenced evaluation and provides up-to-date external data from interdisciplinary scientific literature (i.e. Linguistics; Evidence-Based Practice; Education; Rehabilitation Sciences; Didactics). A typology of relevant and theoretically valid criteria was defined across two distinct levels of analysis.

The first level is designed to globally evaluate the design quality of the selected materials in six domains (general criteria) (i.e *Data* on *expertise and marketing information, Ergonomic and technical qualities, Target population, Global objectives and social validity, Theoretical validity, Measures of equipment/treatment efficacy)* and 22 sub-domains (sub-criteria) (e.g. *Theoretical and social validity of the materials, expertise of the authors, ergonomics, quality of the instructions*, etc.).

The second level specifically evaluates the types of tasks presented as well as the (presumably) complex lexemes selected in the support materials. Ten tasks involving morphological activities were tagged using a taxonomy adapted from Berthiaume et al. (2010). A criterion checking phase was then carried out independently by two expert morphologists to adjust this taxonomy and the related concepts. The grid was constructed in April 2020 and tested with 3 β-testers and the principal designer in May 2020. A phase of adjustment and random re-testing with the 3 β-testers took place in June 2020 with test

[2] ANR-17-CE23-0005-04.

support. The phase of analysis of the SLP resources and tools was held from June to September 2020.

2.2 Selection of SLP resources and tools

The selected materials come from francophone publishers that claim to be specialized in SLP and remedial pedagogy. The materials were identified on the web using the following descriptors: [1] morphology, [2] morphological awareness, [3] derivational morphology, [4] morphological composition, [5] spelling, [6] vocabulary. From the outset, as the identification of such resources and tools is affected by their availability, variety and online exposure, a systematic and controlled approach was not considered.

2.3 Data collection procedures

An identification number was assigned to each of the resources and tools selected for study. For each criterion, the judge's collection procedure followed detailed and systematically referenced guidelines and rationales. In practice, for the first phase of analysis, the manuals presenting the materials or the pedagogical chapters were analyzed: when available, all of the publisher's manuals, instructions, appendices and the promotional website were consulted in their entirety. Judge 1, who designed the grid, carried out an analysis on all of the materials selected for study. Judge 2 then performed a blind analysis of 20% of the data collected from 5 randomly selected sets of materials.

During the first level of the SLP materials analysis (Phase 1), judgements were weighted using a Likert Scale (see Likert, 1932), with the weighting distributed evenly across the sub-criteria, in particular those associated with a vigilance marker. This marker highlights a key criterion in the design of a resource. To consider weighting (i.e., qualitatively assigning a grade), the judge must examine all of the sub-criteria before assigning a qualitative score to the overall criterion in the analysis table. Gradients range from 0-red (unsatisfactory) to 4-dark green (satisfies all criteria) with color coding. In Phase 1, all of the SLP resources and tools were treated consecutively. In Phase 2, a delay occurred to ensure that Judge 1 was not subject to a halo effect or contamination bias. During the second level of the SLP support analysis (phase 2), complex lexemes were collected and labelled according to i) the typology (i.e. suffixed lexeme, prefixed lexeme); ii) their base; and iii) their lexical category (i.e. noun, verb, adverb, adjective), including pseudoword. The occurrences recorded through the SLP supports were then automatically identified and filtered through the AntConc concordance software (Anthony, 2004).

3. Preliminary results

3.1 Results of the general characteristics analysis of the SLP resources and tools

The results concern 15 resources designed for morphological remediation and/or learning. Only two satisfied most of the criteria in the analysis grid and none of the vigilance markers were involved in the scoring. For both of these cases, the authors were careful to systematically provide relevant and substantial information to administrators in their instruction manuals and promotional websites. Both approaches were theoretically valid and the tools had been tested beforehand in qualitative studies. The results of the general characteristics analysis of the SLP morphological resources and tools are the following:

> ***Quality Criterion 1: Data on expertise and marketing information*** indicates that 53.33% (n=8) of the SLP tools achieve a grade of 3 (criterion is mostly satisfied); 40% (n=6) of the materials obtain a grade 0 (criterion unsatisfied) and 6.6% (n=1) a grade 1 (criterion poorly satisfied).
>
> ***Quality Criterion 2: Ergonomics and technical qualities*** reveals that 46.67% (n=7) of the materials analysed obtained a grade of 0 in this category; 40% (n=6) obtained a grade of 2 or 3, (equal distribution), and only two tools obtained a grade of 4.
>
> ***Quality Criterion 3: Target Population*** indicates that 60% (n=9) of the tools are at grade 0, while 13.33% (n=2) are at grade 1 and grade 2. Finally, 26.67% (n=4) of the materials are rated at grade 4.
>
> ***Quality Criterion 4: Global objectives*** is dependent on Criterion 5: Theoretical validity which

is found below. The results obtained indicate that 40% (n=6) of the materials are scored at grade 1. The rest are divided as follows: 20% (n=3) each at grade 0, grade 3, and grade 4.

Quality Criterion 5: Theoretical validity finds a majority of materials at grade 0 with a proportion of 73.33% (n=11), 13.33% (n=2) of the tools came in at grade 2; one tool scored grade 3 and one support obtained grade 4.

Quality Criterion 6: Measure of equipment/treatment effectiveness: The results indicate that 80% of the tools rated grade 0 (n=12). One tool scored grade 2, while one tool scored grade 3. Only one support achieved grade 4.

3.2 Results of the specific analysis of the complex lexemes and morphological activities from the SLP resources and tools

In the second phase, a specific analysis of the data found in the activities on derivational morphology was carried out through a static criterion-referenced grid.

3.2.1 Results of complex lexemes used in SLP supports

In total, the corpus of occurrences compiled from the 15 SLP activities is a collection of 8251 entries. The collected structures were automatically filtered through the AntConc concordancer software and revealed 5134 occurrences of (presumed) complex lexemes. We cite here the most frequent lexemes, i.e. those appearing 10 or more times:

16 occ.	13 occ.	12 occ.	11 occ.	10 occ.
ILLÉGAL "illegal"	INCROYABLE "unbelievable" INJUSTE "unfair" LAITIER "milkman"	FLEURISTE "florist" INCAPABLE "unable"	DENTISTE "dentist"	COUPURE "cut" GUITARISTE "guitar player" ILLOGIQUE "illogical" INHUMAIN "inhuman" IRRÉGULIER "irregular" OUVERTURE "opening" PRÉHISTOIRE "prehistory" SÉCHAGE " drying"

Tab 1: most frequent complex lexemes in the 15 SLP tools

First, we note that these 15 lexemes are highly lexicalized, and that some of them first appeared in French a very long time ago: 11th century (OUVERTURE), 13th century (INJUSTE, LAITIER, COUPURE, IRRÉGULIER), 14th century (ILLÉGAL, INHUMAIN), 16th century (INCROYABLE, INCAPABLE), 17th century (FLEURISTE), 18th (DENTISTE, SÉCHAGE), 19th century (GUITARISTE, PRÉHISTOIRE). Six of them have a Latin cognate (ILLÉGAL, INJUSTE, INCAPABLE, INHUMAIN, IRRÉGULIER, OUVERTURE), so it is not clear whether they were constructed in French or whether they are borrowings.

These 15 lexemes correspond to six morphological patterns. Two of them are prefixations: inX (ILLÉGAL, INCROYABLE, INJUSTE, INCAPABLE, ILLOGIQUE, INHUMAIN, IRRÉGULIER), préX (PRÉHISTOIRE); the other four are suffixations: Xiste (FLEURISTE, DENTISTE, GUITARISTE), Xure (COUPURE, OUVERTURE), Xier (LAITIER) and Xage (SÉCHAGE).

Among these most frequent complex lexemes, the inX pattern is overrepresented (7/15=46,67%, corresponding to 49,4% of the sum of occurrences). The two remediation activities with the best general characteristics also display this kind of over-representation. However, as has been shown (Schwarze, 2007; Dal and Namer, 2014), *in-* prefixation in French is not productive, except with deverbal adjectives in *-able* (e.g. DÉCHIRABLE / INDÉCHIRABLE; TRANSFÉRABLE / INTRANSFÉRABLE). Only two of the seven lexemes in *in-* that are frequently used in SLP, INCROYABLE and INCAPABLE, appear to have this linear structure inXable. Yet in the second, the segment *cap-* is difficult to relate to a French verb (CAPABLE is inherited from Latin CAPABILIS).

Moreover, three of the *in-* lexemes (ILLÉGAL, ILLOGIQUE, IRRÉGULIER) illustrate the phenomenon of graphic regressive assimilation: in these adjectives, the graphic consonant <n> of the prefix is

assimilated to the first consonant C of the radical of the base, if C =<l> or <r>[3]. However, as shown by Buchi (2012), since the 17th century, in such contexts, the prefix *in-* tends to be realized <in> (/ɛ̃/), even when an adjective containing this assimilation is lexicalized. For example, despite the existence of ILLOGEABLE, which is attested from the 16th century, INLOGEABLE appeared in 1784, and INRETROUVABLE (1933) competes with IRRETROUVABLE (1906). A search on the web carried out on July 22, 2021 confirms this competition in contemporary French, with a preference for <in> when the prefixed adjective is not lexicalized (*TLF*):

Base	*in-* = <in>	*in-* = <il> / <ir>
LIVRABLE	*inlivrable(s)*: 516	*illivrable(s)*: 2
LOCALISABLE	*inlocalisable(s)*: 6460	*illocalisable(s)*: 2513
LOUABLE	*inlouable(s)*: 6120	*illouable(s)*: 110
RATABLE	*inratable(s)*: 2 172 000	*irratable(s)*: 121 000
RAYABLE	*inrayable(s)*: 165 000	*irrayable(s)*: 45
RÉCOLTABLE	*inrécoltable(s)*: 185	*irrécoltable(s):* 11

Tab 2: Adjectives in *-able* and their corresponding inX from the web

3.2.2 Results of the specific analysis of the morphological tasks in SLP resources

A qualitative analysis reveals that the derivation task (task. A) is the most frequently occurring task in the instructions (n=8). The main objective of this task is to test the learner's ability to produce correct derivational forms from a given base. In this activity, the teacher/SLP can observe if some previously learned strategies are transferred to similar or equivalent contexts. The labelling task called structural analysis (task. B) is the second most used type (n=4). This task involves identifying and separating the elements involved in complex lexemes. In some cases, the occurrences proposed in those resources are 1) affixed or 2) distractors (pseudo-affixed). The underlying objective is to investigate the learner's ability to analyze the structure of a given word. In order of frequency, the other types of tasks found are the affix choice task (task. G - n = 3) and the relation judgement task (task. D - n = 2). This typology still needs to be refined because the instructions in the materials are sometimes ambiguous, which hinders the process of task categorisation. In fact, the majority of SLP resources and tools (n= 9) indicated one type of task but the activity in fact belonged partially or wholly to another category. In other cases (n=2), the instructions were not precise enough to allow the judge to assign a category label. Except for the two resources that had the best general results, it was not obvious in which area of intervention and for which mode (oral or written), the SLP tools were designed to be used.

4 Discussion

Publishers specialising in speech and language therapy regularly produce tools and resources with a focus on derivational morphology. These products are said to be adapted for Developmental Language Disorders, reading impairments or children with dyslexia. The aim of this study was to identify the main characteristics of materials designed for remediation in derivational morphology used by French-speaking SLPs. First, the results indicate that a majority of SLP tools suffer from a lack of objectivity in their theoretical constructs (operational definitions, theoretical validity, standardized terminology, and developmental knowledge) and in their measurements of social validity. Furthermore, the pedagogical objectives and application procedures of most of these materials are not transparent or are lacking in the instructions (Research Question 1).

Secondly, the instructions for morphological activities are predominantly oriented towards derivation tasks (task A) and structural analysis tasks (task B) (Research Question 3). However, for 13 out of 15 resources it was found that a majority of the tasks proposed either fit into more than one sub-task category or were not specific enough to be formally labelled. Moreover, it is not explicitly stated whether these tasks are to be performed in oral or in written form or both. This is the case for all but the two resources that score very well in their general characteristics.

[3] From a phonological point of view, the digramm <in> is reduced to /i/ in these lexemes.

Finally, an examination of the lexemes frequently used in these SLP tools shows that the occurrences analysed do not stand up to structural and diachronic examination. As said before, the morphological pattern inX is not productive in French, except with deverbal adjectives in *-able.* This calls into question the relevance of employing the most frequent lexemes prefixed with *in-* in SLP supports (Research Question 2). Lexemes such as INHUMAIN or INJUSTE, for example, are not good primary stimuli (task A). In particular, the specific analysis of the derived occurrences raises doubts about the rationales behind the two remediation resources that had the best overall characteristics. The context of the task (task A or B) seems to be in contrast with the target lexemes proposed in most of the supports. At the very least, this raises the question of whether it is in fact the morphological mechanisms that are practiced in tasks associated with this type of pattern, or rather memory skills or the breadth of the learner's mental lexicon. For the field of derivational morphology, Libben et al. (2014) partially support this position, suggesting that a psychocentric perspective should be adopted in order to understand the role that lexical and linguistic variables can play. The psychocentric perspective emphasises the fact that notions such as the morphological structure of lexemes, their frequency or their transparency are, in their essence, abbreviated expressions of the internal cognitive states of a language user, rather than stimulus-independent features. According to this view (if it is indeed appropriate to conceptualize word recognition as a process by which a reader or listener makes sense of a read or heard stimulus), it is the internal states of the language user at the time of recognition that are, in fact, the source of the outcomes that a researcher, clinician, or educator measures as dependent variables in a psycholinguistic experiment or teaching.

The lexicon of a language such as French is largely made up of complex lexemes: prefixed, suffixed, converted or compounds. The question of lexicon invokes different cognitive and cultural representations among speakers - including SLPs and teachers - which may be motivated either by their relationship to the norm (i.e. what cognitive resources or normative tools would justify the use of a particular complex lexeme rather than another?) or by use (i.e. a complex lexeme used within a community or the spontaneous appearance of a meaningful construction). This structural information is usually obtained by speakers in the etymological sections of dictionaries. However, the variability of its formulation makes it difficult to exploit, not to say artificial, from the point of view of usage. Moreover, studies on morphological processing in French do not achieve a consensus and suffer from the same inconsistencies as those found in English (Sánchez-Gutiérrez et al., 2018). Many variables are involved and authors mention that morphological processing in French is influenced by root frequency, word-level morphological structures (suffixed versus prefixed words), semantic or even pre-semantic morphological processing (Colé et al., 2002; Beyersmann et al., 2014).

Some welcome efforts have been made to provide resources to investigate the morphology of French. For example, the French database DériF (Hathout et al., 2002; Namer, 2009) describes a collection of complex lexemes, while the French-Canadian cross-linguistic database MorphoLex-Fr (Mailhot et al., 2020) focuses on characterizing the different variables that impact the morphology of English and French. These types of contributions provide operational descriptions of occurrences and variables that modulate morphological processing. International scientific contributions that describe the derivational properties of words in a systematic way and anchor their approach at the interface of fundamental linguistic research, applied research and societal demand would make it possible to respond to these multiple challenges. This is in particular the case of the research project Demonext, which aims to develop a morphological database with an automated interface to empirically confirm and define hypotheses in morphology, develop tools for automatic language processing (ALP), vocabulary teaching and the treatment of developmental or acquired language disorders. To conclude, the use of a strict taxonomy of tasks and lexematic structural analysis is relevant for the analysis of clinical materials.

5 Conclusion

The preliminary results of this study indicate that the francophone remedial materials used by SLPs for working on morphology and derivational morphology present weaknesses in their general characteristics, in the typology of the morphological tasks provided, and in the efficacy of the choice of derivational lexemes targeted for remedial treatment. A critical analysis of materials for remedial and clinical use is relevant to enabling SLPs to make informed and evidence-based choices in their practice

and in the materials they use. Additional studies should be carried out along the same lines, alongside the development of research at the interface of linguistic expertise and clinical needs.

Acknowledgments

This research was supported by the *Agence Nationale de la Recherche*, grant number: ANR-17-CE23-0005.

The authors thank Kathleen O'Connor, Ph.D., for her help in proofreading this paper.

References

Joseph W. Alba and J. Wesley Hutchinson 1987. Dimensions of consumer expertise. *Journal of consumer research* 13(4): 411–454. Doi: doi.org/10.1086/209080.

Laurence Anthony. 2004. AntConc: A learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. *IWLeL 2004: An interactive workshop on language e-learning*, pages 7–13. http://www.laurenceanthony.net/research/iwlel_2004_anthony_antconc.pdf.

Noémie Auclair-Ouellet, Marion Fossard, Robert Laforce Jr, Nathalie Bier, and Joël Macoir. 2017. *Conception* or **conceivation*? The processing of derivational morphology in semantic dementia. *Aphasiology* 31: 166–188. doi: 10.1080/02687038.2016.1168918.

Christina Bachmann and Lauro Mengheri. 2018. Dyslexia and Fonts: Is a Specific Font Useful? *Brain sciences* 8(5): 89. Doi: 10.3390/brainsci8050089.

Alfondo Badecker and William Caramazza. 2001. Morphology and aphasia. In A. Spencer and A. M. Zwicky (eds.), *The handbook of Morphology*. Blackwell Reference Online, pages 390–405. https://www.blackwellpublishing.com/content/BPL_Images/Content_Store/WWW_Content/9780631226949/22Chap20.pdf.

Rachel Berthiaume, Anne-Sophie Besse, and Daniel Daigle. 2010. L'évaluation de la conscience morphologique : proposition d'une typologie des tâches. *Language Awareness* 19(3): 153–170. Doi: 10.1080/09658416.2010.482992.

James R. Bettman, Eric J. Johnson, and John W. Payne. 1991. Consumer decision making. In T. S. Robertson and H. H. Kassarjian (eds), *Handbook of consumer behaviour*, Prentice-Hall, pages 50–84.

Elisabeth Beyersmann, Galina Iakimova, and Joahannes C. Ziegler. and Pascale Colé. 2014. Semantic processing during morphological priming: An ERP study. *Brain Research*, 1579: 45–55. Doi: 10.1016/j.brainres.2014.07.010.

Éva Buchi. 2012. *Réel, irréel, inréel* : Depuis quand le français connaît-il deux préfixes négatifs *in-* ? In M. Barra-Jover (éd.), *Études de linguistique gallo-romane*. Saint-Denis, Presses universitaires de Vincennes, pages 323–340. https://www.cairn.info/etudes-de-linguistique-gallo-romane--9782842923426-page-323.htm.

Joanne F. Carlisle. 1995. Morphological awareness and early reading achievement. In L. B. Feldman (ed.), *Morphological aspects of language processing*, Hillsdale, NJ: Erlbaum, pages 189–209. Doi: 10.1023/A:1008131926604.

Pascale Colé, Carine Royer, Christel Leuwers, and Séverine Casalis. 2004. Les connaissances morphologiques dérivationnelles et l'apprentissage de la lecture chez l'apprenti-lecteur français du CP au CE2. *L'année psychologique* 104(4): 701–750. https://www.persee.fr/doc/psy_0003-5033_2004_num_104_4_29686.

Georgette Dal and Fiammetta Namer. 2014. Adjectifs positifs en *-able* et négatifs en *in-* correspondants en français : ou pourquoi seuls sont importables les ordinateurs portables. In F. Neveu et al. (eds), *Actes du 4e Congrès Mondial de Linguistique Française, Berlin, Allemagne, 19-23 juillet 2014*, pages 1741–1754. https://doi.org/10.1051/shsconf/20140801341.

Katharina Galuschka and Gerd Schulte-Körne. 2016. The Diagnosis and Treatment of Reading and/or Spelling Disorders in Children and Adolescents. *Deutsches Ärzteblatt International* 113(16): 279–286. Doi:

10.3238/arztebl.2016.0279.

Amanda P. Goodwin and Soyeon Ahn. 2010. A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of dyslexia* 60(2): 183–208. Doi: 10.1007/s11881-010-0041-x.

Britt Hakvoort, Madelon van den Boer, Tineke Leenaars, Petra Bos, and Jurgen Tijms. 2017. Improvements in reading accuracy as a result of increased interletter spacing are not specific to children with dyslexia. *Journal of Experimental Child Psychology* 164: 101–116. Doi: 10.1016/j.jecp.2017.07.010.

Nabil Hathout, Georgette Dal, and Fiammetta Namer. 2002. An experimental constructional database: the MorTAL project. In P. Boucher (ed.), *Many morphologies*, Cambridge Mass, Cascadilla Press, pages 178–209. halshs-00319461v1.

Sanne M. Kuster, Marjolijn van Weerdenburg, Marjolein Gompel, and Anna M. T. Bosman. 2018. Dyslexie font does not benefit reading in children with or without dyslexia. *Annals of Dyslexia* 68(1): 25–42. Doi: 10.1007/s11881-017-0154-6.

James Law, C. Campbell, Sue Roulstone, Catherine Adams, and James Boyle. 2008. Mapping practice onto theory: the speech and language practitioner's construction of receptive language impairment. *International Journal of Language & Communication Disorders* 43(3): 245–263. Doi: abs/10.1080/13682820701489717.

Gary Libben, Kaitlin Curtiss, and Silke Weber. 2014. Psychocentricity and participant profiles: implications for lexical processing among multilinguals. *Frontiers in psychology* 5: art. 557. Doi: 10.3389/fpsyg.2014.00557.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 140: 1–55. https://psycnet.apa.org/record/1933-01885-001.

Gregory L. Lof. 2011. Science-based practice and the speech-language pathologist. *International Journal of Speech-Language Pathology* 13(3): 189–196. Doi: 10.3109/17549507.2011.528801.

Hugo Mailhot, Maximiliano A. Wilson, Joël Macoir,S. Hélène Deacon, and Claudia Sánchez-Gutiérrez. 2020. MorphoLex-FR: A derivational morphological database for 38,840 French words. *Behavior research methods* 52(3): 1008–1025. Doi: 10.3758/s13428-019-01297-z.

Fiammetta Namer. 2009. *Morphologie, lexique et traitement automatique des langues : l'analyseur DériF*. Paris, Hermès.

Fiammetta Namer and Nabil Hathout. 2019. ParaDis and Démonette, From Theory to Resources for Derivational Paradigms. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*. Prague, Czech Republic, pages 5–14. https://hal.archives-ouvertes.fr/hal-02288938.

Kate Nation and Margaret Snowling. 2004. Beyond phonological skills: broader language skills contribute to the development of reading. *Journal of research in reading* 27(4): 342–356. Doi: 10.1111/j.1467-9817.2004.00238.x.

Johanne Profetto-McGrath. 2005. Critical Thinking and Evidence-based Practice. *Journal of professional nursing* 21(6): 364–371. Doi: 10.1016/j.profnurs.2005.10.002.

Deborah K. Reed. 2008. A Synthesis of Morphology Interventions and Effects on Reading Outcomes for Students in Grades K–12. *Learning Disabilities Research & Practice* 23(1): 36–49. Doi: 10.1111/j.1540-5826.2007.00261.x.

Claudia H. Sánchez-Gutiérrez and Natividad Hernández-Muñoz. 2018. Development of derivational morphological awareness in Anglophone learners of Spanish: A relational knowledge study. *Foreign Language Annals* 51: 369–388. Doi: 10.1111/flan.12344.

Christoph Schwarze. 2007. La notion de règle morphologique et les mots complexes non construits. In N. Hathout and F. Montermini (éds), *Morphologie à Toulouse. Actes du colloque international de Morphologie 4èmes Décembrettes*. München: Lincom Europa (LSTL 37), pages 221–244. https://kops.uni-konstanz.de/handle/123456789/45765.

Keith E. Stanovich. 1988. Explaining the Differences Between the Dyslexic and the Garden-Variety Poor Reader:

The Phonological-Core Variable-Difference Model. *Journal of Learning Disabilities* 21: 590–604. Doi: 10.1177/002221948802101003.

Describing valence-increasing constructions with XMG

Valeria Generalova
Heinrich Heine University of Dusseldorf
Dusseldorf, Germany
generalo@hhu.de

Abstract

This paper presents an analysis of syntactic and semantic derivations caused by the addition of a derivational affix to the verb.[1] It examines two valence-increasing constructions – causative and applicative – leading to the formation of three-argument cores. The paper contributes to the project based on Role and Reference Grammar (RRG) and uses eXtensible MetaGrammar (XMG) as the main technical framework.

1 Introduction

Morphological derivation sometimes leads to further structural changes.[2] Namely, in some languages, there are morphological devices to encode valence-modifying operations. As the argument structure changes, syntactic and semantic derivations take place. This paper examines constructions with a morphological derivation of a verb leading to valence increase. It aims to present a formalized description suitable for grammar-engineering purposes of both syntactic and semantic derivations conditioned by the addition of an affix to a verb.

This paper is a position work and does not offer any quantitative experimental data. Nevertheless, it aims to contribute to an existing grammar engineering project, which has been first presented as Generalova and Petitjean (2020). The present paper grounds on Role and Reference Grammar (RRG, Van Valin and LaPolla (1997); Van Valin (2005)). It is a powerful theory created bearing typological distinctions of languages in mind, which is an asset for a multilingual grammar engineering project. For creating syntactic trees, we use the formalized version of it presented by Osswald and Kallmeyer (2018). In contrast to classical RRG, it uses features for encoding properties of syntactic constituents and for linking them to other dimensions. For representing semantics, we use decompositional frames, as suggested in Lichte and Petitjean (2015). These data structures are well compatible with features in the syntactic dimension. They also allow a fine-grained unification of features, which is important for a precise description of complex linguistic phenomena.

The paper is organized as follows. In Section 2, we describe the scope of the present paper and show examples of the data we use. We also introduce the most important definitions. Section 3 presents the theory and the technical framework used in our study. It also compares the present paper to a similar project realized in another framework. Section 4 presents our analysis of syntactic and semantic derivations in causative and applicative constructions. Finally, Section 5 summarizes our claims and indicates directions for further studies.

2 Data

To explore how morphological derivation influences syntactic and semantic structures, we took a case where this effect can be observed clearly. We decided to focus on valence-increasing affixes attached

[1]This study is funded by an ERC Consolidator Grant awarded to Prof. Laura Kallmeyer.

[2]Abbreviations: 1 = first person, 3 = third person, ACC = accusative, APPL = applicative, C1 = noun class 1, C7 = noun class 7, CAUS = causative, DAT = dative, FOC = focus, FV = final vowel, M = masculine, NOM = nominative, PRS = present, PST = past, SG = singular.

to transitive verbs. Since most core structures are two-argument, the addition of a third one is a particular operation, which can not be confused with other grammatical phenomena. This is the reason why we narrowed our scope to two types of constructions: causative-of-a-transitive and three-argument applicative.

Our data is collected from secondary sources: grammars, field notes, typological studies. We also make use of various typological databases such as Dryer and Haspelmath (2013); Hartmann et al. (2013) and others.

2.1 Causative-of-a-transitive

We follow the typology of causative constructions couched in Dixon (2000) since it pays special attention to causatives derived from transitive verbs. We try to use bases that exhibit high Transitivity (according to (Hopper and Thompson, 1980)), but some other two-argument verb bases are also attested in our data.

A typical example of a three-argument causative construction is given in (1). The newly introduced argument, the causer, becomes the PSA and demotes the causee to another position. Nevertheless, syntactic tests show that it is still within the CORE and does not become a peripheral adjunct.

(1) *Avanu-∅ nana-ge bisket-annu tinn-is-id-anu* KANNADA
3SG-NOM 1SG-DAT biscuit-ACC eat-CAUS-PST-3SG.M

'He fed me a biscuit,' or 'He made me eat a biscuit.' (Foley and Van Valin, 1984, 384)

It is not always easy to determine whether the non-macrorole still remains in the core. Many languages do not allow three-argument cores, and thus, causative verbs derived from transitive bases remain two-argument. We try to exclude these cases from our sample by determining the argumenthood of the participants using syntactic evidence.

2.2 Three-argument applicative

We follow Peterson (2007, 1) and define applicative constructions as those encoding "thematically peripheral argument or adjunct as a core-object argument". With respect to RRG terminology, we are going to explore sentences having three arguments in the CORE.

Three-argument applicative constructions are attested in 81 languages, according to Polinsky (2013). However, not all of them are selected for our study. First, we exclude from our sample the so-called "obligatory applicatives" (Peterson (2007, 46), see discussion in Peterson (2007, 50-51)) since we are interested in having a minimal pair of an alternation. Second, we are interested in valency-increasing and not valency-modifying applicatives. In Peterson's terms, both the applicative object and the base object must demonstrate some object properties (Peterson, 2007, 51-53). Examples of this type of construction can be found in Bantu languages, see (2).

(2) *N-ä-ï-lyì-í-à m-kà k-élyà* CHAGA
FOC-1SG-PRS-eat-APPL-FV C1-wife C7-food

'He is eating food for his wife' Bresnan and Moshi 1993:49-50 cited in Pylkkänen (2002, p. 17,(2a))

There is one more important difference between applicative affixes that is relevant to this paper. Peterson (2007, 40-45) distinguishes between affixes that mark an applicative construction where the semantic role of the applicative object is determined ("morphologically distinct") and universal ("morphologically non-distinct") markers that appear in several types of applicative constructions. In our study, we work with both types of applicative affixes and treat them differently in Sec. 4

3 Background

This section briefly overviews the theoretical and methodological background of our study. It features only the most relevant concepts related to the present paper.

3.1 RRG

The present paper is driven by Role and Reference Grammar = RRG (Van Valin and LaPolla, 1997; Van Valin, 2005). This theory has been developed with linguistic diversity in mind and thus suits our goals well.

Syntactic representations in RRG are realized as trees, where each layer corresponds to a syntactic entity. Our study will be dealing with CORE structures – syntactic levels comprising the predicate with all its arguments, but nothing more. The predicate is placed in the NUCLEUS, being the essential part of the CORE. A CLAUSE, which is a well-known unit in any linguistic paradigm, is built upon a CORE and also includes PERIPHERY (non-arguments).

We also use the concept of macroroles from the classical RRG (Van Valin, 2005, 60–68). There exist two macrorles – Actor and Undergoer – that, in fact, are extreme generalizations of semantic roles. This approach helps to trace similarities between different grammatical phenomena and across languages.

Usually, in constructions with transitive verbs, the syntactic subject (called PSA = privileged syntactic argument) bears the Actor macrorole and the direct object – the Undergoer. Since there are only two macroroles[3], in three-argument constructions, one participant does not bear any macrorole (= is a non-macrorole participant, NMR). However, syntactic and semantic derivations can lead to macrorole reassignment: for example, in causative constructions like (1), the former actor becomes NMR, since the newly introduced participant is Actor; the Undergoer macrorole remains assigned to the same participant.

For the semantic representations, we use frames in the form of attribute-value matrices as they allow for keeping track of typed features (Lichte and Petitjean, 2015). We follow the approach suggested by (Osswald and Kallmeyer, 2018) (more discussion and comparison with other solutions can be found in (Kallmeyer and Osswald, 2013)). These data structures tell the type of the predication (more or less corresponding to the Aktionsart, see Van Valin (2005, 32-42)), the list of semantic roles of this predication, the correspondence of these roles to macroroles, and, if available, other semantic features. From the technical side, frames are easy to implement in XMG and thus inherit information from each other to reflect generalizations vs. language-specific information.

3.2 XMG

We use the technical framework called XMG = eXtensible MetaGrammar (Crabbé et al., 2013; Petitjean et al., 2016) that has been designed for describing various grammatical structures. The XMG description language is static and declarative. It means that instead of formulating rules that apply consequently, XMG descriptions comprise immutable constituents and constraints that regulate their combinations. With the use of conjunction and disjunction, XMG gets rid of the “ordering and termination issues" often occurring in procedural approaches and offers “monotonic (no information removal)” descriptions ((Crabbé et al., 2013, 597)).

The basic unit of an XMG metagrammar is a class, which can correspond to an entity of any level, from morpheme to sentence. A class can comprise one or several dimensions (syntactic, semantic, morphological, etc.). Within each dimension, one can declare variables and assign them features. The values of these features can be specified or defined as unification constraints (e.g. value of feature f_1 on variable v_1 is required to be equal to value of feature f_1 on variable v_2).

Classes are organized in hierarchies. This is possible due to the inheritance mechanism that allows to borrow the content of one class and add it to the description of another one. Conjunction and disjunction help fine-tuning the borrowing process to ensure the inheritance of necessary fragments only. The inheritance mechanism contributes to the greater modularity of descriptions. For two classes being alike to even a small extent, it is possible to declare the common traits apart and only once in order to inherit them afterwards.

[3]Some discussion about the third macrorole can be found in Van Valin (2007); Haspelmath (2006); Diedrichsen (2012); Kailuweit (2013).

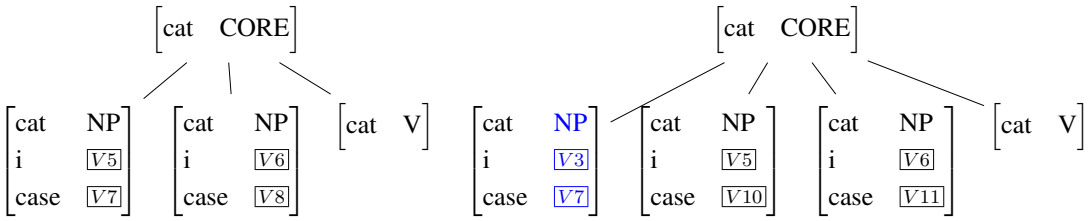


Figure 1: Syntactic trees for a (i) two-argument construction and (ii) the derived three-argument causative construction

3.3 Comparison to Curtis (2018)

A well-developed solution for valence-modifying constructions has been presented in Curtis (2019), and, in more detail, Curtis (2018). It accounts for a small number of typologically varied and genetically unrelated languages, following the typology by **?**. Their project is similar to ours in the respect that it also seeks to decompose complex rules into simpler “building blocks". The difference is that our study is rooted deeper in the typology and does not present benchmark tests (yet).

It is difficult to offer a close comparison of this solution with ours since both the theoretical backgrounds (HPSG vs. RRG) and the used methodologies vary a lot. First of all, Curtis (2019) describes only semantic structures, while one of the key goals of our project is to suggest linked syntactic and semantic representations.

Second, the method in Curtis (2019) is focused on the modification of initial semantic structures. Thus, it suggests “rules" that tell how frame changes have to take place. In contrast, our focus is to determine what morphological items are responsible for what parts of syntactic and semantic derivations. The structures we produce represent the resulting representations and trace all intermediary steps required for their combination.

However, some common traits can be found in both projects. For example, the concept of “axes of variation" (Curtis, 2019, 116-117) implies that each “rule component" (which can be a constituent, a constraint or something else) can have only a limited and pre-defined scope of variation. In our project, limited variation also plays an important role in filtering out ungrammatical structures. But in our architecture, these axes are built with the restriction on the values of features in each XMG class.

4 Analysis

4.1 Syntactic changes

In traditional causative and applicative constructions, the valence-increasing affix adds one argument to the syntactic core of the clause. An example of a tree structure corresponding to the derivation of a three-argument causative construction is shown in Fig. 1, which corresponds to the derivation of 1. Labels $\boxed{V5}$ and $\boxed{V6}$ identify the syntactic arguments of the non-derived two-argument construction. The label $\boxed{V7}$ refers to the case normally used for the PSA (=syntactic subject). It encodes one of the arguments of the non-derived construction and the causer, the added argument in the derived construction (identified as $\boxed{V3}$). Other cases in the derived construction depend on the language and the construction type and are not necessarily identical to those used in the non-derived construction (which is shown by new labels $\boxed{V10}$ and $\boxed{V11}$).

Formally, it could have been realized as (sister) adjunction (Osswald and Kallmeyer, 2018, 362-363). However, the method of adjunction is not suitable for argument addition. It is normally exploited for adding peripheral elements to the core since they do not interact with the part they attach to. In contrast, the procedure of argument addition requires a certain re-analysis of the initial syntactic structure (e. g., macrorole reassignment, change of case marking). Therefore, an intuitively simple derivation through adjunction is not an option for verbal derivation.

Another approach is to perceive argument derivation as a nuclear juncture (Van Valin, 2005, 234-239). The core idea is that one postulates two initial predications that merge into a more complex one. This

approach has received both support and critique (see Cole (1983); Alsina (1992); Kemmer and Verhagen (1994); Horvath and Siloni (2011) *inter alia*). With relation to valence-increasing constructions formed by morphological means, we find this analysis especially farfetched. The increase of valence is realized by means of an affix, which in most languages is not related to another predicative elements. By all means, once it is fully grammaticalized, there is no morphological clue for construing valence-increasing constructions as complex clauses.[4]

What we actually suggest is to delimit tree-specific and construction-specific properties from each other. One separate XMG class determines the shape of the tree, in particular, the number of arguments and their order. The latter is realized thanks to the unification of the value of the feature `word order` with the value of the same feature in a language description in a separate XMG class.

In turn, classes encoding specific constructions import the respective class from the set of three shapes (e.g., the construction `causative-of-a-transitive` would take the three-argument template) and assign values to other features, namely, morphological cases, macroroles, etc. To ensure that classes describing specific constructions do not match unintended tree shapes, one can use the XMG notion of *family* (Petitjean et al., 2016, 594). Trees belonging to the same family can share structural information with each other, but not with other families.

Thus, we suggest grouping syntactic trees by the number of arguments they have. In this solution, `simple transitive` and `causative-of-an-intransitive` would fall into the same family as being two-argument constructions, while, for example, `causative-of-a-transitive` would be in a different, three-argument family.

Using this solution, we approach the very question of syntactic derivation differently. The derivation now does not consist in modifying the initial structure. The derived structure with an increased number of arguments has to be accessed through a different family. To achieve this, one has to ensure that the combination of the verbal stem and the derivational affix indicates the family the resulting tree belongs to unambiguously. This is possible due to the unification of feature values between classes describing the verbal stem, the derivational affix and the resulting derived construction.

This is illustrated in Fig. 2. In both tree structures, on the CORE level, there are boolean features `trans (transitive)` and `caus (causative)`. In the model shown in Fig. 2, they determine the tree shape (other features being omitted for the sake of example's clarity, feature values on arguments are determined as if illustrating one particular language). Thus, the transitive non-causative (upper) tree has only two arguments. The causative-of-a-transitive (lower) tree is more complex. The value of the feature `trans` is defined on the verbal stem and percolates upwards to the V (verb) node and then to the CORE. The feature `caus` is defined on the causative affix since its value is negative in the absence of this affix. It also percolates upwards, and the two positive values at the CORE level make this structure match the three-argument tree shape.

So, the combination of features at the core level determines the construction class, which, in turn, belongs to a single family. This ensures the selection of the appropriate syntactic derivation.

In terms of syntactic mechanisms, all valence-increasing constructions look alike. Moreover, valence-reducing constructions are represented in our solution similarly: stems and affixes are assigned features whose combination is unique for describing a construction and conditions the selection of the correct tree shape.

In the next section, we will show how the semantic derivation is realized and the differences between causative and applicative valence-increasing constructions.

4.2 Semantic changes

As already mentioned, we use decompositional frames to represent semantic structures. Similar to syntactic structures, they comprise features that are assigned values. Moreover, feature values can be unified across dimensions, i. e., syntactic features can share their values with semantic features and vice versa.

[4]Nevertheless, one can perform syntactic tests to attest whether a construction demonstrates some properties of a bi-clausal entity.

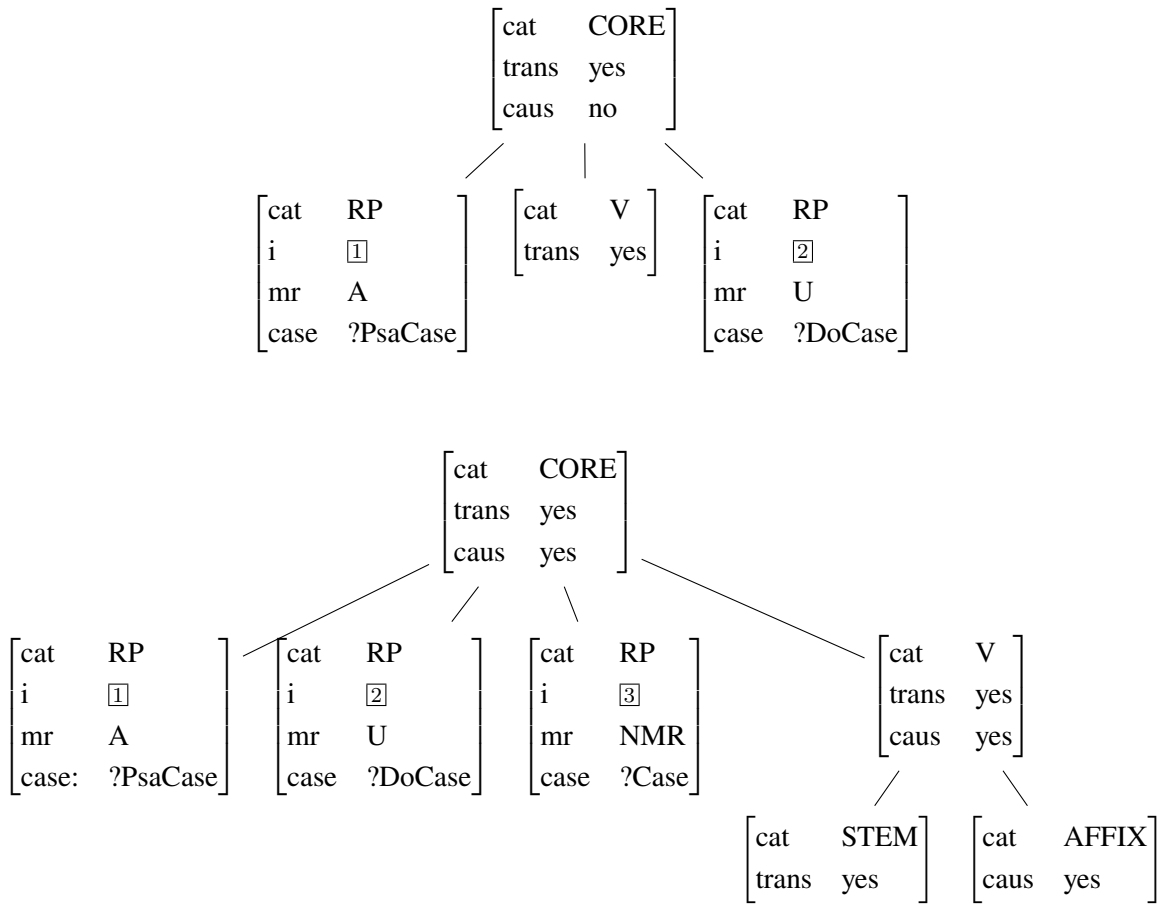


Figure 2: Features percolating up to the CORE level in a (i) simple transitive construction, (ii) causative-of-a-transitive construction

The semantic derivation of a causative construction requires the composition of a complex frame from elementary subframes, see 3. The *effect* being the frame of the base verb with its role structure (we use generalized terms to label two different participants). The *cause* subframe encodes the activity of causation and the participant responsible for it, i. e., the causer. In all languages, the sole participant of the *cause* subframe becomes the Actor of the resulting construction (see label below subframes), while the Undergoer macrorole can be assigned to either participant of the *effect* subframe depending on the language.

In XMG, frames can be inherited from other classes and combine with each other. Since the *effect* subframe encodes the role structure of the base verb, it is clear that it is assigned to the verbal stem in the Lexicon. The *cause* subframe is assigned to the derivational causative affix. They combine in a single frame at the level of the V (Verb) node, and the participant labels are specified at the level corresponding to the CORE (once all participants are already known).

Frames allow the introduction of additional purely semantic features. For example, some languages distinguish between factual, permissive, prohibitive and other constructions where the causation activity receives some additional semantics. These properties can be introduced to the semantic frame of the causation by using the feature *manner* (see Generalova (2020) for details). Importantly, this feature can be introduced at any level. If two constructions with different semantics are derived by means of different verbal affixes, the *manner* feature is going to be part of the frame stored for the respective affix in the Lexicon. If constructions differ at a syntactic level, additional features can be introduced in the construction class (at the level corresponding to the syntactic CORE).

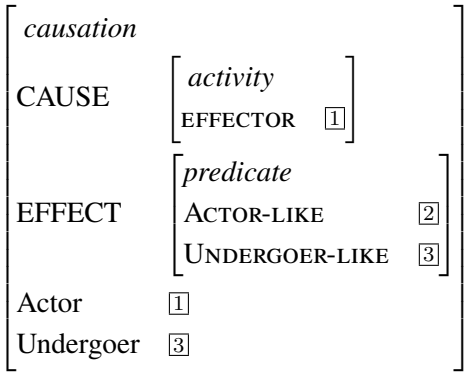


Figure 3: Generalized frame for causative construction

Applicative constructions are more complicated in this respect since the derivation does not involve a second frame. The applicative morpheme does not bear a (sub)frame by itself but operates on the frame of the base verb. Nevertheless, the basic mechanism of splitting features between the verbal stem and the derivational affix can still be applied.

As demonstrated in Sec. 2, there are two main types of applicative affixes. First, some languages have families of applicative constructions, where the suffixes indicate the role of the applied object. In this situation, the postulation of a boolean feature (like `caus: yes`) for applicative constructions is insufficient. Indeed, there is only one way of frame modification introduced by the causative affix, but several different applicative constructions. Therefore, we suggest postulating a categorical feature `applied object` that would take as its value the semantic role of the applicative object (e.g., `ben` in (2)).

A more complicated case is the "morphologically non-distinct applicative construction marker" (Peterson, 2007, 43), i. e. a marker that encodes several constructions. This kind of affixes can not bear the precise semantic role of the applicative object.

One possibility is to list all the semantic roles that can be assigned to the applicative object in a construction with a given suffix is a disjunction. Thus, each applicative verb is potentially the head of the whole family of applicative constructions. The semantic frame with the correct semantic role is thus selected at the latter stage (corresponding to the core level), where all participants are already in the sentence.

However, this approach is problematic since it is difficult to determine the semantic role of the participant in the applicative construction from its syntax. For precise diagnostics, additional lexical features have to be defined in the Lexicon for each noun. Without that, one could only tell that the frame comprises an applied object without specifying its role. Actually, this latter logic is used in our project, while the elaboration of optimal feature sets is the next step of our research.

5 Conclusion

In this paper, we examined constructions where morphological derivation entails syntactic and semantic changes. By exploring three-argument causative and applicative constructions, we demonstrated that the syntactic effect of both types of affixes is fairly similar. In contrast, mechanisms of semantic derivation used in causative constructions differ from those taking place in applicative constructions. We formalized our analysis using Role and Reference Grammar and eXtensible MetaGrammar. Without going deep into technical details, we demonstrated mechanisms of class inheritance and feature unification that result in producing linked syntactic and semantic representations of three-argument constructions.

The present paper is a part of a larger project. Within its scope, the next steps would be to refine the treatment of morphologically non-distinct applicative affixes and extend the methodology to handle not only valence-increasing constructions.

References

Alex Alsina. 1992. On the argument structure of causatives. *Linguistic inquiry* 23(4):517–555.

Peter Cole. 1983. The grammatical role of the causee in universal grammar. *International Journal of American Linguistics* 49(2):115–133.

Benoit Crabbé, Denys Duchier, Claire Gardent, Joseph Le Roux, and Yannick Parmentier. 2013. Xmg: extensible metagrammar. *Computational Linguistics* 39(3):591–629.

Christian Curtis. 2019. A parametric approach to implemented analyses: Valence-changing morphology in the lingo grammar matrix. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. pages 111–120.

Christian Michael Curtis. 2018. *A parametric implementation of valence-changing morphology in the LinGO Grammar Matrix*. Master's thesis, University of Washington.

Elke Diedrichsen. 2012. What you give is what you get? on reanalysis, semantic extension and functional motivation with the german bekommen-passive construction .

Robert M. W. Dixon. 2000. A typology of causatives: form, syntax and meaning. In Robert M. W. Dixon and Alexandra Y. Aikhenvald, editors, *Changing valency: Case studies in transitivity*, Cambridge University Press, Cambridge, chapter 2, pages 30–83.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info/.

William Foley and Jr. Robert D. Van Valin. 1984. *Functional syntax and universal grammar*. Cambridge University Press.

Valeria Generalova. 2020. Towards a cross-linguistic description of morphological causatives: issues in syntax-semantics linking. In Alexandra Pavlova, editor, *Proceedings of the ESSLLI & WeSSLLI Student Session 2020*, Brandies University, pages 55–66.

Valeria Generalova and Simon Petitjean. 2020. A prototype of a metagrammar describing three-argument constructions with a morphological causative. *Typology of Morphosyntactic Parameters* 3(2):2951.

Iren Hartmann, Martin Haspelmath, and Bradley Taylor, editors. 2013. *Valency Patterns Leipzig*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://valpal.info/.

Martin Haspelmath. 2006. Ditransitive constructions towards a new role and reference grammar. *Investigations of the Syntax Semantics Pragmatics Interface* 105:75.

Paul J Hopper and Sandra A Thompson. 1980. Transitivity in grammar and discourse. *language* pages 251–299.

Julia Horvath and Tal Siloni. 2011. Causatives across components. *Natural Language & Linguistic Theory* 29(3):657–704.

Rolf Kailuweit. 2013. Radical Role and Reference Grammar (RRRG). In Brian Nolan and Elke Diedrichsen, editors, *Linking Constructions into Functional Linguistics: The role of constructions in grammar*, John Benjamins Publishing, volume 145 of *Studies in Language Companion Series*, pages 103–142.

Laura Kallmeyer and Rainer Osswald. 2013. Syntax-driven semantic frame composition in lexicalized tree adjoining grammars. *Journal of Language Modelling* 1(2):267–330.

Suzanne Kemmer and Arie Verhagen. 1994. The grammar of causatives and the conceptual structure of events. *Cognitive linguistics* 5(2):115–156.

Timm Lichte and Simon Petitjean. 2015. Implementing semantic frames as typed feature structures with xmg. *Journal of Language Modelling* 3(1):185–228.

Rainer Osswald and Laura Kallmeyer. 2018. Towards a formalization of Role and Reference Grammar. In Rolf Kailuweit, Eva Staudinger, and Lisann Künkel, editors, *Applying and expanding Role and Reference Grammar (NIHIN Studies)*, Freiburg: Albert-Ludwigs-Universität, Universitätsbibliothek, pages 355–378.

David A. Peterson. 2007. *Applicative constructions*. Oxford University Press.

Simon Petitjean, Denys Duchier, and Yannick Parmentier. 2016. Xmg 2: describing description languages. In *International Conference on Logical Aspects of Computational Linguistics*. Springer, pages 255–272.

Maria Polinsky. 2013. Applicative constructions. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Mariliina Pylkkänen. 2002. *Introducing arguments*. Ph.D. thesis, Massachusetts Institute of Technology.

Jr. Robert D. Van Valin. 2005. *Exploring the syntax-semantics interface*. Cambridge University Press.

Jr. Robert D. Van Valin. 2007. The role and reference grammar analysis of three-place predicates. *Suvremena lingvistika* 33(63):31–63.

Jr. Robert D. Van Valin and Randy J. LaPolla. 1997. *Syntax: Structure, meaning, and function*. Cambridge University Press.

Adding Glawinette into Démonette: pratical consequences and theoretical questions

Nabil Hathout
CLLE, CNRS
Université de Toulouse

Fiammetta Namer
Université de Lorraine
ATILF, CNRS

Abstract

Glawinette is a derivational lexicon of French made up of morphological families and morphological series. It has been acquired automatically from GLAWI, a large machine readable dictionary and contains about 100 000 pairs of morphologically related lexemes. In this paper, we present Glawinette and discuss how we plan to include this new resource into the Démonette derivational database, what changes this may bring to the architecture of this database and how this inclusion will raise several theoretical questions regarding the content of Démonette and the nature of derivational paradigms.

1 Introduction

Glawinette (Hathout et al., 2020) is a newly created resource which provides a description of derivational morphology of French on a large scale. In this paper, we discuss its inclusion into the Démonette database (Hathout and Namer, 2014, 2016; Namer et al., 2019; Namer and Hathout, 2020). This will increase the size of Démonette and test the robustness of the principles underlying the structure of Démonette and its description formats by confronting them with the diversity of derivational relations contained in Glawinette. In addition, it will involve a manual revision of Glawinette. Verification and inclusion into Démonette will be done in batches, starting with the most reliable lexeme pairs and lexeme clusters. Batch processing will also make it easier to complement the descriptions of Glawinette and to fill in the semantic fields of Démonette.

2 Glawinette

Glawinette is a derivational morphological lexicon of French built from the GLAWI machine readable dictionary (Sajous and Hathout, 2015; Hathout and Sajous, 2016). Like Démonette, and before it Morphonette (Hathout, 2011a), Glawinette is a lexicon of derivational relations which enables a smooth and easy integration into Démonette. Morphological relations (i.e. pairs of morphologically related lexemes) are acquired from the definitions of GLAWI and the morphological sections of this dictionary. Specifically, these relations are extracted from the so-called morphological definitions (Martin, 1983), i.e., definitions where the *definiendum* is a complex lexeme whose meaning is described by a *definiens* that contains a member of its morphological family, as in (1) that link *glaçon* 'ice cube' to *glace* 'ice' and *développement* 'development' to *développer* 'to develop'. In these examples, the morphological relations are direct (base → derivative), but this is not always the case as in (2) where *conservation* 'conservation' is not the base for *conservateur* 'preservative'.

(1) a. *glaçon = morceau de glace* 'piece of ice'
b. *développement = action de développer, de se développer ou résultat de cette action, au propre et au figuré* 'act of developing or result of this action, literally and figuratively'

(2) *conservateur = substance chimique minérale ou organique, ajoutée aux aliments afin d'améliorer leur conservation* 'chemical substance, mineral or organic, added to food to improve its preservation'

Glawinette proposes a description of morphological relations within two fundamental structures in the paradigmatic organization of derivational morphology (Bochner, 1993; Van Marle, 1985; Bauer, 1997; Štekauer, 2014; Hathout and Namer, 2018, 2019; Bonami and Strnadová, 2019; Namer and Hathout, 2020): morphological families and morphological series (Roché, 2009; Hathout, 2011b; Fradin, 2018). In Glawinette, morphological families are related graphs of derivational relations like (3) which presents the family of the noun *prince* 'prince'. In addition, every relation (i.e. lexeme pairs) is part of a morphological series as in (4) which presents a part of the series that connects agent nouns in *-eur* and action nouns in *-ion*. The series are labeled by patterns consisting of two regular expressions that contain the same number of sequences `(.+)` and where these sequences represent the same strings.

(3) prince=N:princesse=N 'princess' prince=N:princier=A 'princely' prince=N:princillon=N 'petty prince' prince=N:princiser=V 'make become a prince' princesse=N:prince=N princier=A:prince=N princier=A:princièrement=R 'princely' princillon=N:prince=N princiser=V:prince=N princièrement=R:princier=A

(4)

`^(.+)eur$=N`	`^(.+)ion$=N`		
acteur	action	'actor'	'action'
animateur	animation	'animator'	'animation'
classificateur	classification	'classifier'	'classification'
formateur	formation	'trainer'	'training'

On the one hand, Glawinette takes advantage of the fact that lexeme pairs that enter into regular morphological relations form formal analogies (Lepage, 1998, 2004; Stroppa and Yvon, 2005; Hathout, 2008; Langlais and Yvon, 2008; Arndt-Lappe, 2015; Fam and Lepage, 2018, 2021), for example, *acteur=N:action=N::animateur=N:animation=N*. These analogies are directly acquired from the morphological definitions and morphological sections of GLAWI. On the other had, sets of relations such as (4) are made up of two sets of lexemes (the left and right columns) that exhibit morphologically relevant regularities. For example, all words in the left column of (4) contain a final sequence `eur`, all words in the right one contain a final sequence `ion`. Moreover, these sequences are morphologically relevant because the stem of the lexeme pairs in each line are identical (for example, we have the same stem `animat` in the two lexemes of the second line). Glawinette is also distinguished by its ability to describe the morphological series by means of "natural" patterns that are very similar to the ones used by linguists to characterize complex lexemes. For example, the relation *activiste=N:activisme=N* will be characterized by the pattern `^(.+)iste$=N/^(.+)isme$=N` and not `^(.+)te$=N/^(.+)me$=N` nor `^(.+)t(.+)$=N/^(.+)m(.+)$=N`. Glawinette contains 97 293 lexemes connected by 47 712 relations. These relations are divided into 15 904 morphological families and 5 400 series. Note that some of the relations described in Glawinette are already present in Démonette. This intersection will be used to complement the morphological descriptions of the relations in Glawinette.

3 Some "practical" consequences of the inclusion of Glawinette in Démonette

Glawinette will provide Démonette with more complete and varied morphological families. The families of Glawinette contain a large number of relations not yet covered in Démonette. This integration will test the capacity of the database architecture to describe a representative fragment of French morphological relations, which potentially are more complex than the ones currently described in Démonette. For example, they contain derivationally distant pairs such as *déformer=V:indéformable=A* 'to deform:undeformable' where *déformer* is a second level ascendant of *indéformable* (*déformer* → *déformable* 'deformable' → *indéformable*). This type of relation is hardly present in the current version of Démonette. Their inclusion will test the robustness of the tagsets used in Démonette.

The other interesting feature of the Glawinette relations is that they are semantically relevant as they are directly derived from definitions (and morphological sections). However, the relations of Glawinette, like the ones from other resources used to create Démonette, do not contain semantic characterization. Completion of these descriptions will be the main challenge in integrating Glawinette into

Démonette. Several paths will be explored for these descriptions. On the one hand, there will be a semi-automatic shallow completion at the level of the series of specific relations. For example, we can specify that the *-ion* derivatives in (4) are action nouns and propose for the pair *formateur=N:formation=N* 'trainer:training' a gloss such as 'a trainer carries out a training' which could be later completed in 'a trainer carries out a training of people to whom he teaches new skills'. Another strategy will take advantage of clusters of relations within the families to leverage the semantic descriptions of some of them, e.g., to predict the gloss of an indirect relation, or cross-formation (Becker, 1993), or that of a complex relation (e.g. *déformer:indéformable*) from existing, base-to-derivative glosses, defining the lexemes involved in these indirect and complex relations. For example, the pair *déformer:indéformable* can be glossed as 'something undeformable cannot be deformed' by superposition and adaptation of the glosses of the direct relations *déformable:indéformable* 'what is undeformable is not deformable' and *déformer:déformable* 'something deformable can be deformed'. The integration of the pairs from Glawinette will also involve a revision of the exponents of the morphological processes. For example, the pattern `^re(.+)er$=V/^(.+)er$=V` will be replaced by the pattern `^re(.+)$=V/^(.+)$=V` which is a more appropriate level of generalization as prefixation in *re-* is not limited to verbs of the first conjugation (with infinitives ending in *-er*). Series is the right level of granularity to make this kind of decision because it gathers homogeneous sets of similar relations. Moreover, families give a more complete view of all the specific derivational relations that hold between its lexemes.

4 Feeding Démonette with relations from Glawinette

The series of Glawinette will be integrated into Démonette one by one. These series are characterized by their yield (that is, the number of lexeme pairs they contain) and by the properties of the patterns that define them: the (cumulative) length of the patterns, the specificity of the exponents (i.e. the ratio of the number of words that match a pattern in the whole lexicon to the number of pairs contained in the series of relations, (Bybee, 1988)), and their versatility (i.e. the overall number of connections of the lexemes identified by the pattern). These features enable us to estimate the quality of the pairs contained in a series, to process the most reliable ones first and to devote more resources for the ones that are likely to contain errors. For example, the series `^(.+)er$=V/^(.+)age$=N` contains 1465 pairs that normally contains no errors. Conversely, the series `^(.+)tte$=N/^(.+)lle$=N` contains only the erroneous pair *batte=N:balle=N* 'bat:ball'. The very small size of the stem (`ba` contains only 2 characters) is an additional clue to this error. However, not all series that contain few pairs are incorrect, especially the ones with sufficiently long patterns like `^(.+)anisme$=N/^(.+)éen$=A` which only contains *européanisme=N:européen=A* 'europeanism:european'. By combining a number of such criteria, we can quickly identify potentially erroneous pairs and series that are most cost-effective to include in Démonette.

5 New theoretical questions

The inclusion of Glawinette in Démonette also contributes to the debate on several current theoretical issues in morphology. The families and series of Glawinette are the source material from which morphological paradigms can be built. The creation of these paradigms from the morphological series remains an open question that Glawinette will help clarify. They will lead to complement the architecture of Démonette with additional tables that will represent this paradigmatic organization (morphological families, morphological paradigms). This is not a trivial evolution because these structures are defined on top of multiple, redundant and unconstrained relational descriptions. At first, we will only include the derivational relations from Glawinette.

Various future decisions regarding the relations encoded in Démonette will be reconsidered with respect to the content of Glawinette. First, the relations in Démonette are symmetrical by design, whether direct base-to-derivative, complex ancestor-to-descendent or indirect between siblings. Whenever Démonette contains an entry (word1, word2), it also includes the corresponding (word2, word1) entry described by means of feature values that are symmetrical to the ones of (word1, word2). However, we observe that the morphological relations originating from the GLAWI dictionary are not symmetrical, and this will

lead us to rethink the conditions of the systematic symmetrization of the entries in Démonette.

Second, the presence in Glawinette of lexeme pairs that are in complex relations like `^(.+)er$=V/^in(.+)able$=A` confirms the relevance of this type of relations and validates their description in Démonette. Moreover, these pairs empirically validate the intuition of speakers who unconsciously reanalyze these sequences as affixes in their own right (see Stump (2017, 2019) for a theoretical account of this phenomenon he calls "rule conflation").

On the other hand, the observation of indirect relations in Glawinette questions the systematic description of all indirect relations in Démonette. For example, the series `^(.+)eur$=N/^(.+)ion$=N` contains only 285 pairs in Glawinette when the series `^(.+)er$=V/^(.+)ation$=N` contains 1322 ones. Yet when a verb is the base of an action noun in -ation, then it should also be the base of an agent noun in -ateur: therefore, we would have expected similar figures for the two series. The integration of Glawinette thus leads to two questions: (*i*) explain the shift; (*ii*) account for it in Démonette, for example by completing the graphs (i.e. the families) on the fly according to users' wishes.

Finally, Glawinette may call into question theoretical certainties about the indentity of rule exponents. For instance, Glawinette contains 122 `^(.+)er$=V/^(.+)ion$=N` pairs compared to the 1322 `^(.+)er$=V/^(.+)ation$=N` series above; this calls into question the common conception that *-ation* is an allomorphic variant of *-ion* where the sequence /at/ is part of the verb stem(Bonami et al., 2009). In view of these numbers, it seems legitimate to consider *-ation* as an exponent in its own right and to adapt the description of these derivatives in Démonette accordingly. Conversely, the relations in Glawinette are essentially determined by the formal regularities that exist in the lexicon. Their inclusion in Démonette will impose to dissociate their formal, categorical and semantic components and will highlight the multiplicity of the possible generalizations.

6 Perspective

In the short term, we plan to integrate most of the relations of Glawinette into Démonette, which will significantly increase the number of entries in the database and the diversity of indirect and complex relations. This extension will provide additional material to conduct experimental and quantitative morphology experiments. The next step will be to exploit the definitions in GLAWI to generate glosses for the lexeme pairs in Glawinette. These glosses will then be used to feed the semantic section of Démonette. Finally, we plan to build a phonological version of Glawinette by combining the phonological transcriptions in GLAWI and in the lexeme table of Démonette in order to the characterize phonological operations and provide phonological patterns that will be used to complement the phonological fields of Démonette.

References

Sabine Arndt-Lappe. 2015. Word-formation and analogy. In Ingeborg Müller, Peter O.and Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation. An international handbook of the languages of Europe*, de Gruyter Mouton, Berlin/Boston, pages 822–841.

Laurie Bauer. 1997. Derivational paradigms. In *Yearbook of Morphology 1996*, Springer, pages 243–256.

Thomas Becker. 1993. Back-formation, cross-formation, and 'bracketing paradoxes' in paradigmatic morphology. *Yearbook of Morphology* 1992:1–25.

Harry Bochner. 1993. *Simplicity in generative morphology*. Mouton de Gruyter, Berlin & New-York.

Olivier Bonami, Gilles Boyé, and Françoise Kerleroux. 2009. L'allomorphie radicale et la relation flexion-construction. In Bernard Fradin, Françoise Kerleroux, and Marc Plénat, editors, *Aperçus de morphologie du français*, Presses universitaires de Vincennes, Saint-Denis, pages 103–125.

Olivier Bonami and Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2):167–197.

Joan L. Bybee. 1988. Morphology as lexical organization. In Michael Hammond and Michael Noonan, editors, *Theoretical Morphology. Approaches in Modern Linguistics*, Academic Press, San Diego, CA, pages 119–141.

Rashel Fam and Yves Lepage. 2018. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, page 1060–1066.

Rashel Fam and Yves Lepage. 2021. A study of analogical density in various corpora at various granularity. *Information* 12(8).

Bernard Fradin. 2018. Paradigms and the role of series in derivational morphology. *Lingue e Linguaggio* 2/2018:155–172.

Nabil Hathout. 2008. Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the Coling workshop Textgraphs-3*. ACL, Manchester, pages 1–8.

Nabil Hathout. 2011a. Morphonette: a paradigm-based morphological network. *Lingue e linguaggio* 2011(2):243–262.

Nabil Hathout. 2011b. Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*. In *Des unités morphologiques au lexique*, Hermès Science-Lavoisier, Paris, pages 251–318.

Nabil Hathout and Fiammetta Namer. 2014. La base lexicale Démonette : entre sémantique constructionnelle et morphologie dérivationnelle. In *Actes de la 21e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2014)*. ATALA, Marseille, pages 208–219.

Nabil Hathout and Fiammetta Namer. 2016. Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Nabil Hathout and Fiammetta Namer. 2018. Defining paradigms in word formation: concepts, data and experiments. *Lingue e Linguaggio* 17(2):151–154.

Nabil Hathout and Fiammetta Namer. 2019. Paradigms in word formation: what are we up to? *Morphology* 29(2):153–165.

Nabil Hathout and Franck Sajous. 2016. Wiktionnaire's Wikicode GLAWIfied: a workable French machine-readable dictionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Nabil Hathout, Franck Sajous, Basilio Calderone, and Fiammetta Namer. 2020. Glawinette: a linguistically motivated derivational description of French acquired from GLAWI. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, pages 3870–3878.

Philippe Langlais and François Yvon. 2008. Scaling up analogical learning. In *Proceedings of the 22nd international conference on Computational Linguistics (COLING 2008)*. Manchester, UK, pages 51–54.

Yves Lepage. 1998. Solving analogies on words: An algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics*. Montréal, volume 2, pages 728–735.

Yves Lepage. 2004. Lower and higher estimates of the number of true analogies between sentences contained in a large multilingual corpus. In *Proceedings of the 20th international conference on Computational Linguistics (COLING-2004)*. Genève, pages 736–742.

Robert Martin. 1983. *Pour une logique du sens*. Linguistique nouvelle. Presses universitaires de France, Paris.

Fiammetta Namer, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout, and Delphine Tribout. 2019. Demonette2 - une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de la 26e conférence annuelle sur le traitement automatique des langues naturelles (TALN 2019)*. Toulouse, pages 233–243.

Fiammetta Namer and Nabil Hathout. 2020. ParaDis and Démonette – from theory to resources for derivational paradigms. *The Prague Bulletin of Mathematical Linguistics* 114:5–33.

Michel Roché. 2009. Pour une morphologie *lexicale*. In *La morphologie lexicale est-elle possible ?*, Éditions Peeters, Leuven, volume 17 of *Mémoires de la Société de Linguistique, Nouvelle Série*, pages 65–87.

Franck Sajous and Nabil Hathout. 2015. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*. Herstmonceux, England, pages 405–426.

Nicolas Stroppa and François Yvon. 2005. An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*. ACL, Ann Arbor, MI, pages 120–127.

Gregory Stump. 2017. Rule conflation in an inferential-realizational theory of morphotactics. *Acta Linguistica Academica* 64:79–124.

Gregory T. Stump. 2019. Some sources of apparent gaps in derivational paradigms. *Morphology* 29(2):271–292.

Jaap Van Marle. 1985. *On the Paradigmatic Dimension of Morphological Creativity*. Foris, Dordrecht.

Pavol Štekauer. 2014. Derivational paradigms. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, Oxford, Oxford University Press, Oxford, pages 354–369.

The MoNoPoli database
Or how to catch *Macron-itis*

Mathilde Huguin
ATILF (CNRS-UL) UMR 7118 / Nancy, France
mathilde.huguin@univ-lorraine.fr

Abstract

In this article 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 alternative 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.

1 Introduction

We study French deanthroponyms, i.e. words morphologically built on proper names (Schweickard, 1992; Leroy, 2008; Schlücker and Ackermann, 2017) 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 intentionally to meet an immediate need in a given context. In (1a) and (1b), the forms *Macronite* ('Macron-itis') and *aubrycratie* ('aubry-cracy') 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. *Il faut dire à ces militants de ne pas confondre démocratie et* **aubrycratie**. (*Martine Aubry*)
'It is necessary to say to these militants not to confuse democracy and aubry-cracy.'

As these derivatives are most often neologistic nonce-formations, the constitution of the corpus required the elaboration of a specific methodology which we describe in this article. This methodology consists of two steps. First, we automatically generate hypothetical deanthroponym *candidates* (e.g. *françoishollandien*, 'françoisholland-ian', *lepenphobe*, 'lepen-phobic') using the 89 PPNs and 90 suffixes of Huguin (2018). We then look up these candidates in context, using the Web. We also present some of the results we obtained focusing on two types of atypical constructions which characterize deanthroponyms.

Our presentation will be structured as follows. We explain why the Web is the most likely resource to contain deanthroponyms (§2). We discuss two types of possible strategies to collect them and evaluate

Each example of deanthroponym is provided with its context; the base PPN is indicated between brackets and the stem/suffix boundary is marked by a hyphen in the English translation.

Both lists are available online at this address: https://perso.atilf.fr/mhuguin/accueil/these/documents/.

their theoretical implications (§3). We present the architecture of the program that allowed their generation (§4) and the results of our collection (§5). Finally, we present some of the results that emerged from the analysis of our data (§6).

2 Where to find data

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

The list of PPNs used to generate candidates contains names of politicians who have held a prominent position in French politics (Presidents, leaders of political party, etc.). They have occupied this role since 1981, so they are contemporary referents (e.g. Jacques Chirac, Emmanuel Macron). 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).

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, 2015) as in (2). In (2a), alongside their morphological creations *royaliste* ('follower of Ségolène Royal') and *montebourgeois* ('follower of Arnaud Montebourg'), the writer inserts a meta-discursive comment: "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 (X*phobe*$_{\text{Adj}}$), 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).*
'*Lose his mind*, a militant blog. Long time royal-ist, now montebourge-ian (I don't know if that's how you say it).'

b. *Il y a de quoi venir phobe :* **hollandephobe**, **vallsphobe**, **taubiraphobe**, **belkacemophobe**, *gauchophobe, antifaphobe, imamophobe, racaillophobe.* (*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, gangstero-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, 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.

Lüdeling et al. (2007); Fradin et al. (2008); 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 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.

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.

3 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. *Christiane Taubira*, *Emmanuel Macron*, *Jacques Chirac*, *Jean-Marie Le Pen* [...]
b. *able*, *erie*, *ification*, *isme*, *isterie*, *ix*, *logue*, *mètre*, *oïa*, *thon*, *us* [...]

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 example, this strategy leads to build exclusively *taubirie* /tobiri/ ('taubir-land') from the inputs *Taubira* /tobira/ (from *Christiane Taubira*) and *-ie* /i/ so as to: **(i)** avoid the hiatus /ai/ proscribed by the markedness constraint *HIATUS (/tobira/ + /i/ = */tobirai/), **(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 well-formedness 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 /rari/ from /tobirari/, which corresponds to the attested form *Taubirarie* 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). 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 *Sarkozy* (from PPN *Nicolas Sarkozy*) and the French suffix *-able*.

(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.'

(5) *sarkozyable*, *sarkozable*, *sarkozyssable*, *sarkozysable*, *sarkozytable*, *sarkozylable*, *sarkozydable* [...]

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 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.

4 Generating derived forms

We run our candidate generation program on all the graphical forms that realize each PPN in our list and all suffixes from our set. 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. 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, 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 used if the referent is a woman (6).

(6) a. *On a déjà assez à faire pour lutter contre la* **najatisation** *de l'enseignement !* (*Najat Vallaud-Belkacem*)
'We are already busy enough fighting against the najat-ization of education!'

b. *Finalement le clientélisme et le* **clémentinisme** *se rejoignent.* (*Clémentine Autain*)
'Finally, clientelism and clémentin-ism come together.'

Sub-names	**Examples**	**Derived forms**	**Gloss**
Last name	Strauss-Kahn	*strausskahnité*	'strausskahn-ity'
First name	Dominique	*dominiqueur*	'dominiqu-er'
Full name	Dominique Strauss-Kahn	*dominiquestrausskahnien*	'dominiquestrausskahn-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 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 $\texttt{Stem}_{\texttt{i}}^{\texttt{n}}$' and a suffix $\texttt{Suff}_{\texttt{j}}$ (7a) ($0 < \texttt{j} \leq 90$). For a given $\text{PPN}_{\texttt{i}}$ ($0 < \texttt{i} \leq 89$), the symbol $\texttt{Stem}_{\texttt{i}}^{\texttt{n}}$' corresponds to the `stem` of one of its sub-names `n` ($0 < \texttt{n} \leq 6$), or consists of a variation of this `stem` (7b) produced by one of the 36 $\Re$ rules of the program.

(7) a. $\texttt{Candidate} = \texttt{Stem}_{\texttt{i}}^{\texttt{n}}\text{'} \oplus \texttt{Suff}_{\texttt{j}}$

b. $\begin{cases} \texttt{Stem}_{\texttt{i}}^{\texttt{n}}\text{'} = \texttt{Stem}_{\texttt{i}}^{\texttt{n}} \\ \texttt{Stem}_{\texttt{i}}^{\texttt{n}}\text{'} = \Re(\texttt{Stem}_{\texttt{i}}^{\texttt{n}}\text{'}) \end{cases}$

Each of them selects two arguments: $\texttt{Stem}_{\texttt{i}}^{\texttt{n}}$' et $\texttt{Suff}_{\texttt{j}}$. Rules are organized in four blocks, cf. Figure 1.

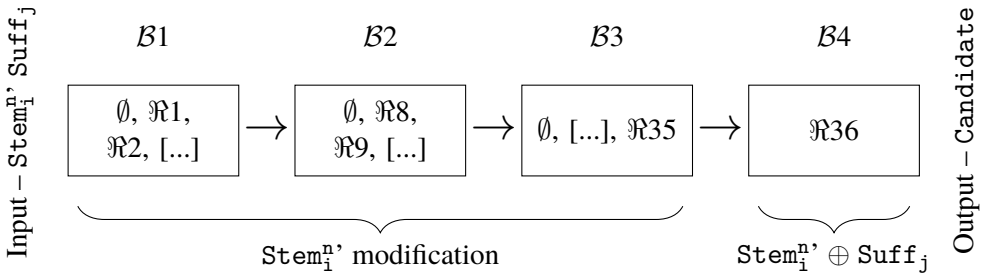


Figure 1: Rule combinations

The rules in the block $\mathcal{B}1$ remove a graphical sequence from $\texttt{Stem}_{\texttt{i}}^{\texttt{n}}$'. The rules of $\mathcal{B}2$ add a graphical sequence to $\texttt{Stem}_{\texttt{i}}^{\texttt{n}}$'. 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 $\Re$36 rule of $\mathcal{B}4$ concatenates the inputs $\texttt{Stem}_{\texttt{i}}^{\texttt{n}}$' and $\texttt{Suff}_{\texttt{j}}$. The program

is oriented and acyclic. In each block, the rules are in complementary distribution. The values of $\texttt{Stem}_{\texttt{i}}^{\texttt{n}}$' et $\texttt{Suff}_{\texttt{j}}$ 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 $\Re 36$ of concatenation in $\mathcal{B}4$ gives the `Candidate` (8). In contrast, the second combination using $\Re 1$ in $\mathcal{B}1$ will be discarded by the program. $\Re 1$ 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. $\Re 2$, can be applied, since inputs `taubira/ique` satisfy the conditions for application of the $\Re 2$ truncation rule: `taubira` ends with a vowel and `ique` begins with a vowel. $\Re 2$ 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 $\Re 2$ is given as input to $\Re 36$ 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) $\Re 36$(`taubira`, `ique`) = `taubiraique`

(9) a. $\Re 2$(`taubira`, `ique`) = (`taubir`, `ique`)
b. $\Re 36$(`taubir`, `ique`) = `taubirique`

(10) $\Re 8$(`taubira`, `ique`) = (`taubirat`, `ique`)
$\Re 36$(`taubirat`, `ique`) = `taubiratique`

(11) $\Re 9$(`taubira`, `ique`) = (`taubiras`, `ique`)
$\Re 36$(`taubiras`, `ique`) = `taubirasique`

5 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, e.g. *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*$_{\text{Nc}}$ '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 permettant à l'INRAP (institut national de recherches archéologiques préventives) d'effectuer des fouilles de sauvetage sur des sites archéologiques menacés par des programmes*

immobiliers. (*Jacques Chirac*, *Jean-Pierre Raffarin*)
'The chiraco-raffarinian parliamentary majority suppresses a good part of the financial means allowing the INRAP (national institute of preventive archaeological researches) to carry out excavations of rescue on archaeological sites threatened by real estate programs.'

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 the 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 presentation of all the information contained in MoNoPoli is beyond the scope of this presentation. However, the reader will be able to find an excerpt of the database and an explanation of each feature online.

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	/pejɔnad/
Rama Yade	*crainte d'une* **ramayadisation** 'fear of ramayad-ization'	Xisation	Nc	/ramajadizasjɔ̃/
François Hollande	*le chimpanzé à cul rose* **homo hollandus** 'the pink ass chimpanzee homo holland-us'	homoXus	Nc	/homoolãdys/

Table 2: Exerpt of MoNoPoli

6 Analysis of deanthroponyms: some remarkable results

Unsurprisingly, the PPNs most often used as bases in MoNoPoli (14a) 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). Jacques Chirac (14b) was President of the French Republic. Dominique Strauss-Kahn (14c) 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. The 13 PPNs in (14a) are the bases of 50% of the deanthroponyms in our corpus.

(14) a. *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*

b. *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.'

c. *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.'

The nonce-formations are identifiable thanks to the meta-discursive signals and the discursive processes (cf. §2) but also sometimes thanks to the morphological processes used. 10% of the deanthroponyms of MoNopoli are produced by extragrammatical processes (Dressler and Kilani Schoch, 2005). We provide an overview of the morphological diversity of the content of MoNoPoli in Table 3. The most frequent types of constructs are formed by suffixation and composition. Derivation by conversion is the least represented, which is certainly an effect of our methodology. Indeed, we only looked for forms corresponding to infinitive verbs.

The excerpt is available at this address: https://perso.atilf.fr/mhuguin/accueil/these/corpus-monopoli/.

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

Table 3: Overview of the processes of MoNoPoli

In the following, we show that PPNs are privileged bases for extragrammatical derivation on several levels. On the one hand, they serve as a bases for processes classified as extragrammatical in French such as reduplication, cf. §6.1. On the other hand, they also sometimes lead to the subversion of more classical i.e., more regular processes. We illustrate our remarks by studying the case of compounding in §6.2.

6.1 Extragrammatical process

MoNoPoli contains 23 deanthroponyms that instantiate the pattern X*o*X*suff* as complex words of (15) when X is the stem of a sub-name of PPN. In (15a) *suff* is *-iste* ('-ist'), in (15b) it is *-ien* ('-ian').

(15) 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 sphera.'

At first sight, the process used to obtain these derivatives is compounding. One could see in *sarkozo-sarkozyste*$_{Adj}$ and *chiraco-chiraquien*$_{Nc}$ a particular case of the compound X*o*Y*suff*. For example, the demonym *franco-canadien*$_{Adj/Nc}$ 'French and Canadian' is compounded from the bases X *français*$_{Adj/Nc}$, which appears truncated, et Y *canadien*$_{Adj/Nc}$ (Dal and Amiot, 2008). As in *franco-canadien*$_{Adj/Nc}$, we find in *sarkozo-sarkyste* the intercalary vowel /o/ typically associated with compounds, ans more specifically *learned* compounds, i.e. including the stem of a lexeme inherited from Greek or Latin. In *franco-canadien*$_{Adj/Nc}$, 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 (15), 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*$_{Adj}$ and *chiraco-chiraquien*$_{Nc}$ as compounds. The second obstacle is semantic. The meanings of deanthroponyms of (15) 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* (15a) means 'very/typically/exclusively sarkozist' and *chiraco-chiraquiens* (15b) 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 (15) 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 (16) 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 *compound salad*). These paraphrases have semantic values equivalent to those we have determined for the complex words of (15). 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'.

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

6.2 Subverted grammatical process

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. (17).

(17) 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 soi-disant 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 so-called 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 (17a) 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 observed in §6.1. 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 (17b), 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 (17a). So, X*o* constituents are truncated forms of deanthroponyms of the same nature as Y*suff* *bessonisme*$_{\text{Nc}}$ ('besson-ism') which refers to the ideology of Eric Besson. They are truncated forms of the common nouns *pasquaïsme*$_{\text{Nc}}$ ('pasqua-ism') and *sarkozysme*$_{\text{Nc}}$ ('sarkoz-ism').

- The examination of (17c), 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 a outbursts, to distinguish himself (cf. §2).

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.

7 Conclusion

The method used to create the MoNoPoli database is reproducible and adaptable to other languages or other inputs (bases or affixes). The database created will eventually be accessible online. It provides a large corpus of contextualized deanthroponyms, which to our knowledge does not exist in French. Moreover, MoNoPoli presents a real morphological diversity, i.e. a large panel of different processes. It is a possible resource for further research on both words based on anthroponyms and French nonce-formations. Analysing it reveals that deanthroponyms are often nonce-formations constructed by extragrammatical processes. We have also shown that grammatical processes can be subverted to satisfy the enunciative needs of the writer. The latter demonstrates, at the same time, their epilinguistic capacity to play with language.

References

Laurie Bauer. 1983. *English Word-Formation*. Cambridge Textbooks in Linguistics. Cambridge University Press. https://doi.org/10.1017/CBO9781139165846.

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

Georgette Dal and Fiammetta Namer. 2015. Internet. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation:An International Handbook of the Languages of Europe*, De Gruyter Mouton, Berlin, New York, volume 3, pages 2372–2386. https://doi.org/10.1515/9783110375732-044.

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

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

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

Jila Ghomeshi, Ray Jackendoff, Nicole Rosen, and Kevin Russell. 2004. Contrastive Focus Reduplication in English (The Salad-Salad Paper). *Natural Language & Linguistic Theory* 22(2):307–357.

John Goldsmith. 1976. *Autosegmental phonology*. Ph. Dissertation, MIT.

Christophe Gérard. 2018. Variabilité du langage et productivité lexicale: Problèmes et propositions méthodologiques. *Neologica* 12:23–45.

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

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

Peter Hohenhaus. 2007. How to do (even more) things with nonce words (other than naming). In Judith Munat, editor, *Lexical Creativity, Texts and Contexts*, John Benjamins Publishing Company, Amsterdam, Philadelphia, Studies in Functional and Structural Linguistics 58, pages 15–38.

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

Mathilde Huguin. 2018. Anthroponyms and paradigmatic derivation in French. *Lingue e Linguaggio* XVII(2):217–232. https://doi.org/10.1418/91866.

Daphné Kerremans. 2015. *A Web of New Words: A Corpus-Based Study of the Conventionalization Process of English Neologisms*. Peter Lang GmbH, Internationaler Verlag der Wissenschaften. https://doi.org/10.1515/ang-2016-0076.

Georges Kleiber. 1990. *La Sémantique du prototype. Catégories et sens lexical*. PUF, Paris.

Anke Lüdeling, Stefan Evert, and Marco Baroni. 2007. Using web data for linguistic purposes. In Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, editors, *Corpus Linguistics and the Web*, Rodopi, Amsterdam, New York, 59, pages 7–24. https://doi.org/10.1163/9789401203791_003.

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

Stéphanie Lignon, Fiammetta Namer, Nabil Hathout, and Mathilde Huguin. 2019. 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.

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

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

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

Michel Roché. 2010. Base, thème, radical. *Recherches linguistiques de Vincennes* 39:95–134. https://doi.org/10.4000/rlv.1850.

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

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

Barbara Schlücker and Tanja Ackermann. 2017. The morphosyntax of proper names: An overview. *Folia Linguistica* 51(2):309–339. https://doi.org/10.1515/flin-2017-0011.

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

Ludovic Tanguy. 2012. *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.

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

Developing a resource for *-ance* nouns, and related verbs and adjectives

Marie Laurence Knittel
Université de Lorraine
ATILF - CNRS
23, Boulevard Albert 1°
BP 60446
54 001 Nancy Cedex
marie-laurence.knittel@univ-lorraine.fr

Rafael Marín
CNRS & Université de Lille
STL - UMR 8163
Rue du Barreau
BP 60149
59653 Villeneuve d'Ascq
rafael.marin@univ-lille.fr

Abstract

This article introduces the AdVeNance[1] resource, which includes 112 French nouns ending in -ance (Nance) (e.g. *résistance* 'resistance'), extracted from Lexique3 (New et al. 2001), and their related verbs and/or adjectives, and how it enabled us to verify the Extended Boundedness Hypothesis, an enlarged version of the Boundedness Hypothesis (Jackendoff, 1991).

We describe the procedure we followed to extract relevant data from Lexique3 and the tests we used to pair Nance with their relevant bases. The correlation between the mass/count properties of Nance, the (a)telicity of the related verbs, and the open/closed scale of the related adjectives is then discussed in detail.

Our results show that over 90% of Nance are mass. Mass Nance are mostly related to stative verbs and unbounded adjectives, in line with the Extended Boundedness Hypothesis. As for count Nance, all are related to telic verbs, but a significant number of them are unexpectedly paired with non-degree (vs bounded) adjectives. Therefore, the EBH is only partially confirmed by count Nance.

So as to expand AdVeNance, we began to examine nouns in *-ence* (Nence, e.g. *préférence* 'preference'). A preliminary analysis of these nouns and their verbal bases reveals that, similarly to Nance, most Nence are mass, and most of mass Nence derive from stative verbs.

1 Introduction

French *-ance* nominals (Nance) constitute a relatively small but most interesting noun class, since they can be related to verbs (1a), adjectives (1b) or both (1c) (Dal & Namer 2010, Knittel 2016).

(1) a. *appartenance*$_N$ / *appartenir*$_V$ 'belonging' / 'to belong'
b. *constance*$_N$ / *constant*$_{Adj}$ 'consistency', 'steadyness' / 'constant'
c. *abondance*$_N$ / *abonder*$_V$ / *abondant*$_{Adj}$ 'abundance' / 'to abound' / 'abundant'

As a consequence, they constitute a useful set of data to study the possible semantic relationship between three categories. In particular, they allow to check an extended version of the Boundedness Hypothesis (Bach, 1976; Mourelatos, 1978; Jackendoff, 1991; Brinton, 1998), which takes into account not only nouns and verbs, but also adjectives.

There is a broad agreement in the linguistic community that each category prototypically conveys a certain type of meaning. Verbs would then denote eventualities, nouns entities and adjectives properties. Each of these concepts is in turn characterized by a typical semantic property. Verbs fall into aspectual classes; nouns can be mass or count, and adjectives, when gradable, are either bounded or unbounded.

[1] The AdVeNance project is supported by the Maison des Sciences de l'Homme de Lorraine (MSHL-USR 3261).

In its original version, the Boundedness Hypothesis (Bach, 1976; Mourelatos, 1978) predicts a parallelism between verbal and nominal properties. More specifically, it predicts that mass nouns, as describing unbounded entities, should be related to atelic verbs, that describe unbounded events; conversely, count nouns parallel telic verbs, both describing bounded entities.

The Extended Boundedness Hypothesis, that we propose here, adds adjectives (Paradis, 2001) to the initial hypothesis, and predicts that open-scale adjectives should be paired with mass nouns and atelic verbs, while closed-scale adjectives are paired with count nouns and telic verbs. In line with Gumiel-Molina et al. (2020), we assume that closed-scale adjectives are those having a closed upper bound, i.e. a scale with a final boundary.

This work aims at verifying empirically the Extended Boundedness Hypothesis by means of the analysis of a significant number of Nance nouns, extracted from Lexique 3, and their related verbs and/or adjectives. The annotation of these nouns, verbs and adjectives has led to the elaboration of AdVeNance, a morpho-semantic resource that will be described here.

In the following section, we describe the bases of AdVeNance; particularly, the selection of Nance (ant their related verbs and adjectives), and the annotation of their relevant properties. In section 3, we present the main results we have found. In section 4, we include a comparison between Nance and a new group of nouns, those ending in *-ence* (Nence), that we will also include in AdVeNance. Finally, section 5 summarizes our main findings and points out some questions for further research.

2 The AdVeNance resource

2.1 Aims

The aim of the AdVeNance resource is to provide a list of Nance with the semantically related verbs and/or adjectives (1), annotated according to the properties relevant to the Extended Boundedness Hypothesis: nominal countability, verbal aspect and adjectival scalarity. The resource appears as a database providing columns presenting each category and its relevant feature, as shown in Table 1.

Nouns	Countability	Verbs	Aspect	Adjectives	Scalarity
confiance	mass			*confiant*	closed-scale
persistance	mass	*persister*	state	*persistant*	open-scale
vengeance	count	*venger*	achievement		

Table 1. Sample of the AdVeNance resource.

2.2 Nance selection

The nouns that we analyzed were extracted from the resource Lexique (New et al. 2001). From a total of 244 Nance listed in Lexique, we had to discard the nouns that were irrelevant for our study. Thus, we excluded non-suffixed nouns (2a), nouns without morphological relation with verbs or adjectives (2b), nouns build on other Nance by prefixation (2c), and spelling doublets (2d).

(2) Sample of discarded nouns
a. *chance* 'luck', *substance* 'substance'
b. *délinquance*$_N$ 'delinquency' > *délinquant*$_N$ 'offender'
c. *auto-surveillance* 'self-surveillance' > *surveillance* 'surveillance'
d. *becquetance / bectance* 'food, meal'

2.3 Matching Nance with their base(s)

The first annotation step was to pair the Nance kept at the end of the above selection with their related category in an appropriate manner, so as to discard improper pairs or triplets.
To do so, we used the tests provided in the literature. On the one hand, to identify Nance related to adjectives, we used the tests provided by Rainer (1989), Van de Velde (1995), Flaux & Van de Velde (2000), Beauseroy (2009). On the other hand, Nance related to verbs were identified using tests from Grimshaw (1990), Melloni (2007), Balvet et al. (2012), Fradin (2011, 2012, 2014), Kerleroux (2012).

In doing so, we were led to check if the formal closeness between Nance and the potential verbal or adjectival basis reflected a regular semantic pattern, which was not always the case. For example, *ambiance* 'atmosphere' does not react in the appropriate manner to the above tests, and cannot be easily related to *ambiant*$_{Adj}$, 'ambient', 'surrouding'[2]. That is why we choose to discard this pair and similar ones.

Another necessary step, as far as adjectives are concerned, was to identify and eliminate forms that behave rather as present participles or as nouns. For example, in the case of *gérance* 'stewardship', the potential base *gérant* does not qualify as an adjective, but either as a noun or as the present participle of *gérer* 'to manage'.

At the end of this stage, we gathered a list of 112 nouns, among which 72 are related to verbs, 97 to adjectives, and 56 to both, as shown in Table 2.

Nance [112]	Related to V	Non-related to V	Total
Related to Adj	56	41	97
Non-related to Adj	16	—	
Total	72	41	

Table 2. Distribution of Nance and their related categories

The availability of triplets (N/V/Adj) confirmed previous observations (Dal & Namer, 2010; Knittel, 2016), that Nance can be related to verbs and/or adjectives. Although in a reduced number, the nominals of the AdVeNance resource constitute a reliable list of nouns paired with their related verbs and/or adjectives.

2.4 Annotation

After the matching of the 112 Nance with their corresponding verb and/or adjective, we proceeded to the examination of their relevant characteristics, namely mass/count opposition for nouns, aspect for verbs and scalarity for adjectives.

The mass/count distinction was annotated following a methodology similar to that of Dugas et al. (2021), mostly applying the same tests. Unlike mass nouns, count nouns accept plurals (3a), count quantifiers (3b) and definite numerals (3c). On the other hand, unlike count nouns, mass nouns allow partitive articles (4a), as well as modification by intensifiers (4b).

(3) *table* ‘table’
  a. *tables* ‘tables’
  b. *plusieurs tables* ‘several tables’
  c. *trois tables* ‘three tables’

(4) *joie* 'happiness'
  a. *de la joie* ‘lit. of the happiness’
  b. *beaucoup de joie* ‘a lot of happiness’; *une joie intense* ‘an intense happiness’

On the other hand, verbs have been annotated with respect to the four Vendlerian classes (states, activities, accomplishments and achievements). As illustrated in Table 3, these four classes are characterized by means of three features: dynamicity, telicity and duration. Thus, states are the only verbs denoting non-dynamic situations; both states and activities are atelic, while accomplishments and achievements are telic; as for achievements, they are the only verbs denoting punctual (i.e. non durative) situations.

[2]The definitions found in the CNRTL (https://www.cnrtl.fr) confirm this discrepancy.
AMBIANT: Qui entoure ou circule autour, qui environne. [Engl. *ambient, surrouding*]
AMBIANCE: Qualité du milieu (matériel, intellectuel, moral) qui environne et conditionne la vie quotidienne d'une personne, d'une collectivité. [Engl. *atmosphere*]

	Dynamicity	Telicity	Duration
State	–	–	+
Activity	+	–	+
Accomplishment	+	+	+
Achievement	+	+	–

Table 3. Aspectual features of the Vendlerian verb classes.

Regarding the annotation of verbs, we used a battery of standard aspectual tests (Dowty, 1979), following a general procedure as the one described in Balvet et al. (2012).

States (*préférer* 'to prefer'), unlike dynamic predicates, are not compatible with the progressive form *être en train de* 'to be V-ing' (5a), and are not good answers either to questions of the type – *Que s'est-il passé hier?* 'What happened yesterday?' (5b).

(5) States
- a. **Il est en train de préférer les bettes.* lit.: 'He is preferring chard.'
- b. *– Que s'est-il passé hier?* 'What happened yesterday?'
  **– Il a préféré les bettes.* lit.: – He preferred chard.'

Atelic predicates are only compatible with *for x time* modifiers (6a), while telic predicates are compatible with *in x time* modifiers (6b). For similar reasons, telic predicates combine with expressions such as 'to take x time to V' (6c).

(6) Telic vs atelic predicates
- a. *Il s'est promené {pendant/*en} trois heures.* 'He walked {for/*in} three hours.'
- b. *Elle a réparé la voiture en trois heures.* 'She repaired the car in three hours.'
- c. *Il m'a fallu trois heures pour réparer la voiture.* 'It took me three hours to repair the car.'

Finally, among telic predicates, achievements are not compatible with the aspectual semi-auxiliaries *continuer* 'to keep on' or *arrêter* 'to stop' (7).

(7) **Marie a {continué/arrêté} de trouver le vaccin.*
'Marie {continued/stopped} to find the vaccine.'

All these tests are summarized in Table 4, which illustrates the way we have used them so as to assign aspectual classes to verbs.

	State	Activity	Accomplishment	Achievement
Progressive	–	+	+	–
What happened	–	+	+	+
for x time	+	+	–	–
in x time / take x time	–	–	+	+
keep on	+	+	+	–
stop	–	+	+	–

Table 4. The behavior of aspectual classes according to a battery of tests.

Finally, adjectives have been examined with respect to gradability and scalarity. The first distinction bo be made is between degree and non-degree adjectives (8a): only the former accept modification by *très* 'very' and similar adverbials (Paradis, 2001). After that, we distinguished, among degree adjectives, those encoding open scales from those encoding closed scales (Kennedy & McNally, 2005). Following standard views (Kennedy & McNally, 2005), we used diagnostics oriented towards upper bounds and others oriented to lower bounds. Thus, for example, closed upper bounds accept modification by *complètement* 'completely' (8b), while closed lower bounds accept

modification by *légèrement* 'sightly' (8c). However, according to EBH, we assume that bounded (or closed) adjectives are those having a closed upper bound. On the other hand, open-scale adjectives, unlike closed-scale ones, accept diagnostics on comparison (8d-e).

(8) a. *très petit* 'very small' ; **très mortel* 'very mortal'
b. *{complètement / légèrement} transparent* 'completely / slightly transparent'
c. **{complètement / légèrement} étranger* 'completely / slightly foreign'
d. *Marie est grande pour une enfant de neuf ans.* 'Marie is tall for a nine year old girl.'
e. *Par rapport à son ami, Marie est grande.* 'Compared to her friend, Marie is tall.'

3 Results

Our first result concerns the number of unbounded items in the three categories under examination. We noticed indeed a significant proportion of mass nouns (103/112) (9a), atelic verbs (62/72), among which 54 are stative) (9b), and unbounded adjectives (72/97 upper open) (9c).

(9) a. *élégance* 'elegance'; *connaissance* 'knowledge'; *méfiance* 'distrust', 'suspicion'
b. $consister_{Stative}$ 'to consist'; $dominer_{Stative}$ 'to dominate' vs. $croître_{Dynamic}$ 'to grow'; $errer_{Dynamic}$ 'to wander'
c. *arrogant* 'arrogant'; *important* 'important'; *répugnant* 'disgusting'

This provides a first confirmation of the accuracy and coverage of the Extended Boundedness Hypothesis. The following sections describe our results in more details.

3.1 Deverbal Nance and their corresponding verbs

The distribution of the aspectual properties of verbs with respect to the countability of Nance are shown in Table 5.

		Verbs					
Nouns [72]		Telicity			Aspectual class		
Mass	63	Atelic	62	98.4%	State	54	85.7%
					Activity	8	12.7%
		Telic	1	1.6%	Achievement	1	1.6%
Count	9	Atelic	0	0%			
		Telic	9	100%	Achievement	7	77.8%
					Accomplishment	2	22.2%

Table 5. Aspectual properties of verbs related to Nance.

The data in Table 5 clearly confirm that the overwhelming majority of mass Nance (98.4%) are related to atelic verbs (10a), which is in line with Balvet et al. (2012). More precisely, we notice that 85,7% of these atelic verbs are stative (10b), an observation also made by Fábregas & Marín (2017) for Spanish. Conversely, the verbs related to count nouns are systematically telic (11). These results clearly confirm the Boundedness Hypothesis as far as nouns and verbs are concerned.

(10) a. *assistance* 'assistance' > *assister* 'to assist'; *ignorance* 'ignorance' > *ignorer* 'to be unaware'; *maltraitance* 'abuse'> *maltraiter* 'to abuse'
b. *dominance* 'dominance' > *dominer* 'to dominate'; *gouvernance* 'governance' > *gouverner* 'to govern', 'to rule'; *nuisance* 'nuisance', 'disturbance' > *nuire* 'to harm', 'to affect'

(11) *délivrance* 'delivery', 'deliverance' > *délivrer* 'to issue', 'to set free'; *soutenance* 'defense (of a thesis)' > *soutenir* 'to defend (a thesis)'; *vengeance* 'revenge' > *(se) venger* 'to retaliate'

3.2 Deadjectival Nance and their corresponding adjectives

Table 6 presents the distribution of degree and scalar properties of adjectives with respect to the countability of Nance.

Nouns [97]		Adjectives		
Mass	91	Non-degree	9	9,89%
		Unbounded	72	**79.12%**
		Bounded (upper)	10	10.98%
Count	6	Non-degree	5	**83.3%**
		Unbounded	0	0%
		Bounded (upper)	1	16.7%

Table 6. Scale properties of adjectives paired with Nance.

According to the Extended Boundedness Hypothesis, we should find unbounded adjectives in relation with mass nouns, and bounded adjectives related to count nouns. Table 6 shows that this prediction is not completely borne out. On the one hand, 79.12% of mass nouns are related to unbounded adjectives (12), while there is no count nouns related to an unbounded adjective, in line with the Extended Boundedness Hypothesis. However, count nouns are mostly paired with non-degree adjectives (13a), where bounded adjectives were expected (13b). This result, although unexpected and in need of closer analysis, has to be weighed against the reduced number of count nouns paired with adjectives (6/97).

(12) *abondance* 'abundance' > *abondant* 'abundant'; *endurance* 'endurance' > *endurant* 'enduring'; *répugnance* 'disgust' > *répugnant* 'disgusting'

(13) a. *naissance* 'birth' > *naissant* 'nascent'; *renaissance* 'revival' > *renaissant* 'reviving'; *suppléance* 'replacement', 'substitution' > *suppléant* 'substitute'
b. *défaillance* 'failure' > *défaillant* 'defective'

3.3 Nance related with verbs and adjectives

Finally, Table 7 sums up the properties of Nance related with both verbs and adjectives, and confirms our previous results.

Nouns [56]		Verbs			Adjectives		
Mass	52	Atelic	51	**98%**	Non-degree	10	19.6%
					Unbounded	37	**72.5%**
					Bounded	4	7.8%
		Telic	1	2%	Unbounded	1	100%
Count	4	Atelic	0	0%			
		Telic	4	**100%**	Non-degree	3	**75%**
					Bounded	1	25%

Table 7. The properties of verbs and adjectives related with Nance.

As before, we observe that mass nouns are mostly paired with atelic verbs and unbounded adjectives (14a), while the few count nouns that we have found are paired with telic verbs and non-degree adjectives (14b). In the latter case, however, the reduced number of examples prevents us to draw a firm conclusion. Similarly, the unavailability of count nouns related to atelic verbs is an interesting result, corresponding to our expectations. However, no strong conclusion can be drawn from such a number of cases.

(14) a. *condescendance* 'condescendance' > *condescendre* 'to condescend' / *condescendant* 'condescending'; *médisance* 'slander' > *médisance* 'slander' / *médisant* 'slandering'

b. *naissance* 'birth' > *naître* 'to be born' / *naissant* 'nascent'; *suppléance* 'replacement', 'substitution' > *suppléer* 'to substitute'/ *suppléant* 'substitute'

4 A comparison with Nence

As a further step, we began to expand AdVeNance by also including nouns ending in *-ence* (Nence), that stand in the same relation with verbs and adjectives as Nance, as shown in example (15).

(15) a. $préférence_N$ / $préférer_V$ 'preference' / 'to prefer'
b. $éloquence_N$ / $éloquent_{Adj}$ 'eloquence' / 'eloquent'
c. $négligence_N$ / $négliger_V$ / $négligent_{Adj}$ 'negligence', 'carelessness' / 'to neglect' / 'negligent', 'careless'

From the 212 Nence extracted from Lexique3, we discarded nouns that are not paired with verbs or *-ent* adjectives (*conférence* 'conference') and prefixed nouns (*incohérence* 'inconsistency', from *cohérence* 'consistency'), similarly to what we did for Nance. The few Nence referring to concrete objects (cf. *semence* 'seed') were also eliminated. We finally obtained a list of 109 forms related to verbs (15a), adjectives ending in *-ent* (15b) and both (15c), as displayed in Table 8.

Nence [109]	Related to V	Non-related to V	Total
Related to Adj	23	80	103
Non-related to Adj	6	—	
Total	29		

Table 8. Distribution of Nence and their related categories

Table 8 shows that most Nence are related to adjectives (103/109), whereas verbs are less represented (29/109, among which 6 Nence paired with verbs only). By contrast, we found 97 Nance related to adjectives, 72 to verbs and adjectives, and 16 to verbs only, on a total of 112 nouns (see Table 2).

Regarding the relevant semantic features of each category, we have for now completed the annotation process for both nouns and verbs, but we do not yet have gathered all the data for adjectives.

In the case of nouns, we observed a prevalence of mass nouns (99/109), which was also the case for Nance. Similarly, verbs, although less represented for Nence than Nance, are mostly stative (25/29).

Table 9 provides a comparison of the distribution of nominal countability and verbal aspect for Nence and Nance.

	Features	Nence (109)	Nance (112)
Nouns	Mass	99 (90,82%)	102 (91,07%)
	Count	10 (9,17%)	9 (8,03%)
Verbs	**Total**	**29 (100%)**	**72 (100%)**
	State	25 (86,2%)	54 (75%)
	Activity	2 (6,9%)	8 (11,11%)
	Accomplishment	—	2 (1,78%)
	Achievement	2 (6,9%)	8 (11,11%)

Table 9. Comparison of the properties of Nance and Nence and their related verbs.

These data lead us to conclude that Nance and Nence, as well as their related categories, display similar characteristics. In both cases, we observe a predominance of mass nouns and stative verbs. Thus, unboundedness is a consistent pattern of at least two of the categories involved.

Although more research is needed to examine thoroughly the relations between the mass/count properties of Nence and the aspectual properties of their related verbs on the one hand, and the

gradable / scalar properties of their related adjectives on the other hand, we can conclude that the two sets of data we examined tend to confirm the Extended Boundedness Hypothesis.

5 Concluding remarks and further research

The AdVeNance resource contains 112 Nance associated with their corresponding verbs and/or adjectives. All its items are annotated for the features prototypical to their category: mass/count for nouns. lexical aspect for verbs, scalarity for adjectives. The results provided by the annotation evidence that mass nouns are mostly related with atelic verbs and/or unbounded adjectives, thus confirming the Extended Boundedness Hypothesis (Bach 1976, Mourelatos 1978, Paradis 2001). Although we do not have a full picture of Nence and their related verbs and adjectives yet, the first results we obtained seem to point in the same direction.

The next step of the AdVeNance project is to make the resource we built available to the community as a remotely interrogatable database.

As a further development, we consider comparing the properties of Nance and their related verbs and adjectives with their Italian and Spanish counterparts, for which similar cross-categorial relations are observed.

References

Emond Bach. 1976. An extension of classical transformational grammar. In: R. Saenz (ed.), *Problems of linguistic metatheory: Proceedings of the 1976 conference*, East Lansing, MI: Michigan State University, 183–224.

Antonio Balvet, Lucie Barque, Marie-Hélène Condette, Pauline Haas, Richard Hughe, and Aurélie Merlo. 2012. La ressource Nomage. Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus. *Traitement Automatique des Langues (TAL)* 52(3): 129–152.

Delphine Beauseroy. 2009. *Syntaxe et sémantique des noms abstraits statifs : des propriétés verbales et adjectivales aux propriétés nominales.* PhD. Thesis, Université Nancy 2 & UMR 7118-ATILF.

Laurel Brinton. 1998. Aspectuality and countability. *English language and linguistics* 2(1): 37–63.

Georgette Dal and Fiammetta Namer. 2010. Les noms en *-ance/-ence* du français : quel(s) patron(s) constructionnel(s) ? *Actes du CMLF 2010*, Paris: ILF, 893–907

David Dowty. 1979. *Word meaning and Montague grammar*. Dordrecht: Reidel.

Edwige Dugas, Pauline Haas and Rafael Marín. 2021. Héritage sémantique des noms aux verbes : Étude des verbes dénominaux en français. *Verbum* 43: 69–95.

Antonio Fábregas and Rafael Marín. 2017. Lexical categories and aspectual primitives: The case of Spanish *-ncia*. In Maria Bloch-Trojnar & Anna Malicka-Kleparska (eds), *Aspect and Valency in Nominals,* Berlin: De Gruyter, 157–179.

Nelly Flaux & Danièle Van de Velde. 2000. *Les noms en français : esquisse de classement.* Paris : Ophrys.

Bernard Fradin. 2011. Remarks on state denoting nominalizations. *Recherches Linguistiques de Vincennes* 40: 73–100.

Bernard Fradin. 2012. Les nominalisations et la lecture ‘moyen’. *Lexique* 20, 129–156.

Bernard Fradin. 2014. La variante et le double. In : S. David, F. Villoing (eds) *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*. Nanterre: Presses Universitaires de Paris Ouest, 111–148.

Jane Grimshaw. 1990. *Argument structure.* Cambridge MA: MIT Press.

Silvia Gumiel-Molina, Norberto Moreno-Quibén, and Isabel Pérez-Jiménez. 2020. On degree minimizers in Spanish. *Borealis* 9(1), 69–86.

Ray Jackendoff. 1991. Parts and boundaries. *Cognition* 41: 9-45.

Christopher Kennedy and Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2): 345–381.

Françoise Kerleroux. 2012. Il y a nominalisation et nominalisation. *Lexique* 20: 157–172.

Marie Laurence Knittel. 2016. Les noms en *-ance* : un panorama. Actes du CMLF 2016. Paris: ILF. DOI: 10.1051/SHS2shsconf/

Clara Melloni. 2007. *Polysemy in Word Formation : The Case of Deverbal Nominals*. Tesi di dottorato, Dipartimento di germanistica e slavistica, Università di Verona, Verona.

Alexander Mourelatos. 1978. Events, processes and states. *Linguistics and Philosophy* 2: 415–434.

Boris New, Christopher Pallier, Ludovic Ferrand, and Rafael Matos. 2001. Une base de données lexicales du français contemporain sur internet : LEXIQUE. *L'Année Psychologique* 101, 447–462. http://www.lexique.org

Carita Paradis. 2001. Adjectives and boundedness. *Cognitive Linguistics* 12(1): 47–65.

Franz Rainer. 1989. *I nomi di qualità nell'italiano contemporaneo*. Wien: Braunmüller.

Danièle Van de Velde. 1995. *Le spectre nominal : des noms de matières aux noms d'abstractions.* Louvain / Paris: Peeters.

Designing a derivational resource for non-concatenative Morphology: the Hebrewnette database

Lior Laks
Bar-Ilan University
Ramat-Gan, Israel
Lior.Laks@biu.ac.il

Fiammetta Namer
UMR 7118 ATILF & Université de Lorraine
Nancy, France
fiammetta.namer@univ-lorraine.fr

Abstract

This paper presents a Derivational Database of Modern Hebrew (and more generally of Semitic languages) called Hebrewnette. The methodology adopted is based on adjusting the structure and properties of a database developed for the description of the derived lexicon of a Romance language (Démonette), and completing it to account for the specificities of the morphology of Semitic languages. We present the properties of Hebrewnette and the type of information in consists of, with special emphasis on both structural and semantic relations between words. Through a case study, we show how the annotations that are used allow us to verify theoretical hypotheses about non-concatenative morphology. The design of Démonette's annotation system makes its features, initially designed for French, suitable for capturing both morphological and semantic relations between Hebrew words, regardless of the type of morphology (concatenative or non-concatenative).

1 Introduction

This study presents the methodology of a derivational database of Hebrew (and more generally of Semitic languages) called Hebrewnette. The methodology adopted consists in adapting the structure of Démonette (Hathout and Namer, 2016; Namer and Hathout, 2020) a database developed for the description of the derived lexicon of French, and completing it to account for the specificities of the morphology of Semitic languages. Through a case study, we show how the annotations used allow us to verify theoretical hypotheses about non-concatenative morphology. The design of Hebrewnette relies on a word-based approach to morphology, whereas the tradition in the creation of tools (Daya et al., 2008) and lexical resources for Semitic languages (Neme, 2011)[1] is rather root-based (for an overview of theoretical approaches to Semitic morphology, see (Bat-El, 2017; Goldenberg, 1994; Ussishkin, 2005; Aronoff, 2007; Ravid, 2008; Berman, 2012; Faust, 2015; Kastner, 2020)). Hebrewnette provides a description of the derivational relations between Hebrew words in contrast to other types of database that relate mainly to inflectional paradigms. Finally, some works, even recent ones, point out the scarcity of freely available resources in Semitic languages, eg. in Arabic (El Haj et al., 2015). Hebrewnette (which is currently in a prototype stage) will contribute to fill this gap. Démonette, on which Hebrewnette is based, has been designed and implemented to represented the derivational relations within the French lexicon. Its realization is based on the following principles: (*i*) each entry is the relationship between two members of a derivational family; (*ii*) the same word participates in more than one entry; (*iii*) beside the classical base-to-derivative relations, entries in the database may correspond to cross-formations, or express a broader ancestor-descendant relation; (*iv*) both the words and their relation are identified by a set of morphological, phonological and semantic features.

2 Hebrew Morphology

Hebrew word formation relies highly on non-concatenative morphology, i.e. the formation via root and pattern (Berman, 1978; Bolozky, 1978; Schwarzwald, 1981; Ravid, 1990; Aronoff, 1994). The pattern

[1] see also: https://www.pealim.com/ for Modern Hebrew

indicates the prosodic structure of the word and it consists of the following elements: (i) consonantal slots; (ii) vocalic pattern; and in some cases (iii) affixes (Bat-El, 1994, 2017). For example, the verbs *diber* 'speak$_V$' and *tipes* 'climb$_V$' are formed in the CiCeC pattern. They share the vocalic pattern i-e and differentiate in their roots, `d.b.r` and `t.p.s` respectively. The verbs *hitraxec* 'wash oneself$_V$' and *hitragel* 'get used to$_V$' are formed in the hitCaCeC pattern, which consists of the prefix *hit-*, in addition to the vocalic pattern a-e. Words that share the same consonantal root typically share some semantic relations with different degrees of transparency, for example *hidpis* (hiCCiC) 'print$_V$', *hudpas* (huCCaC) 'be printed$_V$', *madpeset* (maCCeCet) 'printer$_N$' and *tadpis* (taCCiC) 'printout$_N$'. Hebrew verbal patterns typically differ from each other with respect to transitivity and the semantic types of verbs that they host (see (Berman, 1978; Bolozky, 1978; Borer, 1991; Aronoff, 1994; Doron, 2003; Schwarzwald, 2008) and references therein). For example, CiCeC typically hosts active transitive verbs, e.g. *kivec* 'shrink', *nigev* 'wipe' and *xibek* 'hug', while hitCaCeC typically hosts intransitive verbs like inchoatives (*hitkavec* 'become shrunk'), reflexives (*hitnagev* 'wipe oneself') and *hitxabek* 'hug each other'). However, these only represent tendencies and there is no one-to-one correspondance between form and meaning of the patterns. For example, *hitʔalel* 'abuse' is formed in hitCaCeC but does not belong to any of the above mentioned semantic classes.

Within verb formation, non-concatenative formation is obligatory and every verb that enters the language must conform to one of the existing patterns. In contrast, the formation of nouns and adjectives is based on a variety of word formation strategies. Nouns, for example, can be raw (*cav* 'turtle'), borrowed (*krason* 'croissant'), and can be formed in both patterns and by affixation. For example, agent nouns are formed in the CaCaC pattern (*cayar* 'painter', *nagar* 'carpenter') and by affixation (*yam* 'sea' - *yamay* 'sailor').

3 From Démonette to Hebrewnette: overview

The founding principles of Démonette (Hathout and Namer, 2016; Namer and Hathout, 2020) that have been applied to Hebrewnette are the following:

- Each entry describes a derivational relationship between two lexemes.
- The entries form derivational families represented by connected graphs.
- A derivational relation regards any pair of members of the same family: it can connect an ancestor to a descendant (e.g. a derivative: *dansable*$_A$ 'danceable' and its base: *danser*$_V$ 'dance') or two derivatives of the same base (e.g. *danseur*$_{Nm}$ 'male dancer' and *danseuse*$_{Nf}$ 'female dancer'), or two more distant elements of the family (e.g. *danser*$_V$ and *indansable*$_A$ 'undanceable'). Each relation is coded according to its orientation (does it connect a derivative to its base? Two words derived from the same base? etc.) and complexity (i.e. the number of derivational steps required to connect the two words).
- The base is deliberately highly redundant: each lexical unit has as many derivational descriptions as it has connections within its family.
- In addition to the properties of its relation with other words, each lexical unit is defined by features independent of the relations in which it is found (e.g. its inflectional paradigm, part of speech, ontological category, frequency...)
- The (lexical and relational) properties are grouped into patterns that generalize the different levels of regularities that can be found in the constructed lexicon: phonological, semantic, morphological.

Like Démonette, Hebrewnette is represented in a tabulated format. Each entry is a pair of (noninflected) words (W_1, W_2) belonging to the same derivational family. The morphological properties are divided between descriptions of the relations and descriptions of the words involved in these relations. The excerpt in Tab.1 gives an overview of the general organization of a Hebrewnette entry according to its

different properties. They are detailed in the following sections, in particular the features necessary for the expression of the non-concatenative morphology within Semitic languages.

As shown on the left part of Tab.1, each word is identified by its graphic form and phonetic transcription, its part of speech, and its English gloss. Formally, it is described by the pattern it belongs to, its root (and the type of root) and its vocalic structure, that is, its morphological structure (see §.4.2). When relevant, a feature encodes the variation between the morphological structure of a word and that of its pattern: for instance, the fact that the vowel /e/ in the noun *lemida* 'learning' is not predicted by its pattern CCiCa (see details in §.4.1 and Tab.2.4, column P_i to W_i).

Finally, each word is annotated by means of its ontological properties (Semantic Type, Semantic Subtype) and its argument structure (features Transitivity and Argument Structure). In Tab.1, the value 'dyn' of Semantic Type for *lamad* and *limed* indicates that both verbs are dynamic predicates. *lamad* is a regular active transitive predicate (Semantic Subtype= act, Transitivity=trans.), which is reflected by the value XY of its Argument Structure (someone$_X$ studies something$_Y$). The Semantic Subtype and Transitivity features of *limed* are valued causative and transitive, respectively, because *limed* introduces a causative argument W in its argument structure, with respect to *lamad* argument structure (someone$_W$ teaches someone else$_X$ something$_Y$).

The relation between two words is described according three dimensions, for reasons explained in §.4.3 (see right part of Tab.1):

- The orientation and complexity of the relation (is W_1 derived from W_2, W_2 from W_1 ? none of them is derived from the other? How many derivational steps are there between W_1 and W_2?) are examined separately from the structural and semantic points of view, see also Tab.4;
- The phonological dimension of the relation concerns the possible variation between the two words, and/or between their roots, see also Tab.3;
- The semantic relation is paraphrased by a gloss that cross-defines W_1 and W_2. Here, the cross-definition of *lamad* and *limed* illustrates the causative relation between the two verbs and between their arguments.

	$Word_1$	$Word_2$
Written form	למד	לימד
Phon transc	*lamad*	*limed*
Transl	study	teach
PoS	v	v
Pattern	CaCaC	CiCeC
Root	l.m.d	l.m.d
Root type	regular	regular
Morphological representation	|aa|	|ie|
Pattern-to-Word phon. altern.	NA	NA
Semantic type	dyn	dyn
Semantic subtype	act	caus
Transitivity	trans.	trans.
Argument structure	XY	WXY

Relation between $Word_1$ and $Word_2$	
Formal orientation	NA
Formal complexity	simple
W_1/W_2 Phon alternation	NA
Relation bwn roots	=
Semantic orientation	$W_1 \rightarrow W_2$
W_1/W_2 cross-definition	"when W limed X Y, then X lamad Y"

Table 1: The Hebrewnette database: an excerpt

4 The Hebrewnette database

In the following, we provide some examples of information needed to accurately represent the properties of words constructed by non-concatenative morphology. These features serve various purposes: represent each derivational relation and each word involved in it in terms of roots and patterns (§.4.1) as well as the (relation between) root classes (§.4.2), and describe meaning-form asymetry between the formal and the semantic orientations of the derivational relation (§.4.3).

4.1 Roots, patterns, affixes and structural variations

As we have just seen, the representation of non-concatenative derivations involves different annotations illustrated in Table 2. Some features relate to the words involved in the relation: they are distinguished according to whether or not they have a pattern (columns **Pattern P_1** and **Pattern P_2**). When the word has no pattern, it may be 'borr.(owed)' (e.g. *spam* in T2.1) or 'raw' (e.g. *yam* and *yami* 'of sea', in T2.3). When relevant, the representation of the pattern is completed by the description of the root (e.g. `l.m.d` for column $\mathbf{R}_2$ in T2.4), and that of each word structure (at columns $\mathbf{W}_1$ **Struct.** and $\mathbf{W}_2$ **Struct.**). The structure of a word consists in a vocalic pattern (e.g. *|oe|* for *lomed* in T2.4), possibly completed by affixes belonging to the pattern (e.g. *ti* and and *et* in *ti|0o|et*, in T2.2) and autonomous affixes (e.g. the suffix *-i* in *|iu|+i*, in T2.6). When relevant, the indication of a phonological shift between the representation of the word and that of its pattern is also provided. For instance, the annotation: `0`/$\mathtt{e}^{V1}_{W2}$ in T2.4, column $\mathbf{P}_i$ **to** $\mathbf{W}_i$ indicates the insertion of the vowel /e/ in position V_1 of W_2, that is, between the first and the second consonants of the word root. On the CCiCa P_2 pattern, V_1 is empty (the absence of the vowel is represented by the value '`0`') whereas it is filled with /e/ in the W_2 *l̲emida*. The CCiCa pattern typically has an initial consonantal cluster (CC) and vowel insertion does not occur, as illustrated with šmira, in T2.5. Other features are used to describe the structure of the relation itself (column **Structure of relation**), and the phonological variation between W_1 and W_2. For example, in T2.2, column $\mathbf{W}_1$**/**$\mathbf{W}_2$ **phono. alt.**, there is a /v/ to /b/ variation on the consonant position C_2, between *gav̲ar* and *tig̲boret*.

	W_1	W_2	Pattern P_1	Pattern P_2	R_1	W_1 struct.	R_2	W_1 struct.	P_i to W_i	Structure of the relation	W_1/W_2 phono. alt.
T2.1	*spam* 'spam$_N$'	*hispim* 'spam$_V$'	borr.	hiCCiC		|0a|	s.p.m	hi|0i|		CCaC/P2	
T2.2	*gavar* 'increase$_V$'	*tigboret* 'reinforcement$_N$'	CaCaC	tiCCoCet	g.b.r	|aa|	g.b.r	ti|0o|et		P1/P2	v/b^{C2}
T2.3	*yam* 'sea$_N$'	*yami* 'marine$_A$'	raw	raw		W		W+i		W/W+i	
T2.4	*lomed* 'learner$_N$'	*lemida* 'learning$_N$'	CoCeC	CCiCa	l.m.d	|oe|	l.m.d	|0i|a	0/$\mathtt{e}^{V1}_{W2}$	P1/P2	
T2.5	*šomer* 'guard$_N$'	*šmira* 'guarding$_N$'	CoCeC	CCiCa	š.m.r	|oe|	š.m.r	|0i|a		P1/P2	
T2.6	*limud* 'teaching$_N$'	*limudi* 'educational$_A$'	CiCuC	CiCuC+i	l.m.d	|iu|	l.m.d	|iu|+i		P1/P2	

Table 2: Formal representation of (relations between) words and patterns

4.2 Root types and inter-family relations

Words sharing the same root typically belong to the same morphological family (on the other hand, some families may consist of words without roots, as in T2.3). Morphological families form paradigms. The root description (Table 3) is information specific to each word. Roots are classified according to different types. They are regular ('r') if they contain three consonants (for example, `d.r.x`, in T3.2), 'r-4' if they are quadriliteral. In this case the dot '.' is used to group clusters (as `t.dr.x` in T3.4). When a pattern surfaces as a wordform with only 2 consonants (for instance *rac* in T3.5) the historical value of the missing root consonant is noted in capitals (e.g. /w/ of `r.W.c` in T3.5). Other values, not illustrated here, complete this tagset: for example, they indicate when the same root (e.g. `s.p.r`) corresponds to disjoint families with homonyms (*siper* 'tell$_V$' vs. *siper* 'cut hair$_V$') or polysemes (*xafar* 'dig$_V$' vs. *xafar* 'talk too much$_V$, drill one's mind (metaphorically)$_V$').

By default, a relation connects two items that share the same root (provided they belong to a pattern, compare T2.3 to T2.4). However, there are relations that connect items with different roots. These particular relations are characterized by adding a consonant to the root on word W_2, as in T3.3, where

	W_1	W_2	Patt. P_1	Patt. P_2	R_1	R_1 type	R_2	R_2 type	R_1 to R_2
T3.1a	*mahal* ‘mix_V’	*tamhil* ‘mix_N’	CaCaC	taCCiC	m.h.l	r	m.h.l	r	
T3.1b	*hidpis* ‘print_V’	*tadpis* ‘printout_N’	hiCCiC	taCCiC	d.p.s	r	d.p.s	r	
T3.2a	*hidrix* ‘guide_V’	*tadrix* ‘briefing_N’	hiCCiC	taCCiC	d.r.x	r	d.r.x	r	
T3.2b	*hidrix* ‘guide_V’	*hadraxa* ‘guidance_N’	hiCCiC	haCCaCa	d.r.x	r	d.r.x	r	
T3.2c	*tadrix* ‘briefing_N’	*hadraxa* ‘guidance_N’	taCCiC	haCCaCa	d.r.x	r	d.r.x	r	
T3.3	*tadrix* ‘briefing_N’	*tidrex* ‘debrief_V’	taCCiC	CiCeC	d.r.x	r	t.dr.x	r-4	CCC/tCCC
T3.4a	*tidrex* ‘debrief_V’	*tidrux* ‘debriefing_N’	CiCeC	CiCuC	t.dr.x	r-4	t.dr.x	r-4	
T3.4b	*tidrex* ‘debrief_V’	*tudrax* ‘be debriefed_V’	CiCeC	CuCaC	t.dr.x	r-4	t.dr.x	r-4	
T3.5	*rac* ‘run_V’	*rica* ‘running_N’	CaCaC	CCiCa	r.W.c	$r_{C2=W}$	r.W.c	$r_{C2=W}$	

Table 3: Root classification

`d.r.x` → `t.dr.x`. In that case, the variation between roots R_1 and R_2 is specified, for example, with CCC/tCCC. This type of relationship creates a new family, and its members share the new root. The two families form different paradigms. We can illustrate this observation with *tadrix* ‘briefing_N’/ *tidrex*‘debrief_V’ (T3.3):

- The taCCiC pattern, which includes the prefix *ta-*, is used for the formation of different kinds of nouns that can be related to verbs in different patterns, e.g. *mahal* ‘mix_V (liquids)’ - *tamhil* ‘mix_N’ (T3.1a), *hidpis* ‘print_V’ - *tadpis* ‘printout_N’ (T3.1b). The noun *tadrix* ‘briefing’ is formed in the taCCiC pattern, and is semantically related to the hiCCiC transitive verb *hidrix* ‘guide_V’ (T3.2a) and the haCCaCa action noun *hadraxa* ‘guidance_N’ (T3.2c). The three words are interconnected (T3.2a, 2b, 2c) and form a derivational family sharing the consonantal root `d.r.x`.

- As T3.3 shows, the verb *tidrex* ‘debrief_V’ is formed in the CiCeC pattern based on the noun *tadrix*, taking the `t` consonant of the derivational prefix *ta-* as part of the new root `t.dr.x`. The CiCeC pattern is paradigmatically connected to the CiCuC pattern of action noun (*tidrux* ‘debriefing_N’ in T3.4a) and to the verbal passive CuCaC pattern (*tudrax* ‘be debriefed_V’ in T3.4b).

We can see that the pattern CiCeC of W_2 (*tidrex*) induces new types of relations within its new family. These relations are paradigmatically determined. We can therefore say that a relation like *tadrix/tidrex* serves to connect two paradigms.

4.3 Meaning-form discrepancies: relations with diverging orientations

In Démonette, the value of the orientation feature indicates which of the two related words is the base (or the ancestor) of the other. Non-concatenative morphology is such that the formal orientation is often impossible to determine. For instance, in the *cilem/cilum* relation there is no formal clue to decide if *cilem* ‘photograph_V’ is the base of *cilum* ‘photography_N’ or is derived from it. By distinguishing semantic orientation and formal orientation these two aspects are dissociated. Therefore each derivational relation in a family can be properly described according to the value combination of these two independent features. Table 4 shows several cases of such combinations; orientations (columns 4 and 5) are symbolized by arrows, f_1 and f_2 stand for the form of W_1 and W_2 respectively, s_1 and s_2 represent their semantic content.

- base word → derived word *regular* orientation (T4.1): *maclema* ‘camera_N’ is more complex both formally and semantically than *cilem* ‘photograph_V’ (we assume that W_2 is semantically more complex than W_1 if the semantic content of W_2 includes at least one additional predicate or operator compared to W_1: here, W_2 denotes the instrument used to perform the action described by W_1).

- base word → derived word *semantic* orientation (T4.2): *šavir* 'breakable$_A$' is more complex than *šavar* 'break$_V$', whereas the formal orientation cannot be determined.
- base word → derived word *formal* orientation (T4.3): the structure of *hitkavec* 'get shrunk$_V$' is more complex than that of *kivec* 'shrink$_V$'. On the other hand, no semantic orientation can be assigned to the relation: it is unclear whether the intransitive predicate is built from the transitive one, or vice-versa (Haspelmath, 1993).
- indirect semantic relation (T4.4): the formal orientation between *kuvac* 'be shrunk$_V$' and *kavic* 'shrinkable$_A$' is indeterminate, and the semantic content of the two words are defined based on a common morphosemantic base (*kivec* 'shrink$_V$')
- two more combinations are illustrated in T4.5. The semantic contents of the agent noun *calam* 'photographer$_N$' and the instrument noun *maclema* are not directly related to one another, but they are semantically linked to the common verb ancestor *cilem*, and *maclema* can be formally derived from *calam*.
- double indeterminacy: in T4.6 the noun *šuman* and the adjective *šamen* are of the same formal complexity and share the same semantic content ('fat').

	W_1	W_2	Struct. of the rel.	Form. orient.	Sem. orient.
T4.1	*cilem* 'photograph$_V$'	*maclema* 'camera$_N$'	CiCeC/maCCeCa	$f_1 \rightarrow f_2$	$s_1 \rightarrow s_2$
T4.2	*šavar* 'break$_V$'	*šavir* 'breakable$_A$'	CaCaC/CaCiC	–	$s_1 \rightarrow s_2$
T4.3	*kivec* 'shrink$_V$'	*hitkavec* 'become shrunk$_V$'	CiCeC/hitCaCeC	$f_1 \rightarrow f_2$	–
T4.4	*kuvac* 'be shrunk$_V$'	*kavic* 'shrinkable$_A$'	CuCaC/CaCiC	–	$s_1 \leftrightarrow s_2$
T4.5	*calam* 'photographer$_N$'	*maclema* 'camera$_N$'	CaCaC/maCCeCa	$f_1 \rightarrow f_2$	$s_1 \leftrightarrow s_2$
T4.6	*šamen* 'fat$_A$'	*šuman* 'fat$_N$'	CaCeC/CuCaC	–	–

Table 4: Structural vs. semantic orientation of a relation

5 Case study: maCCuC formation

Some Hebrew adjectives have doublets that are formed in the maCCuC pattern, mostly in a jocular manner. The adjectives *maxrid* (1a)[2] and *maxrud* (1b)[3] , both denote 'awful', share the consonants `x.r.d`, but are formed in different patterns. A similar case is presented in (2) for *misken*[4] and *maskun*[5] 'poor$_A$'.

(1) a. lavašti jins **maxrid**
'I wore an awful pair of jeans'
b. hi xorešet al oto jins **maxrud**
'she wears the same awful jeans'

(2) a. eyze **misken**, ma hu asa la
'what a poor (guy), what did he do to her?'
b. eyze **maskun**, kol paam ani yocet alexa
'what a poor (guy), I lash out at you every time'

Not all speakers accept maCCuC forms like the ones in (1b) and (2b) (Bolozky, 1999, 2000), yet web searches reveal that they are productive. In contrast to cases like (1)-(2), there are many adjectives that do not have maCCuC counterparts, e.g. *metunaf* – **matnuf* 'filthy'. Why is it so? maCCuC formation (and

[2]https://bike.co.il/?p=2239
[3]http://tmi.maariv.co.il/style/Article-609396
[4]https://www.tiktok.com/@einabl_253/video/6948081577649818881
[5]https://www.inn.co.il/Forum/Forum.aspx/t851240

lack thereof) can be predicted based on structural and semantic properties of the base adjective. From the semantic point of view, maCCuC adjectives must have negative meaning, and therefore adjective like *maksim* 'charming' and *meratek* 'fascinating' do not have such doublets (**maksum*, **martuk*). maCCuC adjectives can be derived from adjectives in different patterns that are not marked for specific semantic meaning, e.g. maCCiC, muCCaC. This derivation is not oriented formally because both patterns are equally complex as they both consist of a prefix. The derivation is semantically oriented from maCCiC or muCCaC to maCCuC because a negated property is semantically more complex than the corresponding unmarked one.

On the structural dimension, adjectives with maCCuC doublets must have medial consonant clusters. maCCuC formation is faithful to the base, as it involves vowel(s) changes and preserves the syllabic structure (T5-a). This brings about structural transparency between the forms. maCCuC formation based on adjectives without medial clusters involves more changes of the base, especially modification of the syllabic structure, and therefore it is highly rare (T5-b) or unattested (T5-c,d). Unattested forms are not included in Hebrewnette, we add them here just for the sake of demonstration.

This difference can be predicted from the string distance between the 'regular' form (W_1) and its doublet W_2. The greater the difference, the higher the probability that W_2 is either very rare, or unattested. Distances can be computed by means of a string metric. In Table 5, we use a measure parametrized such that string modification is weighed according to the distance from the original syllabic structure. Therefore, vowel substitution is twice "cheaper" as prefix insertion or deletion. Moreover, it weights four times less than vowel deletion or insertion, because the latter transformation involves consonant (de)clusterization, that is, either breaking consonant clusters that exist in the base, or creating consonant clusters that are not part of the base. A maCCuC adjective occurs when the distance with respect to the 'regular' negative adjective is smaller than 4 or equals to it. Since Hebrewnette encodes both semantic and structural information of each word and the relations between words, this allows to predict which adjectives are more likely to have maCCuC doublets.

	W_1	W_2	W_1 Str.	W_2 Str.	W_1/W_2 string operations	W_1/W_2 string distance
a. Frequent maCCuC formations						
T5.a	*maxrid* 'awful$_A$'	*maxrud*	ma|0i|	ma|0u|	V_{subs}: /i/ > /u/	1
	misken 'poor$_A$'	*maskun*	mi|0e|	ma|0u|	V_{subs}: /i/ > /a/; /e/ > /u/	2
b. Unfrequent (T5.b) or unattested (T5.c,d) maCCuC formations						
T5.b	*metoraf* 'crazy$_A$'	*matruf*	me|ua|	ma|0u|	V_{subs}: /e/ > /a/; /a/ > /u/ V_{del}: /u/ > 0	6
T5.c	*metunaf* 'filthy$_A$'	**matnuf*	me|ua|	ma|0u|	V_{subs}: /e/ > /a/; /a/ > /u/ V_{del}: /u/ > 0	6
T5.d	*satum* 'blockheaded$_A$'	**mastum*	|au|	ma|0u|	Prefix: *ma-*; V_{del}: /a/ > 0	6

Table 5: Likeliness of maCCuC adjective doublets formation

Interestingly, adjectives with negative meaning without maCCuC counterparts have semi-counterparts in Segolate (Bat-El, 2012; Shany-Klein and Ornan, 1992) patterns like CeCeC or CaCeC. We relate to them as semi-counterparts or "semi-doublets" because unlike maCCuC, which is used for the formation of adjectives, these Segolate patterns usually serve for the formation of nouns, e.g. *satum* 'thickheaded', *setem* 'a thickheaded person', *metunaf* 'filthy', *tanef* 'a filthy person'. These segolate forms have peculiar behavior as they are not inflected for gender and number, unlike Hebrew animate nouns. For example, *metunaf* modifies only masculine nouns and its feminine form is *metunef-et*, while *tanef* relates to both genders. Regardless of the special status of these Segolate forms, they tend to be in complementary distribution with maCCuC forms with respect to marking the negative meaning of existing adjectives.

Similarly to the case of maCCuC doublet formation, the formation of CeCeC or CaCeC forms does not modify the syllabic structure of the base. Both types of formation involve faithfulness to the base.

Examine again the attested adjective *metunaf*. It doesn't have a maCCuC counterpart (**matnuf*, T5-c) because such formation would involve vowel deletion (in addition to vowel substitution), which creates a consonant cluster and therefore infringes syllabic structural faithfulness with respect to *metunaf*: this results in a distance of 6 between the two forms. In contrast, the formation of *tanef* (T6-a) is less pricy, because its distance from *metunaf* is only 4: the prefix *ma-* is deleted and vowels are substituted, but the syllabic structure of the two stems is the same. There are some cases in which the Segolate pattern formation is even less pricy, e.g. *satum – setem* (T6-b) where the two stems share the same syllabic structure. In both cases in Table 6, there is no modification of the syllabic structure of the base and therefore the formation of Segolate forms is cheaper than maCCuC forms.

Existing adjectives with a medial consonant cluster do not have Segolate counterparts for the same reason, namely such formation would change the syllabic structure of the base by breaking a consonant cluster, in addition to other changes. The adjective *maxrid* 'awful', for example, does not have a Segolate semi-counterpart like **xered* (T6-c) because this relation would imply vowel insertion that breaks the `xr` cluster, deletion of the prefix *ma-* and vowel substitution, corresponding to a distance of 7 between them.

	W_1 W_2	W_1 Str.	W_2 Str.	W_1/W_2 string operations	W_1/W_2 string distance
T6.a	*metunaf* *tanef* 'filthy$_A$'	me|ua|	|ae|	V_{subs}: /e/ > /a/; /a/ > /u/ Prefix del.: *me-*	4
T6.b	*satum* *setem* 'thickheaded$_A$'	|au|	|ee|	V_{subs}: /a/ > /e/; /a/ > /u/	2
T6.c	*maxrid* **xered* 'awful$_A$'	ma|0i|	|ee|	Prefix del.: *ma-* V_{subs}: /i/ > /e/; V_{ins}: /0/ > /e/	7

Table 6: CeCeC and CaCeC doublets formation

6 The Hebrewnette prototype

Hebrewnette is a prototype of 250 entries. The description of each entry is the product of 37 features. The Hebrewnette core is made up of 160 entries, corresponding to 127 lexemes and 19 families. They have been encoded to test the robustness of the database. These entries combine one or several of the characteristics specific to Hebrew derivation that we have presented in this article: mismatched formal and semantic orientations, non-triconsonant roots, absence of pattern, phonological alternations, structural variations, etc. The 10 other derivational families included in the current version of Hebrewnette have been generated and annotated semi-automatically. Based on an initial list of 10 CiCeC verbs, we relied on the nature fundamentally paradigmatic of the Hebrew verbal lexicon to implement the following predictions:

- CiCeC verbs are likely to realize active, transitive, dynamic predicate , e.g. *xibek* 'hug$_V$', *kivec* 'shrink$_V$', *nihel* 'manage$_V$';
- they are related to a CiCuC action noun (*xibuk* 'hug$_N$', *nihul* 'management$_N$'), a resultative adjective in the meCuCaC participle pattern (*menohal*[6] 'managed$_A$'). CiCeC is also derivationnaly related to the meCaCeC participle pattern that can surface as an adjective (*mexabek* 'hugging$_A$'), a agent noun (*menahel* 'manager$_N$') or an instrument noun;
- when attested, their hitCaCeC related verb is intransitive, typically inchoative (*hitkavec* 'become shrunk$_V$'), reflexive (*hitraxec* 'wash oneself$_V$') or reciprocal (*hitxabek*, 'hug each other$_V$').

[6]The /u/ to /o/ variation between the pattern meCuCaC and the word *menohal* is due to the fact that the second consonant of the root /h/ is a glottal stop.

From these 10 CiCeC verbs, the program produced 70 new annotated lexemes (after manual verification, 20 of them are discarded): each CiCeC verb is the source of a family of 6 members on average. Insofar as each member in a family is linked to all the others, this amounts to supplement the 160 initial wordpairs with 90 new fully documented entries.

7 Conclusions

This paper presented the main principles of designing Hebrewnette, a derivational database for Hebrew, and its properties. We accounted for the adaptations that were made on the Démonette database, which was originally designed for Romance Morphology. Focus was on non-concatenative formation, which is highly typical of Hebrew and Semitic languages in general. We outlined the way words were coded with respect to their root and pattern. Taking a word-based approach for word formation, Hebrewnette is also based on coding relations between words, and specifically for Hebrew, relations between roots and patterns. It is based on separate description of semantic and structural relations so that each type of relation can be examined according to different criteria, e.g. direction of derivation (if any). We examined a case study of doublet formation of adjectives in the Hebrew maCCuC pattern, and showed that the way words and their relations are coded in Hebrewnette can account for the likelihood of such doublet formation. While such doublet formation is semantically motivated in order to mark adjectives as carrying negative meaning, the likelihood of doublet formation is based on structural relations between the existing adjective and its doublet and the degree of faithfulness between them, namely the types of changes that the doublet formation requires. We showed that the proposed design of Hebrewnette allows the representation of the role of faithfulness in word formation.

The features and feature values in the Hebrewnette database intertwine with the content of Démonette, to account for the particularities of languages with non-concatenative morphology. However these additions do not compromise the architecture of Démonette, the global structures of the two databases are superimposable, which allows us to envisage a total interoperability between the two systems (and more widely between the morphologies of Romance languages and Semitic languages). We have shown that the combinability of features allows us to empirically verify hypotheses, which confirm the validity of Word-based approaches in non-concatenative morphology. Nonetheless, just like the Démonette database from which it is inspired, Hebrewnette allows for a multi-theoretical consultation / analysis of derivational relations, in the sense that it gives access not only to word-and-pattern relations (in order to be suited to the family and paradigms principles of derivation), but also to roots and root-and-pattern relations (in accordance with the needs of the root-based approaches to Semitic morphology).

Acknowledgments

This research has been realized as part of the project Demonext[7], supported by the Agence Nationale de la Recherche, grant number:ANR-17-CE23-0005. It has been also been supported by the Chateaubriand Fellowship Program of the French Embassy in Israel.

References

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

Mark Aronoff. 2007. In the Beginning was the word. *Language* 83:803–830.

Outi Bat-El. 1994. Stem modification and cluster transfer in Modern Hebrew. *Natural Language and Linguistic Theory* 12:572–596.

Outi Bat-El. 2012. Prosodic alternations in Modern Hebrew segolates. In Malka Muchnik and Zvi Sadan, editors, *Studies on Modern Hebrew and Jewish Languages*, Carmel Press, Jerusalem, pages 116–129.

Outi Bat-El. 2017. Word-based items-and processes (WoBIP): Evidence from Hebrew morphology. In Claire Bowern, Laurence Horn, and Raffaella Zanuttini, editors, *On Looking into Words (and beyond)*, Language Science Press, Berlin, pages 115–135.

[7]https://www.demonext.xyz/

Ruth Berman. 1978. Modern Hebrew structure. Report, University Publishing Projects.

Ruth Berman. 2012. Revisiting roots in Hebrew: A multi-faceted view. In Malka Muchnik and Zvi Sadan, editors, *Studies on Modern Hebrew and Jewish Languages*, Carmel Press, Jerusalem, pages 132–154.

Shmuel Bolozky. 1978. Word formation strategies in MH verb system: denominative verbs. *Afroasiatic Linguistics* 5:1–26.

Shmuel Bolozky. 1999. *Measuring productivity in word formation: the case of Israeli Hebrew*. Brill, Leiden.

Shmuel Bolozky. 2000. Stress placement as a morphological and semantic marker in Israeli Hebrew. *Hebrew Studies* 41:53–82.

Hagit Borer. 1991. The causative-inchoative alternation: a case study in parallel morphology. *The Linguistic Review* 8:119–158.

Ezra Daya, Dan Roth, and Shuly Wintner. 2008. Identifying semitic roots: Machine learning with linguistic constraints. *Computational Linguistics* 34(3):429–448.

Edit Doron. 2003. Agency and voice: The semantics of the Semitic templates. *Natural Language Semantics* 11:1–67.

Mahmoud El Haj, Udo Kruschwitz, and Chris Fox. 2015. Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation* 49:549–580. https://doi.org/https://doi.org/10.1007/s10579-014-9274-3.

Noam Faust. 2015. A novel, combined approach to semitic word-formation. *Journal of Semitic Studies* LX(2):287–316.

Gideon Goldenberg. 1994. Principles of Semitic word-structure. In Gideon Goldenberg and Schlomo Raz, editors, *Semitic and Cushitic Studies*, Harrassowitz Verlag, Wiesbaden, pages 10–45.

Martin Haspelmath. 1993. *More on the typology of inchoative / causative verb alternations*, John Benjamins, Amsterdam/Philadelphia, pages 87–111.

Nabil Hathout and Fiammetta Namer. 2016. Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for french. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Itamar Kastner. 2020. *Voice at the interfaces: The syntax, semantics and morphology of the Hebrew verb*. Language Science Press, Berlin. https://doi.org/10.5281/zenodo.3865067.

Fiammetta Namer and Nabil Hathout. 2020. Paradis and démonette –from theory to resources for derivational paradigms. *The Prague Bulletin of Mathematical Linguistics* 114:5–33. https://doi.org/10.14712/00326585.001.

Alexis Amid Neme. 2011. A lexicon of Arabic verbs constructed on the basis of semitic taxonomy and using finite-state transducers. In Benoît Sagot, editor, *WoLeR 2011*. pages 79–86. halshs-01186723.

Dorit Ravid. 1990. Internal structure constraints on new-word formation devices in Modern Hebrew. *Folia Linguistica* 24:289–347.

Dorit Ravid. 2008. Parsimony and efficacy: The dual binyan system of hebrew. In *13th International Morphology Meeting (13th IMM)*.

Ora R. Schwarzwald. 1981. *Grammar and Reality in the Hebrew Verb*. Bar Ilan University Press, Ramat Gan.

Ora R. Schwarzwald. 2008. The special status of nif'al in hebrew. In Sharon Armon-Lotem, Gabi Danon, and Susan Rothstein, editors, *Current Issues in Generative Hebrew Linguistics*, John Benjamins, Amsterdam, pages 61–75.

Michal Shany-Klein and Uzzi Ornan. 1992. Analysis and generation of hebrew segolate nouns. In Uzzi Ornan, Gideon Arieli, and Edit Doron, editors, *Hebrew Computational Linguistics*, Ministry of Science and Technology, Jerusalem, pages 39–51.

Adam Ussishkin. 2005. A fixed prosodic theory of nonconcatenative templatic morphology. *Natural Language and Linguistic Theory* 23:169–218.

The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources

Matteo Pellegrini Eleonora Litta Marco Passarotti Francesco Mambrini Giovanni Moretti
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 - 20123 Milan, Italy
{matteo.pellegrini}{eleonoramaria.litta}{marco.passarotti}
{francesco.mambrini}{giovanni.moretti}@unicatt.it

Abstract

In this paper, we propose a model to include a derivational lexicon for Latin (Word Formation Latin) within the LiLa Knowledge Base of interlinked linguistic resources for Latin. After a brief introduction on the architecture of LiLa, we discuss the differences between the flat organization of derivational information in LiLa's Lemma Bank and the hierarchical structure of Word Formation Latin, showing that the latter contains potentially useful information that is not already available in the former. We describe the modelling of such information in LiLa, exemplifying how different word formation processes are treated. We conclude the paper by showing the complementarity of the two approaches, and outlining the advantages offered by their interconnection.

1 Background and Motivation

In recent years, the principles of the so-called Linked Data paradigm[1] are increasingly being applied to language data and metadata, aiming to improve interoperability between resources originally developed for different purposes, hence characterised by different formalisms and conceptual models. As a consequence, a Linguistic Linked Data Cloud is being developed, to which several resources are continuously being added (Cimiano et al., 2020). Within this framework, the aim of the *LiLa* project[2] 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 like 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). As a consequence, 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. More specifically, the backbone of the architecture of the LiLa KB is the Lemma Bank, a large collection of lemmas – i.e. citation forms – to which both the tokens of textual resources and the entries of lexical resources can be connected, as well as the output of NLP tools. The Lemma Bank initially included a limited amount of derivational information on lemmas from the Word Formation Latin (WFL) lexical resource (Litta and Passarotti, 2019). A choice was made not to include the entire information provided by WFL, that, however, might prove useful in certain circumstances.

In this contribution, we describe a model designed to include all the information contained in WFL in the LiLa KB. In Section 2, we detail the architecture of the KB on the one hand and of WFL on the other hand. In Section 3, we describe the model that we propose in order to include WFL within the architecture of LiLa, showing how different word-formation processes are treated. Also, this section describes how our work interacts with other models developed by the Linked Data community – namely, the LexInfo ontology of data categories (Cimiano et al., 2011), the OntoLex-Lemon vocabulary for describing lexical

[1] https://www.w3.org/DesignIssues/LinkedData.html.
[2] https://lila-erc.eu.

resources (McCrae et al., 2017; Buitelaar et al., 2011) and, more specifically, its Morphology Module (Klimek et al., 2019). We conclude in Section 4 by reviewing the dissimilarities between the modelling of the original derivational information in the LiLa Lemma Bank and the one of the WFL resource linked to the KB, showing how the application of Linked Data principles and techniques can benefit the communication between diverse linguistic resources.

2 LiLa and Word Formation Latin

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[3], 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 130,000 Latin lemmas taken from the database of the morphological analyser Lemlat (Passarotti et al., 2017). 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, on its part, is a derivational lexicon of Latin, characterised by a step-to-step morphotactic approach: 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. More specifically, 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. In turn, within derivation rules, affixation (more specifically, prefixation and suffixation) and conversion are distinguished, depending on the presence of an affix and its nature. 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.

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

Table 1: Examples of Word Formation Rules in WFL.

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 – and branching out to all derivatives by means of the successive application of individual WFR. For example, Figure 1 shows a portion of the family taking root from the ancestor lexeme FELIX '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 currently includes only a selection of the derivational information contained in WFL. Besides `Lemmas`, two other classes are involved, namely `Affixes` – in their turn divided into `Prefixes` and `Suffixes` – and `Bases`, merely defined as abstract connectors between lemmas that belong to the same family. 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`.[4] As a consequence, 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 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.

[3] https://lila-erc.eu/lodview/ontologies/lila/.
[4] These properties are all defined in LiLa's ontology.

Figure 1: Word Formation in WFL.

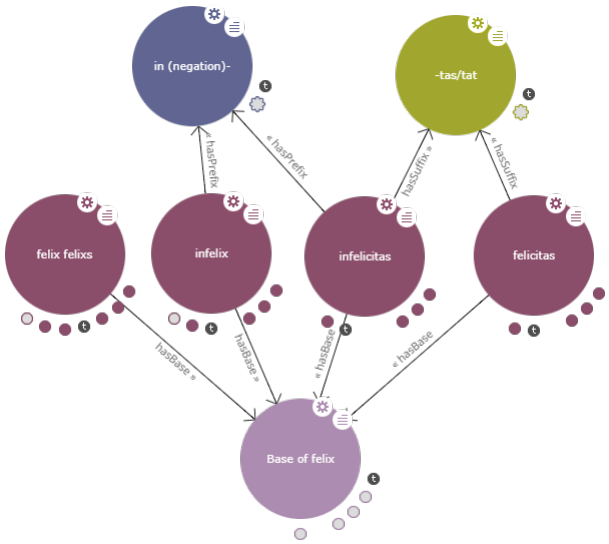


Figure 2: Word Formation in the Lemma Bank.

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.[5]

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, like Construction Morphology (Booij, 2010). Furthermore, such an approach allows for a more natural treatment of cases that were 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 included into the LiLa KB.

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

[5] Kyjánek (2020)'s 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.

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`),[6] 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 take the class `PartOfSpeech` from LexInfo (see again Section 1 for references), an ontology created to provide data categories for the OntoLex model, and we also 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.

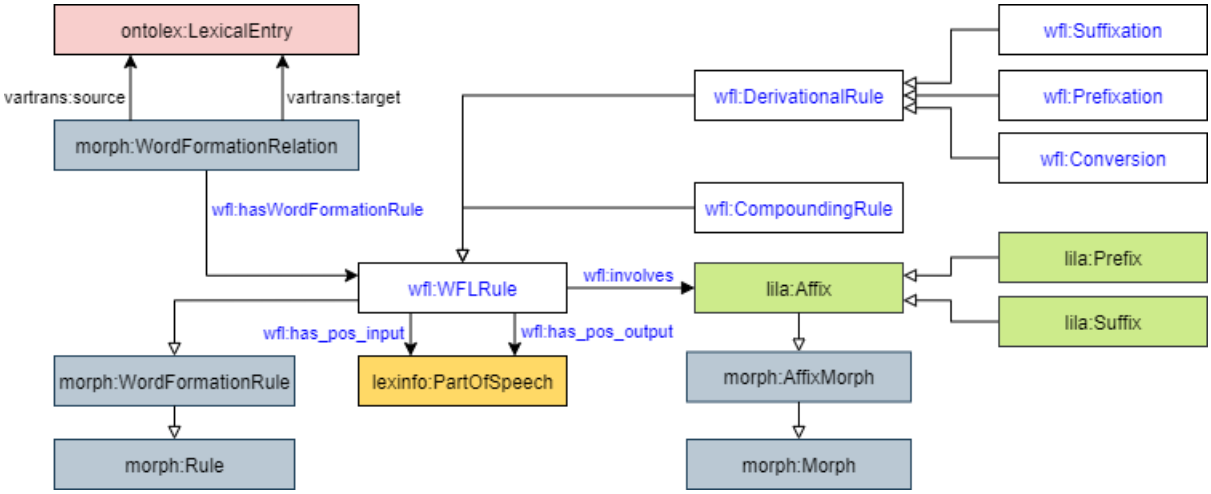


Figure 3: Architecture of the WFL ontology.

Let us now delve into some detail on the architecture of our model. We have 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`. The class `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.[7] For the same reason, rules are distinguished according to the lexical categories of the source and derivative, by providing a link to the `PartOfSpeech` of LexInfo through the properties `wfl:has_pos_input` and `wfl:has_pos_output`. Lastly, a 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`)[8] 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 Lila and the ones of `morph` (`morph:AffixMorph`) on the other hand.

To show the model at work with specific pairs of related words, Figure 4 shows the Linked Data treatment of the derivation of INFELIX 'unhappy' from FELIX 'happy' on the one hand (left side of the

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

[7] 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.

[8] 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.

image), of INFELICITAS 'unhappiness' from INFELIX 'unhappy' on the other hand (right side of the image).

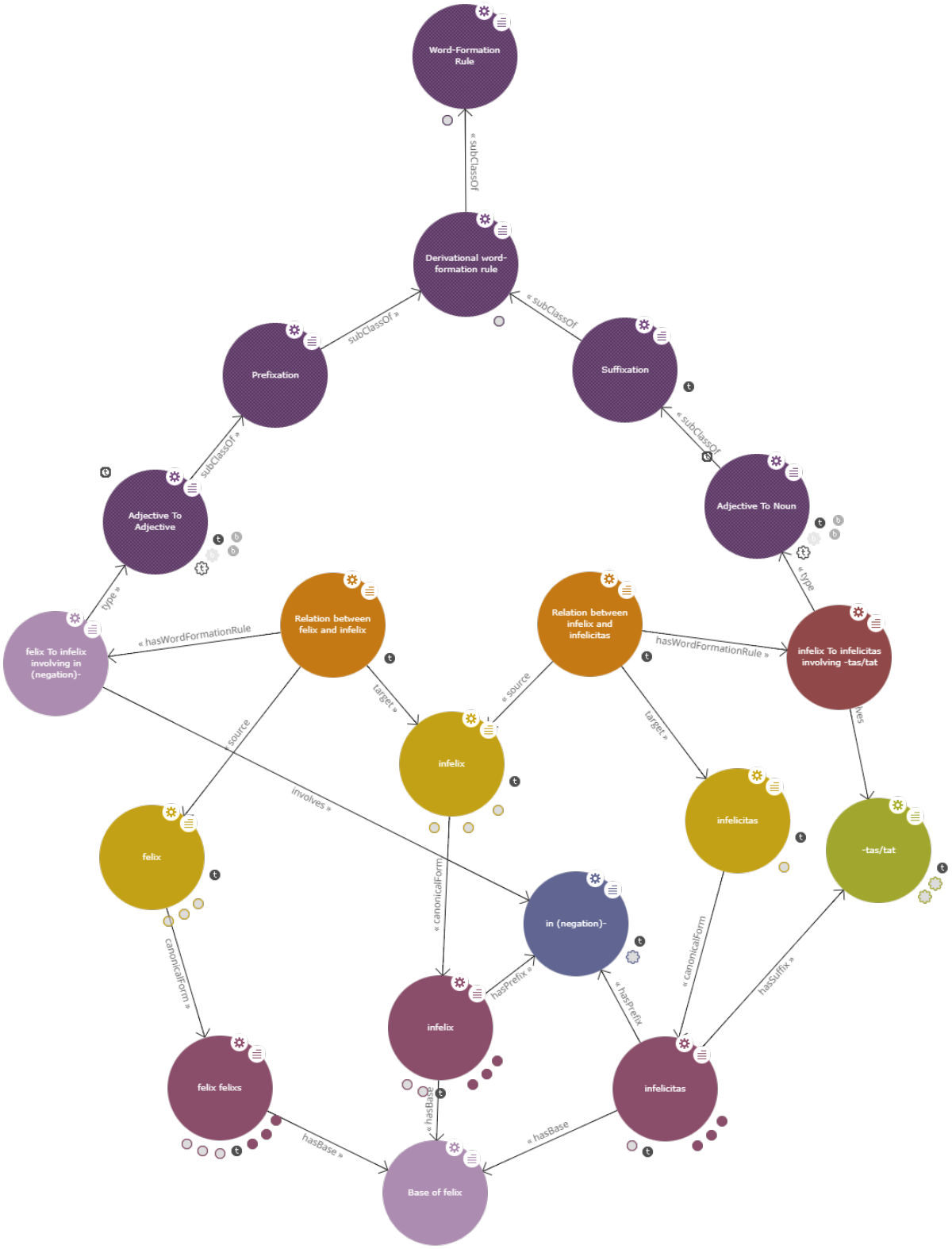


Figure 4: Modelling of prefixation and suffixation in the WFL ontology.

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 and INFELIX and one

between INFELIX and INFELICITAS. Each relation is instantiated by a specific WFR: see the nodes labelled as "felix To infelix involving in (negation)-"[9] and "infelix To infelicitas involving -tas/tat",[10] respectively. Starting from the one that forms INFELIX from FELIX, 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 from INFELIX, 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 'farmer' (from AGER 'field' + COLO '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 the value of this property is 1 for the relation between AGER and AGRICOLA, 2 for the relation between COLO and AGRICOLA.

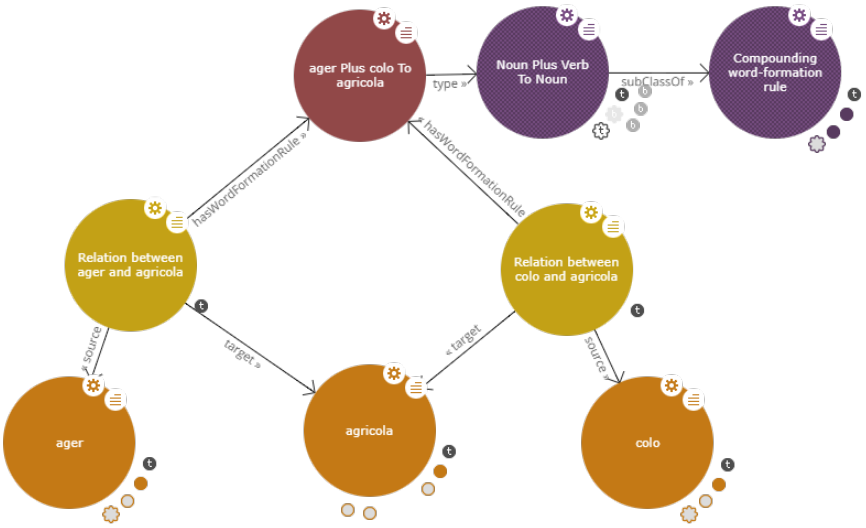


Figure 5: Modelling of compounding in the WFL ontology.

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.

[9]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'.

[10]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*.

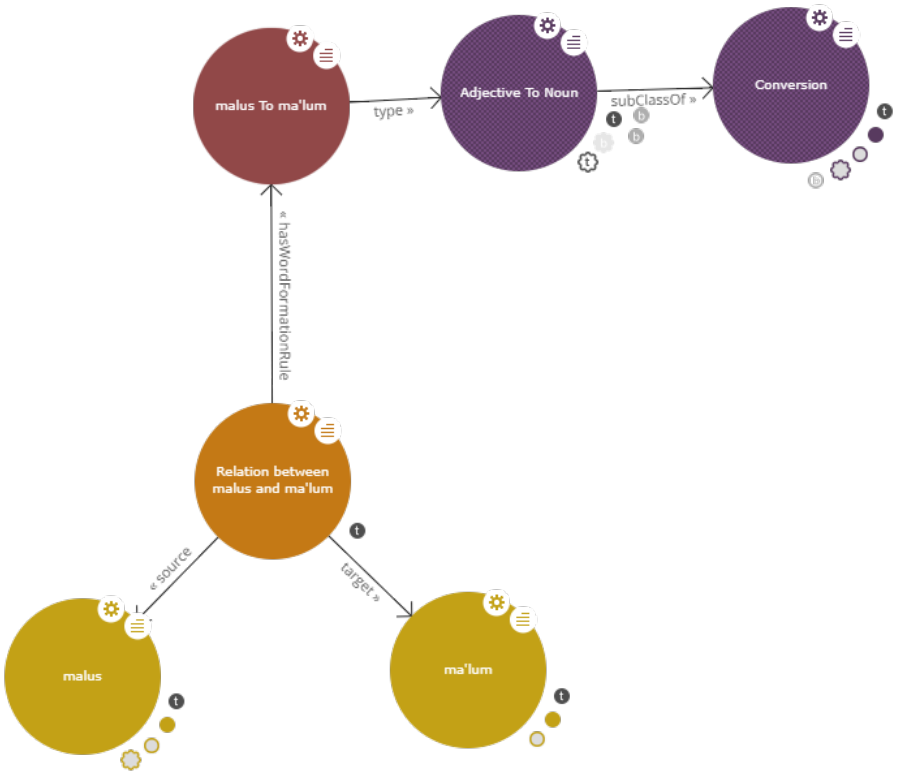


Figure 6: Modelling of conversion in the WFL ontology.

4 Discussion and Conclusion

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'[11]). 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 purposes 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

[11]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.

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.

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

References

Geert Booij. 2010. Construction morphology. *Language and linguistics compass* 4(7):543–555.

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

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

Christian Chiarcos and Maria Sukhareva. 2015. OLiA – Ontologies of Linguistic Annotation. *Semantic Web* 6(4):379–386.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Journal of Web Semantics* 9(1):29–51.

Philipp Cimiano, Christian Chiarcos, John McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data*. Springer.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *Proc. 12th International Semantic Web Conference, 21-25 October 2013*. Sydney, Australia.

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

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

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

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

Eleonora Litta and Marco Passarotti. 2019. (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Lemmata Linguistica Latina. Volume 1. Words and Sounds*, De Gruyter, Berlin, Boston, pages 224–239.

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

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

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

Scenarios and frames in derivation: a case study of derivational families based on animal names

Daniele Sanacore
CLLE, CNRS
Université de Toulouse
daniele.sanacore@univ-tlse2.fr

Nabil Hathout
CLLE, CNRS
Université de Toulouse
nabil.hathout@univ-tlse2.fr

Fiammetta Namer
ATILF, CNRS
Université de Lorraine
fiammetta.namer@univ-lorraine.fr

Abstract

This paper presents an ongoing work on a descriptive device for the characterization of semantic relations in the French derivational lexicon. We call this device the "morphosemantic frame" (MF). In order to describe morphosemantic regularities in the lexicon, we take inspiration from Fillmore's Frame Semantics. We use a case study of derivational families based on animal names to introduce "morphosemantic frames" and to illustrate how derivational families could be described by means of these script-like scenarios.

1 Introduction

Morphological relations are relations of form and meaning. While the formal properties of these relations have been the object of numerous studies, the organization of sense relations and of the structures from which meaning is calculated have been less explored. Our work deals with this latter point. We present a method for the representation of morphosemantic relations in derivational families and provide some elements for its automation and application to large amount of data. More precisely, in this work we present a case study where we apply this methodology on derivational families initiated by animal names in French.

2 Theoretical background

In the framework of paradigmatic derivational morphology (Štekauer, 2014, for a panorama), two notions are central: derivational family and paradigm. Families are sets of derivationally related lexemes (Hathout, 2011). An example is the family of *laver* 'to wash' provided in (1). Derivational families form paradigms, which are sets of families containing the same morphosemantic relations. An example of derivational paradigm is provided in (2), where the families of *laver*, *former* 'to train' and *gonfler* 'to inflate' present the same content relations (Bonami and Strnadová, 2019).

(1) *laver.v* 'to wash', *laveur.n* 'washer', *laveuse.n* 'female washer', *lavage.n* 'washing', *laverie* 'laundromat', *lavable* 'washable'.

(2) *laver.v* 'to wash', *laveur.n* 'washer', *lavage.n* 'washing';
former.v 'to train', *formateur.n* 'trainer', *formation.n* 'training';
gonfler.v 'to inflate', *gonfleur.n* 'inflater', *gonflement.n* 'inflating';

The approach that we adopt for the representation of morphosemantic relations in derivational families is inspired by the principles of Frame Semantics (Fillmore, 1976) and *FrameNet* (Ruppenhofer et al., 2006), a lexical resource that implements it. In Frame Semantics, frames are defined as structures that represent cognitive situations or objects along with their participants or features (called "frame elements"). A semantic frame is described by a sort of "story" that makes the semantic relations between the frame elements explicit. For instance, the COMMERCE-PAY frame is described by the gloss in (3) and is characterized by the frame elements (i.e. the participants) BUYER, SELLER, MONEY and GOODS. Moreover, frames are evoked by some lexical units (LUs), for example by the verb *to pay* or the noun *payment* in the case of , and realized by corpus sentences like the ones presented in (4).

(3) COMMERCE-PAY: This frame involves BUYERS paying MONEY for GOODS to a SELLER. In this frame the MONEY is the direct object, and is mapped to the theme of the transfer.

(4) a. I PAID her 50 dollars for a video game.
BUYER SELLER MONEY GOODS

b. Eurotunnel has offered PAYMENT in shares but TML doesn't want shares.
BUYER MONEY SELLER MONEY

3 Frames for morphosemantic description

This work is based on an approach to semantics in derivational morphology that adapts the semantic frames presented in Section 2 to morphosemantic description. Since we are interested in morphosemantic relations existing in derivational families, the paradigms that we want to represent are morphosemantic paradigms (i.e. sets of families structured by the same semantic relations). In this approach, frame-like structures contribute to the semantic characterization of the relations in derivational families. Derivational families are seen as implementations of such frames, in the same way as corpus sentences are seen as the concrete realizations of frames in *FrameNet* . Let us consider the example of the animal morphosemantic frame in (5). Such a frame contains as frame elements several features, concepts and participants that are generally associated with animals. As it can be seen, the frame is rather general; some aspects related to animals are missing; it involves concepts that are related to animals in different ways. This frame can however be used as a starting point to illustrate the method we propose.

(5) ANIMAL: An animal is a living being with certain PHYSICAL FEATURES (SIZE, COLOR, FUR, FLESH, OTHER PHYSICAL PECULIARITIES) and with a BEHAVIOR. The animal has a certain relationship with humans and it can be involved in human activities such as HUNTING, FISHING, BREEDING and SCIENTIFIC RESEARCH conducted by HUNTERS, FISHERMEN, BREEDERS and SCIENTISTS. If the animal is hunted, fished or bred, it is usually eaten by humans and used in RECIPES to prepare FOOD. The animal can be associated with some STEREOTYPES. If a the animal is a pest, it can be the target of a REMOVAL PROCEDURE realized by SPECIALISTS which use INSTRUMENTS.

A morphosemantic frame is a structure that describes a set of concepts related to a sort of semantic pivot, in this case, animal. If we consider derivational families initiated by animal names, such a conceptual structure can help characterize the meaning of derived nouns, relational adjectives and derived verbs. For example, verbs such as *zébrer* 'to stripe' and *léopardiser* 'to stain' are semantically related to the FUR of the animal, while a verb like *renarder* 'to fox' is semantically related to a STEREOTYPE associated with the fox (to be a cunning animal). Such a frame could also help the description of polysemy. For example, a verb like *saumoner* may mean 'to add salmon to something (for example in recipes)' or 'to give something the color of salmon'.

Others concepts contained in the morphosemantic frame in (5) are directly realized in the derivational families. For example, *renardier* 'fox hunter' and *louvetier* 'wolf hunter' concretely realize the HUNTER frame element, while *chevrier* 'goat breeder' or *apiculteur* 'beekeeper' realize the BREEDER frame element. Finally, lexemes like *dératisation* 'rodent control' or *démoustication* 'mosquito control' realize the REMOVAL PROCEDURE frame element.

4 Methodology for frame creation

We collected derivational families initiated by animal names using the derivational resource *Glawinette* (Hathout et al., 2020) and the *GLÀFF* lexicon (Hathout et al., 2014; Sajous and Hathout, 2015). We also collected lexicographic definitions from two electronic dictionaries, *Wiktionnaire* and *TLFi* (Pierrel et al., 2004).

Families built around animal names often seem to evoke distinct scenarios. For example, let us consider the derivational family built around *sardine* 'sardine' in (6). The lexeme *sardine* has two meanings. The first is the fish itself, while the second defines an object (a metallic pin) that has been named after the fish because of its shape (7). The derived noun *sardinier* is associated with four possible meanings, as shown by the lexicographic definitions in (8). It can denote a fisherman specialized in sardines, a ship used to

fish sardines, the owner of a factory that stocks and sells sardines or a worker of such a factory. These four senses, in a frame-based perspective, seem to realize two distinct scenarios. The first concerns the fishing activity and involves participants such as the fisherman, the tools used for fishing, the boat used for fishing, the fishing activity, the fish itself etc. The other scenario concerns an industrial activity, in this case the production and distribution of sardine cans in factory, the owner of that factory, the product, the grocery shops where this product will be sold, etc. These two scenarios are also evoked by the derived relational adjective *sardinier*, which can refer either to the fishing activity or the industrial distribution of sardines.

(6) *sardine.n, sardinade.n, sardinerie.n, sardinier.n, sardinier.a, sardinière.n, sardinière.a, sardinal.n, sardiner.v, ensardiner.v*[1]

(7) a. SARDINE.N: *Poisson de mer au corps fuselé d'une vingtaine de centimètres de long.* 'sea fish with a streamlined body and around twenty centimeters long'.
b. SARDINE.N: *Broche métallique servant à fixer une tente de camping au sol.* 'metal pin used to fix a camping tent to the ground'.

(8) a. SARDINIER: *Pêcheur de sardines* 'sardine fisherman'.
b. SARDINIER: *Ouvrier, ouvrière qui prépare les sardines* 'worker that prepares sardines'.
c. SARDINIER: *Industriel de la sardine* 'sardine industrialist'.
d. SARDINIER: *Bateau qui se consacre à la pêche à la sardine* 'boat used for fishing sardines'.

(9) SARDINIER.A: *Relatif à la pêche ou aux industries de la sardine* 'related to fishing or to fish industry'.

The derived noun *sardinerie* (10) denotes the factory where sardines are canned and thus realizes a concept inscribed in the scenario of an industrial activity, while the noun *sardinal* (11) is used to refer to the nets used for fishing sardines (along with anchovies and or other species of similar size) and is inscribed in the scenario of the fishing activity.

The derived noun *sardinade* belongs to a third scenario. It denotes a recipe that makes use of sardines and the meal prepared using this recipe (12). In this case, the scenario concerns the preparation of meals with recipes that make use of the flesh, the grease, or other parts of a given animal.

(10) SARDINERIE.N: *Usine où l'on prépare les sardines pour les conserver* 'factory where sardines are prepared in order to be conserved'.

(11) SARDINAL.A: *filets dont les mailles sont calibrées pour prendre des sardines, des anchois, etc.* 'nets whose knits are calibrated to catch sardines, anchovies, etc'.

(12) SARDINADE.N: *Recette de cuisine méditerranéenne où des sardines sont cuites entières.* 'Mediterranean cuisine recipe where the whole body of sardines is cooked'

The denominal verbs *se sardiner* (13a) and *ensardiner* (13b) have a similar meaning related to a stereotype associated with the animal, in this case, the fact of being crammed in cans.

(13) a. SARDINER.V: *(Pronominal) S'entasser comme des sardines dans une boîte de conserve* 'to cram like sardines in a sardine can'.
b. ENSARDINER.V: *Entasser comme des sardines* 'to cram something like sardines'.

The derivational family of *sardine* seems to realize several different scenarios and, on this basis, its lexemes could be semantically described by means of five morphosemantic frames described in (14, 15, 16, 17, 18). These subframes can be merged into a general frame like the one presented in (5).

(14) ANIMAL_FISHING: An ANIMAL is fished for its FLESH, for its GREASE, its SKIN or OTHER BODY PARTS. This ANIMAL is fished by some FISHERMEN, who may use some special BOATS or some special FISHING TOOLS in their activity.

[1]We excluded from the analysis terms that where marked as aged or not representative of contemporary French

(15) ANIMAL_INDUSTRIAL_PREPARATION: The FLESH, the GREASE, SKIN of an animal is generally prepared by some WORKERS in some INDUSTRIES in order to be COMMERCIALIZED in some GROCERIES.

(16) ANIMAL_RECIPES: The FLESH, the GREASE, the ORGANS or other parts of an animal are generally used in some RECIPES to prepare some FOOD.

(17) ANIMAL_OBJECT_ASSOCIATION: An OBJECT resembles an ANIMAL or a PART OF THE ANIMAL in its appearance.

(18) ANIMAL_BEHAVIOR: An ANIMAL is associated with a given stereotyped BEHAVIOR. A PERSON that adopts this BEHAVIOR can be given the name of the ANIMAL in a metaphoric sense.

These subframes describe a set of concepts that are concretely realized by the lexemes in the family of *sardine*. Their relations with the other members of the family could be glossed as in (19a, 19b, 19c, 19d, 19e), where these lexemes are substituted for the frame elements they realize.

(19) a. Une SARDINE est pêchée par un SARDINIER qui se trouve à bord d'un SARDINIER et utilise un SARDINAL.
'a sardine is fished by a sardine fisherman who is sailing on a sardine boat and using a sardine net'.

b. Un SARDINIER est un industriel qui possède une usine où des SARDINIERS entassent des SARDINES dans des boîtes de conserve.
'a sardine industrialist owns a factory where workers cram sardines in cans'.

c. Une personne prépare un plat à base de SARDINES en suivant une recette, la SARDINADE.
'a person prepares a dish based on sardines following a recipe that uses sardines'.

d. Une SARDINE est un objet qui rassemble à une SARDINE.
'a metal pin (sardine) is an object that resembles to a sardine'

e. Une SARDINE est associée à l'état d'être ENSARDINÉ.
'a sardine is associated with the state of being crammed in cans'

We extended this operation to other derivational families in an iterative way in order to (*i*) validate the subframes already created, and (*ii*) create other subframes based on the concepts realized in families.

To illustrate the point (*i*), we can find other derivational families that fit the scenarios that we proposed. These families are initiated by animals that are involved in fishing, industrial distribution or cuisine like *saumonier* 'fisherman specialized in salmons ', *morutier* 'fisherman specialized in cods', and *carpiste* 'fisherman specialized in carps' which realize the FISHERMAN frame element, while *harenguier* 'ship used for fishing herrings', *homardier* 'ship used for fishing lobsters', and *thonier* 'ship used for fishing tuna' realize the BOAT frame element. In other words, several derivational families realize the same ANIMAL_FISHING morphosemantic frame and as a consequence, can be aligned with respect to the frame.

To illustrate the point (*ii*), let us consider the family of *renard* 'fox' in (20). The analysis of this family involves a hunting scenario (25) which includes the fur of the hunted animal (*renard*), the animal itself (*renard)*, the hunter (*renardier*), and the lair where the animal hides (*renardière*). In addition, the family of *renard* presents a group of lexemes related to the stereotyped behavior associated with the animal (*renard*, *renardie*, *renardise*).

(20) *renard.n, renarder.v, renardie.n, renardise.n, renardier.n, renardière.n*

(21) a. RENARD: *Mammifère carnivore, au museau pointu et aux oreilles droites.*
' carnivorous mammal, with a pointed snout and straight ears'

b. RENARD: *fourrure de renard* ' fox fur'

c. RENARD: *Personnage cauteleux, fin et rusé* ' cunning, shrewd person'

(22) RENARDIER: *(Chasse) Celui qui est chargé de prendre les renards*
'person in charge of catching foxes'

(23) RENARDIÈRE: *Tanière du renard.* ' fox lair'.

(24) RENARDIE[2]: *Ruse, déloyauté, action de renard* ' disloyalty, cunning action'

(25) ANIMAL_HUNTING: An ANIMAL is hunted by a HUNTER because of its FLESH, FUR, GREASE or BODY PARTS. The hunter make use of WEAPONS, TRAPS and HUNTING ANIMALS.

5 Lexicographic information for an automatic feeding of morphosemantic frames

Another question we are intersted in is the (partial) automation of our method. What are the elements present in lexicographic definitions that could be helpful to assign a derived lexeme to the frame element FISHERMAN in the ANIMAL_FISHING frame? How can we find the derivational families that realize a given morphosemantic frame?

We can use some recurring structures in definitions and various keywords and labels. Consider the definitions of *carpiste* ' carp fisherman', *saumonier* 'salmon fisherman' and *morutier* 'cod fisherman' in 26) which realize the FISHERMAN frame element in the ANIMAL_FISHING subframe. Their definitions have a similar structure and contain regular key-phrases like "*pêcheur*+ [gerundive]" or "*marin-pêcheur*". Similar regularities are identifiable for the lexemes that realize the BOAT, the FOOD and the FUR frame elements, as shown in (27), (28) and (29).

(26) a. CARPISTE: *Pêcheur se consacrant uniquement à la pêche de la carpe* 'fisherman who dedicates himself to carp fishing'
b. SAUMONIER: *Personne pratiquant la pêche au saumon* 'person that fishes salmon'
c. MORUTIER: *Marin-pêcheur pratiquant la pêche à la morue* 'fisherman who fishes cod'

(27) a. HARENGUIER: *Bateau spécialisé dans la pêche du hareng* 'boat specialized in herring fishing'
b. THONIER: *Bateau destiné à la pêche au thon* 'boat used for tuna fishing'.

(28) a. ANCHOÏADE: *Préparation culinaire à base d'anchois pilés et de câpres.* 'culinary preparation made using piled anchoivies and and capers'.
b. HOMARDINE: *(Cuisine) Sauce à base de homard* 'sauce made with lobsters'

(29) a. VISON: *(Par métonymie) Fourrure de cet animal* '(metonymy) mink fur'
b. LOUTRE:*(Par métonymie) Fourrure de cet animal* '(metonymy) otter fur'

Some examples of markers that can denote the realizations of the cited frame elements are provided in Table 1.

FISHERMAN	BOAT	FUR	FOOD
pêcheur de	*bateau*	*fourrure de*	*préparation culinaire*
pêcheur+gerund.	*bateau*+ [part.]	*par métonymie* + fourrure	*cuisine* + à base de
pêche	*pêche*	*par ellypse*+ fourrure	*sauce*

Table 1: Frame element markers in dictionnary definitions

6 Conclusion

In this work, we used a case study of derivational families based on animal nouns in order to introduce and illustrate morphosemantic frames. We showed that such families seem to realize distinct scenarios where animals are involved and that several families can be aligned with respect to those scenarios on the basis of their morphosemantic regularities.

[2]The same definition is provided for *renardise*.

Acknowledgements

This work benefited from the support of the project DEMONEXT ANR-17-CE23-0005 of the French National Research Agency (ANR).

References

Olivier Bonami and Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2):167–197.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*. pages 20–32.

Nabil Hathout. 2011. Morphonette: a paradigm-based morphological network. *Lingue e linguaggio* 10(2):245–264.

Nabil Hathout, Franck Sajous, and Basilio Calderone. 2014. Glàff, a large versatile french lexicon. In *Conference on Language Resources and Evaluation (LREC)*. pages 1007–1012.

Nabil Hathout, Franck Sajous, Basilio Calderone, and Fiammetta Namer. 2020. Glawinette: a linguistically motivated derivational description of French acquired from GLAWI. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*.

Jean-Marie Pierrel, Jacques Dendien, and Pascale Bernard. 2004. Le tlfi ou trésor de la langue française informatisé. *Actes de EURALEX* 4.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Franck Sajous and Nabil Hathout. 2015. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*. Herstmonceux, England, pages 405–426.

Pavol Štekauer. 2014. Derivational paradigms. *The Oxford handbook of derivational morphology* 354:369.

Deriving the Graph: Using Affixal Senses for Building Semantic Graphs

Matea Filko
Faculty of Humanities
and Social Sciences
University of Zagreb
matea.filko@ffzg.hr

Krešimir Šojat
Faculty of Humanities
and Social Sciences
University of Zagreb
ksojat@ffzg.hr

Vanja Štefanec
Faculty of Humanities
and Social Sciences
University of Zagreb
vstefane@ffzg.hr

Abstract

In this paper, we will present the semantic graphs as a new way of exploring semantic relationships in CroDeriv. CroDeriv is a morphological database developed for the Croatian language. In previous phases of its development words were segmented into morphemes and derivational links among the base word and the derivative were marked. Currently, we focus on the analysis of affixal meanings. This analysis is the basis for the production of semantic graphs. Semantic graphs are used to capture semantic similarities within various derivational families.

1 Introduction

Language resources dealing with derivational morphology are nowadays being developed for numerous languages (Kyjánek, 2018; Filko, 2020). However, these resources differ significantly in terms of the way they are composed and the type of data they contain. Although the analysis of derivational processes is inextricably linked to semantics, semantic description is usually out of focus in the first stages in the building of such resources due to its complexity and frequent unpredictability. In the existing derivational resources, mainly regular semantic relations between derivationally connected words are marked. For example, Démonette (Hathout and Namer, 2014) provides semantic characteristics of words and indicates whether they denote an action, an agent or a property. The authors of Derivancze (Pala and Šmerk, 2015) mark the semantic type of derivational relations. However, only those relations that have regular and transparent meanings are identified, such as the relation between the action and the agent of the action or an adjective and the properties of the attributes marked by the adjective. Semantic relations have also been added to DeriNet, as the latest phase of its expansion. Ševčíková and Kyjánek (2019) describe a semi-automatic procedure for assigning semantic relationships to units in DeriNet. The semantic relations they have added are in line with the relations listed by Bagasheva (2017), which allows for a later comparison in various languages. At this initial stage, the authors focus on five semantic categories: diminutive and female (for nouns), possessive (for adjectives), and iterative and aspect (for verbs).

In this paper, we present the first steps in the marking of semantic categories of affixes in CroDeriv.[1] The development of CroDeriv in the first phases focused on a complete morphological and derivational analysis of verbs (Šojat et al., 2013), nouns (Šojat et al., 2014; Filko, 2020) and adjectives (Filko and Šojat, 2017). This means that the lexemes were segmented into morphs at the surface layer and all morphs were connected to the corresponding morphemes at the deep layer of presentation (for example, *učiteljica* 'female teacher': *uč-i-telj-ic-a* (surface layer) – *uk-i-telj-ic-a* (deep layer). These procedure is done manually for the approximately 14 000 verbs, 1 500 adjectives and 5 500 nouns, due to low precision rates of the automatic procedures (Šojat et al., 2014).

In the next phase, the starting word in the derivational process was marked, as well as the type of the derivational process that was used for the derivation of particular derivatives (*učitelj* 'male teacher' + *-ica* → *učiteljica* 'female teacher'). Besides, it was indicated whether this is a derivational process that changes the part-of-speech category of derivatives or not (Filko et al., 2020). This procedure is also done

[1]CroDeriv is available at croderiv.ffzg.hr.

manually for the subset of verbs (with more than 5 representatives in the derivational family), and for the nouns and the adjectives in the database.

Semantic categories are marked on derivational affixes (see Section 2). Labelling of these categories in graphs within derivational families provides an insight into semantic processes taking part in derivation. Same semantic processes can be realized by various means of derivation and within different derivational families. For example, the derivational semantic pattern **action** → **agent** → **female agent** is realized within the derivational family of the root *uk-* as:

učiti 'to teach' → *učitelj* 'male teacher' → *učiteljica* 'female teacher',

but also within the derivational family of the root *voz-* as:

voziti 'to drive' → *vozač* 'male driver' → *vozačica* 'female driver'

or within the derivational family of the root *da-* as:

izdavati 'to publish' → *izdavač* 'male publisher' → *izdavačica* 'female publisher'.

The aim of such a procedure is to establish semantic paths, i.e. regular semantic shifts / patterns in the derivation, in addition to the derivational paths. This kind of information is essential for the description of the Croatian morphotactics, which is still an under-investigated area of Croatian linguistics (Filko, 2020).

The paper is structured as follows: in the next section we will explain the principles of assigning semantic categories to affixes. In Section 3, we will describe how affixal meanings are encoded in CroDeriv. In Section 4, the semantic graphs will be introduced. We will present how the semantic graphs can reveal semantic properties of derivational families on the one hand, and particular affixes on the other. We will conclude with final remarks.

2 Affixal meanings

Two basic approaches to affixal meanings differ depending on whether they place emphasis on the process of homonimization or on the process of polysemization of affixes. The process of homonimization splits affixes into two or more separate units, while the polysemization reinterprets homonyms as a single unit (Raffaelli, 2015, 187). As a consequence, homonimization multiplies the number of units, while polysemy results in the meaning networks of particular units, in which it is possible to detect the links between different affixal meanings.[2] We believe that speakers recognize these links between meanings and that this enables them to use vocabulary economically and effectively. In addition, we believe that such an approach is methodologically more justified because it does not multiply the number of units. Thus, we consider suffixes as polysemous units, i.e. in the analysis we give preference to polysemy rather than homonymy. This approach is well described and substantiated in the reference literature (Rainer, 2014; Aronoff and Fudeman, 2011; Lieber, 2004; Lehrer, 2003; Babić, 2002). As indicated by Filko (2020), this approach is in line with the cognitive-semantic view that polysemy is linguistically and cognitively more economical than homonymy (Raffaelli, 2015).

We determine the meanings on the basis of the synchronic semantic analysis, combining the semasiological and onomasiological approaches at the same time (Bagasheva, 2017). The semasiological approach is manifested in the analysis of the polysemous structures of individual affixes, e.g., nominal suffix *-ba* can have five meanings in its meaning network:

1. action (*ploviti* 'to sale' → *plovid-ba* 'sailing')
2. result (*skladati* 'to compose' → *sklad-ba* 'composition')
3. event (*svat* 'wedding guest' → *svad-ba* 'wedding')
4. location (*nastaniti* 'to dwell' → *nastam-ba* 'dwelling')
5. non-transparent meaning (*opor* 'harsh' → *opor-ba* 'political opposition') (Filko, 2020, 172-173).

[2] In the example of the Croatian suffix *-ba* below, we can notice several cognitive metonimies leading to the diversification of its meaning network, e.g. ACTION FOR RESULT, ACTION FOR PLACE OF ACTION. Numerous types of regular polysemy patterns are recognized in Apresjan (1974).

This example also shows that suffixal meanings can be directly recognized and determined when dealing with semantically transparent motivated words (examples 1-4). On the other hand, semantically non-transparent derivatives and suffixes used in their derivation are treated as a separate group. This holds for derivatives for which it is not possible to clearly determine which part of their meaning is shaped on the basis of the meaning of the suffix (example 5).

The onomasiological approach, however, is used for the determination of the affixes used in the formation of the same semantic categories, e. g. the meaning 'property, quality' can be expressed by at least three different nominal suffixes[3]:

1. -ina (*bijel* 'white' → *bjelina* 'whiteness')
2. -oća (*slijep* 'blind' → *sljepoća* 'blindness')
3. -ost (*slab* 'weak' → *slabost* 'weakness') (Filko, 2020, 191).

It is important to emphasize that, when determining the meaning of affixes, we distinguish the lexical meaning and the derivational meaning of words (Rainer, 2005; Babić, 2002). The lexical meaning is regularly much more complex than the derivational meaning. The derivational meaning enables us to determine the type of semantic shifts between words used as stems for adding affixes and derivatives. In these processes, affixes play a very important role. The meaning of affixes is determined within derivational patterns, that is, when determining the meaning of the affix, it is necessary to observe the relationship between the meaning of stems and derivatives (cf. also Bauer and Valera, 2015). For example, Croatian noun *stolar* 'carpenter' is derived from the word *stol* 'table' and the suffix *-ar.* Thus, the derivational meaning of the noun *stolar* would be 'the one who makes/produces tables', and the semantic shift from the stem to the derivative is the one of 'male agent'. We can conclude that this particular meaning is actually the meaning provided by the suffix *-ar*. This conclusion is further supported by other words in which the meaning 'male agent' is mediated by the suffix *-ar*:

kuhar 'cook' (← *kuhati* 'to cook')
kipar 'sculptor' (← *kip* 'sculpture')
mesar 'butcher' (← *meso* 'meat')
ribar 'fisherman' (← *riba* 'fish')...

Although the lexical and the derivational meaning can be the same, as in *kuhar* 'the one who cooks', sometimes they differ in a way that lexical meaning becomes more diversified than derivational meaning. This is the case with the above mentioned noun *stolar*. It has broadened its meaning to denote male agents who produce any kind of wooden furniture, window frames or doors, not only tables. However, the meaning relation between the stem and the derivative is revealed through the derivational meaning. Thus, we have to take the derivational meaning into account when determining derivational patterns and semantic shifts within them, because this particular shift is usually mediated by the affix.

Additionally, the affixal meaning is determined according to only one of possible meanings of polysemous lexemes. For example, the word *upravljač* can denote both an agent 'controller' and an object 'control device'. However, since their morphological and derivational properties are the same, we didn't want to multiply number of entries in CroDeriv at this point. Thus, we mark only the most prominent meaning (i.e. the meaning that is listed first) as indicated by the extensive online dictionary of the Croatian language available at Hrvatski jezični portal (`hjp.znanje.hr`), and in the example above, only the meaning of an object will be marked. Although this approach could be criticized, we believe that this decision is methodologically justified when manually annotating the meaning of affixes in several thousands of words.[4]

Related to the principle of determining affixal meanings according to the one meaning of the polysemous lexeme, is the principle of distinguishing the meaning of the suffix from the meaning of derivatives. We believe that suffixal meanings lie between the meaning of stems and derivatives. Although the

[3]Only 20 most productive nominal suffixes were semantically analyzed at this phase of the research.

[4]It is possible that in the next phases of CroDeriv development lexical units will be analyzed for their polysemous structure. However, this information is already available in Croatian dictionaries, so we currently focus on the data which has not been available so far.

meaning of derivatives is extensively discussed in Croatian grammar books, general semantic categories such as agent, means or location are mediated by suffixes and therefore require more attention (see the examples with the suffix *-ar* above).

We distinguish suffixes that 1) change the part-of-speech category of stems, 2) modify the meaning of stems without changing their part-of-speech category and 3) modify the meaning of stems and change their part-of-speech category. Some suffixes only change the part-of-speech category of stems without any alternation or modification of their meaning. We treat these suffixes as the category of substantivizing, adjectivizing etc. suffixes. The prototypical example of the substantivizing suffix is the suffix *-je*, predominantly used for the derivation of gerunds with the meaning 'action', although its polysemous structure is more complex. Suffixes that modify the meaning of stem words without changing their part-of-speech category are the suffixes that are, for example, used for the derivation of nouns with marked meanings (diminutives, pejoratives, augmentatives) or suffixes used for the derivation of feminine/masculine pairs. The most complex role is played by suffixes that both change the meaning and the part-of-speech category of stems. Such is, for example, a suffix *-ač*, as in *bacač* 'thrower' ← *baciti* 'to throw'. In this derivational process the part-of-speech category was changed from the verb to the noun, and the meaning from 'action' into 'agent'.

Suffixal meanings are determined in respect to generalized semantic categories, as described in Filko (2020) for nouns, Šojat et al. (2012) for verbs and Filko and Šojat (2017); Bagasheva (2017) for adjectives. Generalized semantic categories are recognized as fundamental by numerous linguists dealing with affixal meanings. The most extensive list of affixal meanings is the list presented in Bagasheva (2017). The semantic categories in her list are determined for various language families, mainly for those in Europe. Generalized semantic categories are additionally specified and divided into subcategories when it is important to determine semantic differences among suffixes (see Šojat et al., 2012).

3 Affixal meanings in CroDeriv

As extensively described in Filko et al. (2020), CroDeriv data model enables lexemes to be explicitly annotated with three layers of description, i.e. morphological, word-formation, and compounding-derivational description. Word-formation description consists of sequence of clusters, which are multi-morphemic units corresponding to notions of stems and derivational affixes. Since they are associated with morphemes they consist of, clusters also serve as a link between morphological and word-formation layer of description.

Similar to morphemes, clusters are stored as independent units in CroDeriv database. They can be of different types so we distinguish stems, and prefixing and suffixing formants, i.e. affixal clusters, which roughly correspond to notions of stems, derivational prefixes and derivational suffixes from morphological theory, respectively. Clusters, but affixal clusters in particular, can be associated with multiple meanings due to polysemic nature of derivational affixes, as already elaborated in Section 2. Instances of clusters, which make up the word-formation description of a particular lexeme, have their associated meanings reduced to one, i.e. the one which is realized in that lexeme.

Compounding segments are objects that form the third layer of description, the compounding-derivational layer. They are composed of sequences of clusters in which there can be only one stem cluster, and one or more affixal clusters. Compounding segments serve as child members in derivational relations, thus allowing compound words to be a part of multiple derivational families, e.g. the lexeme *naredbodavac* 'commander' [lit. 'command-giver'] is a part of two derivational families, *red-* and *da-*. This is reflected in the fact that it is composed of two compounding segments: *naredbo* and *davac*. First compounding segment consists of the stem cluster *naredb*, and the interfixal cluster *o*. Second compounding cluster consists of the stem cluster *dav*, and the suffixal cluster *ac*. Because of their association with clusters on word-formation level, compounding segments also serve as a link between word-formation and compounding-derivational layer of description.

Additionally, we can identify one more, the fourth, lexical, layer on which grammatical information is added to the entire lexeme.

Figure 1 on the example of the word *mrtvačnica* 'mortuary' (← *mrtvac* 'dead'$_{N}$) in a simplified manner

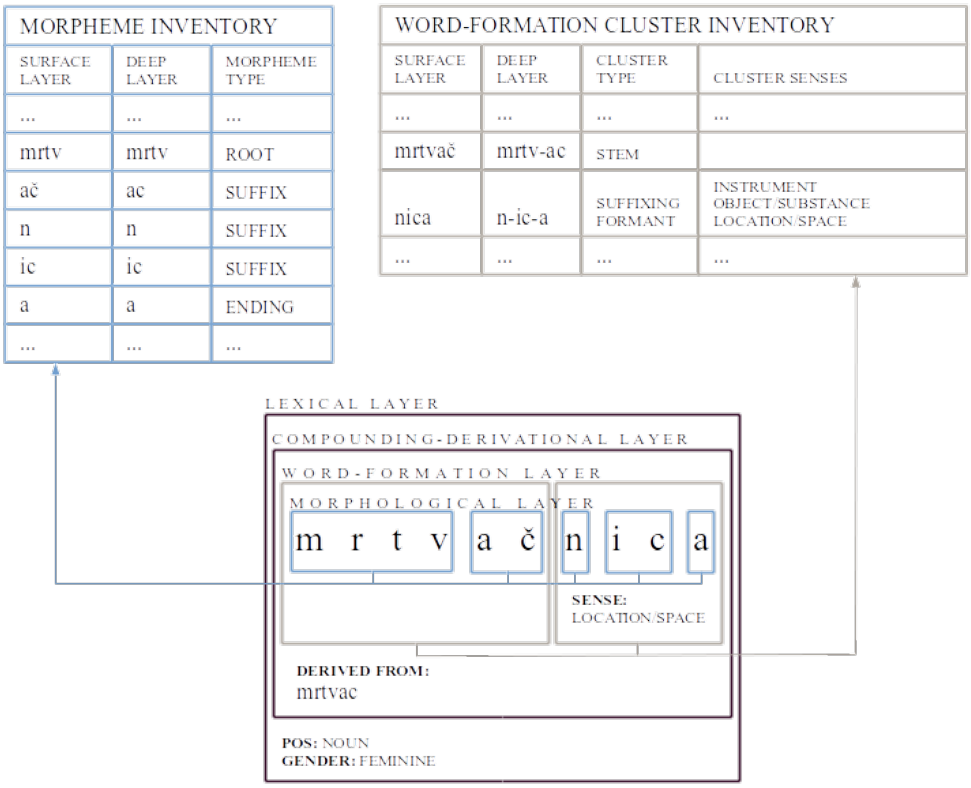


Figure 1: Schematic illustration of how CroDeriv data is organized

schematically illustrates the interconnections between the layers of description in CroDeriv, as well as the type of morphological information included in each of the layers.
The innermost set of borders around groups of letters in the center of the picture represents the morphological layer of description. These borders split the word in allomorphs which are then associated with the allomorphs and morphemes from the morpheme inventory. One layer up is the word-formation layer in which the borders enclose the word-formation units, i.e. "alloclusters". Similarly, they are associated with the "alloclusters" and clusters from the cluster inventory. At this level, affixal clusters are marked for their sense. The next layer is the compounding-derivational layer which, besides splitting the elements of a potential compound, contains information about the base word in the derivational process. The outermost layer is the lexical layer which adds lexeme's grammatical information.
As it can be seen from the illustration, in our annotation scheme layers are independent one from another, i.e. annotation on one layer can exist without the annotation on the other, but also higher layers have access to annotation on lower layers which makes the layers interconnected.

Up to this point, ca. 6500 lexemes in our database have been manually analyzed and annotated for word-formation, out of which, sense of the derivational affix was annotated for ca. 3000 lexemes. All publishing-ready CroDeriv data will be made available through regular contributions to the Universal Derivations dataset.

4 Semantic graphs

The interconnection of different layers of description in CroDeriv facilitates exploring other, not explicitly annotated, phenomena that occur at the intersections of these layers. One of them is derivational semantics which can be studied with the help of semantic graphs. The CroDeriv data model allows for structuring the derivational families in two graph representations. One we shall refer to as

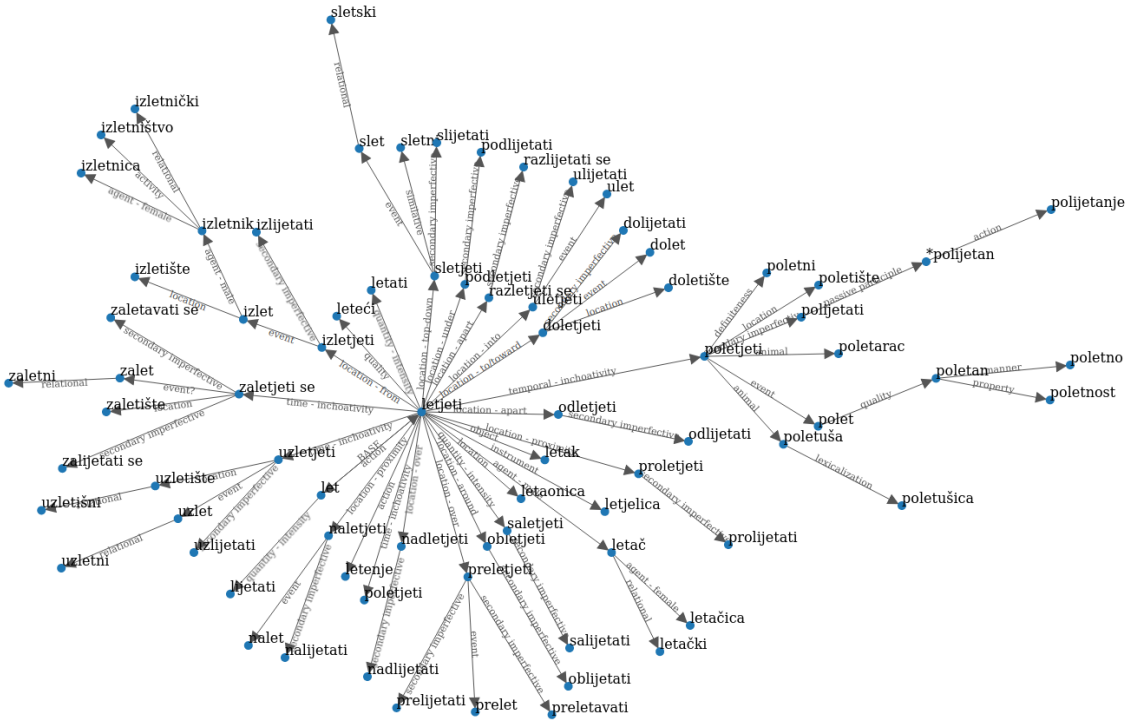


Figure 2: Lexeme-semantic representation of derivational family *let-*.

lexeme-semantic representation (Figure 2), and the other **structure-semantic representation** (Figure 3).

The **lexeme-semantic representation** is a directed acyclic graph with labeled nodes representing lexemes, and labeled edges representing derivational meaning of the lexeme they connect to. This representation is essentially a derivational graph with semantic labels attached to edges and is more-or-less in line with graphs created by Ševčíková and Kyjánek (2019) for Czech. In addition to providing a more informative representation of derivational families, these graphs also show the semantic motivation for the expansion of the derivational tree.

In the Figure 1 example, combining information from the word-formation and compounding-derivational layer enables labeling of the derivational link between the base word and the derivative as

$$mrtvac \xrightarrow{\text{LOCATION/SPACE}} mrtvačnica.$$

The **structure-semantic representation** is also a directed acyclic graph with semantic labels attached to edges, but here the nodes are labeled with derivational affixes involved in the last derivational step of a particular lexeme.

In the Figure 1 example, combining information from the word-formation layer of both base word and the derivative, and from the compounding-derivational layer of the derivative, reveals the derivational mechanism

$$\text{-}ac \xrightarrow{\text{LOCATION/SPACE}} \text{-}nica.$$

Structure-semantic graphs, as derivational generalizations of some sort, show derivational mechanisms used for semantic build-up. Having derivational trees represented in such way, by means of approximate graph matching algorithms, this directly allows for 1) quantifying the derivational similarity between derivational trees, 2) exploring the distribution of particular affixes involved in certain semantic change and vice-versa, and 3) calculating causal distribution of derivational mechanisms. The approximate graph

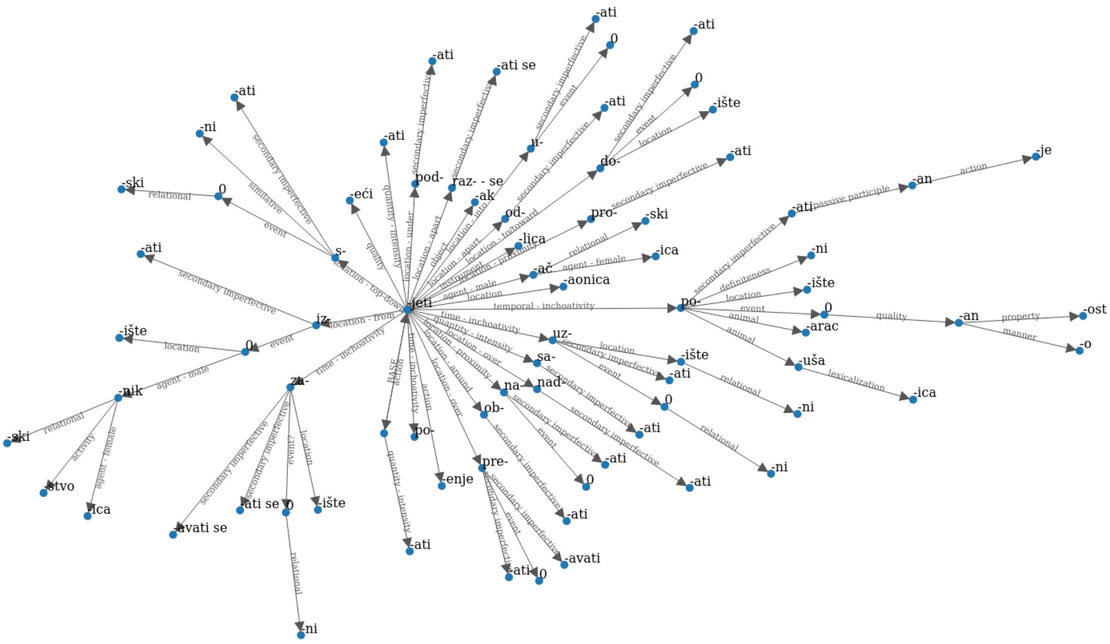


Figure 3: Structure-semantic representation of derivational family *let-*.

matching algorithms, such as G-Finder (Liu et al., 2019), TALE (Tian and Patel, 2008) and SAGA (Tian et al., 2007), facilitate inexact (sub)graph matching and are able to overcome problems of missing (or intermediate) nodes and edges, or different labels. For example, in diverse semantic categories, such as 'property' (see Section 2) this meaning can be expressed with various derivational suffixes, namely *-ina*, *-oća* and *-ost*, e.g. *uredan* 'tidy' → *urednost* 'tidiness' vs. *gust* 'thick' → *gustoća* 'thickness'. Although derivational affixes, i.e. nodes in their respective derivational graphs, are different, semantic pattern **quality** → **property**, i.e. the sequence of edges, is the same in both examples. The matching algorithm needs to be able to capture that similarity. The ability to match two not completely isomorphic graphs, allows for comparing derivational families and finding those that exhibit large derivational similarity among themselves. Further on, this representation facilitates distributional analysis of both derivational affixes and semantic changes in a direct derivational sequence. For example, for highly polysemous affixes, such as the nominal suffix *-ba* (see Section 2), it is possible to determine its productivity in each of the semantic categories across derivational families, while for diverse categories, such as 'property', it is possible to examine its distribution within affixes that can carry its meaning. Finally, based on analysis of several derivational families, it might be fruitful to investigate the surroundings of certain common derivational patterns for possible facilitators of certain tree expansions. As far as to our best knowledge, this type of derivational semantic analysis has not yet been attempted in Croatian morphological theory.

5 Concluding remarks

This paper presents a novel way of representing CroDeriv data for the purpose of investigating word-formation mechanisms in the Croatian language. It also presents a new line of research on Croatian derivational morphology with the help of generalized morphological data in the form of semantic graphs. However, it is important to mention that the data annotated for word-formation so far was chosen on the criterion of belonging to large and productive derivational families. Our intention goes towards annotating the entire CroDeriv dataset in this way.

References

Ju. D. Apresjan. 1974. Regular Polysemy. *Linguistics* 12(142). https://doi.org/10.1515/ling.1974.12.142.5.

Mark Aronoff and Kirsten Anne Fudeman. 2011. *What is morphology?*. Fundamentals of linguistics. Wiley-Blackwell, Chichester, West Sussex, U.K. ; Malden, MA, 2nd ed edition. OCLC: ocn608491860.

Stjepan Babić. 2002. *Tvorba riječi u hrvatskome književnome jeziku*. Number knjiga 2 in Velika hrvatska gramatika. Hrvatska akademija znanosti i umjetnosti : Nakladni zavod Globus, Zagreb, 3., poboljšano izdanje edition. OCLC: ocm53895583.

Alexandra Bagasheva. 2017. Comparative semantics concepts in affixation. In Juan Santana-Lario and Salvador Valera-Hernández, editors, *Competing Patterns in English Affixation*, Peter Lang, Bern : Berlin : Bruxelles : Frankfurt am Main : New York : Oxford : Wien, Linguistic Insights, pages 33–66.

Laurie Bauer and Salvador Valera. 2015. Sense Inheritance in English Word-Formation. In Laurie Bauer, Lívia Körtvélyessy, and Pavol Štekauer, editors, *Semantics of Complex Words*, Springer, Heidelberg : New York : Dordrecht : London, number 3 in Studies in Morphology, pages 67–84.

Matea Filko. 2020. *Unutarleksičke i međuleksičke strukture imeničkoga dijela hrvatskoga leksika (Intralexical and Interlexical Structures of the Nominal Part of the Croatian Lexicon)*. Phd thesis, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.

Matea Filko and Krešimir Šojat. 2017. Expansion of the Derivational Database for Croatian. In Eleonora Litta and Marco Passarotti, editors, *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*. EDUCatt, Milan, pages 27–37.

Matea Filko, Krešimir Šojat, and Vanja Štefanec. 2020. The Design of Croderiv 2.0. *The Prague Bulletin of Mathematical Linguistics* 115:83–104. https://doi.org/10.14712/00326585.006.

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5):125–168.

Lukáš Kyjánek. 2018. *Morphological resources of derivational word-formation relations*. Technical Report 61, ÚFAL - Charles University, Prague.

Adrienne Lehrer. 2003. Polysemy in derivational affixes. In Brigitte Nerlich, Zazie Todd, Vimala Herman, and David C. Clarke, editors, *Polysemy. Flexible Patterns of Meaning in Mind and Language*, De Gruyter Mouton, Berlin : New York, pages 218–232.

Rochelle Lieber. 2004. *Morphology and lexical semantics*. Cambridge University Press, New York.

Lihui Liu, Boxin Du, Jiejun xu, and Hanghang Tong. 2019. G-Finder: Approximate Attributed Subgraph Matching. In *2019 IEEE International Conference on Big Data (Big Data)*. pages 513–522. https://doi.org/10.1109/BigData47090.2019.9006525.

Karel Pala and Pavel Šmerk. 2015. Derivancze — Derivational Analyzer of Czech. In Pavel Král and Václav Matoušek, editors, *Text, Speech, and Dialogue: 18th International Conference, TSD 2015*. Springer, Berlin: Heidelberg, pages 515–523. https://doi.org/10.1007/978-3-319-24033-6_58.

Ida Raffaelli. 2015. *O značenju: uvod u semantiku*. Biblioteka Theoria / Matica hrvatska Novi niz. Matica hrvatska, Zagreb.

Franz Rainer. 2005. Semantic change in word formation. *Linguistics* 42(2):415–441.

Franz Rainer. 2014. Polysemy in Derivation. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, Oxford University Press, Oxford, Oxford Handbooks in Linguistics, pages 338–353.

Yuanyuan Tian, Richard C. McEachin, Carlos Santos, David J. States, and Jignesh M. Patel. 2007. SAGA: a subgraph matching tool for biological graphs. *Bioinformatics* 23(2):232–239. https://doi.org/10.1093/bioinformatics/btl571.

Yuanyuan Tian and Jignesh M. Patel. 2008. TALE: A Tool for Approximate Large Graph Matching. In *2008 IEEE 24th International Conference on Data Engineering*. pages 963–972. ISSN: 2375-026X. https://doi.org/10.1109/ICDE.2008.4497505.

Magda Ševčíková and Lukáš Kyjánek. 2019. Introducing Semantic Labels into the DeriNet Network. *Journal of Linguistics* 70(2):412–423.

Krešimir Šojat, Matea Srebačić, and Tin Pavelić. 2014. CroDeriV 2.0.: Initial Experiments. In Adam Przepiórkowski and Maciej Ogrodniczuk, editors, *Advances in Natural Language Processing*, Springer International Publishing, Cham, volume 8686, pages 27–33. https://doi.org/10.1007/978-3-319-10888-9_3.

Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling* 0(1):111. https://doi.org/10.15398/jlm.v0i1.34.

Krešimir Šojat, Matea Srebačić, and Vanja Štefanec. 2013. CroDeriV i morfološka raščlamba hrvatskoga glagola. *Suvremena lingvistika* 75:75–96.

Splitting and Identifying Czech Compounds: A Pilot Study

Emil Svoboda
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
svoboda@ufal.mff.cuni.cz

Magda Ševčíková
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
sevcikova@ufal.mff.cuni.cz

Abstract

We present pilot experiments on splitting and identifying Czech compound words. We created an algorithm measuring the linguistic similarity of two words based on finding the shortest path through a matrix of mutual estimated correspondences between two phonemic strings. Additionally, a neural compound-splitting tool (*Czech Compound Splitter*) was implemented by using the Marian Neural Machine Translator framework, which was trained on a data set containing 1,164 hand-annotated compounds and about 280,000 synthetically created compounds. In compound splitting, the first solution achieved an accuracy of 28% and the second solution achieved 54% on a separate validation data set. In compound identification, the *Czech Compound Splitter* achieved an accuracy of 91%.

1 Introduction

Compounding refers to "the formation of a new lexeme by adjoining two or more lexemes" (Bauer, 2003, p. 40). For many languages, including Sanskrit, English and German, the process has been mapped and modelled extensively in static data resources and procedural tools, but this is not the case for Czech.

The present paper focuses on compounding in Czech, which is a language where compounds are nearly always represented in writing as a single string of graphical symbols unbroken by whitespace (from here: *graphical word*). The problem we tackle is twofold: a) upon being given a graphical word, to decide whether or not it is a compound; and b) upon being given a confirmed compound, to return the citation forms of its base words (from here: *parent words* or *parents*). Task a) will be referred to as *compound identification* and is approached as an instance of binary classification; and task b) will be referred to as *compound splitting*. The tasks can be seen as part of the more general problem of morphological segmentation, which refers to the splitting of a word into morphemes (affixes, roots, endings).

The following study constitutes, to the best of our knowledge, the first foray into automatic compound identification and compound splitting in Czech. Section 2 is a non-exhaustive overview of existing accounts of compounding relevant to this study. After a brief report on the compilation of the data set including examples of some challenges of Czech compounding (Section 3), the experiments are described and their performance is compared in Section 4. The solutions we implemented include a baseline solution which performs compound splitting only. A more advanced approach based on phonemic string similarity we call *Interlexical Matrices of Likeness*, or $IML()$, is also limited to compound splitting. Finally, a deep learning based tool dubbed *Czech Compound Splitter* was trained, which simultaneously carries out both compound identification and compound splitting. Section 5 contains the summary of this study.

2 Related work

2.1 Compounding in Czech

Theoretical descriptions of compounding in Czech are optimized for human readers. Bozděchová (1997) distinguishes two types of compounding in Czech, depending on whether the words entering the composition are formally modified or not. *Compounding proper*, which requires morphological adjustment of

the input words, and *compounding improper*, which is the result of simple concatenation of a syntactic phrase with no morphological adjustments. In addition, Bozděchová puts forth a multi-level classification, starting from the part-of-speech category of the output compound and then proceeding to semantic criteria (considering the meanings of the input items, of the output compounds and the relationship between the output and the inputs). Moreover, it is taken into account whether the compound is a result of composition only or whether also other word-formation processes (derivation, conversion) were at play. For instance, the compound adjective in (1) was coined through composition proper, when the ending *-ý* in the first input adjective (*tmavý* 'dark') was dropped and an *-o-* interfix was used to concatenate it with the second adjective (*modrý* 'blue'). In (2), the input adjective (*tvrdý* 'hard') undergoes a similar formal modification, but the second item (the noun *hlava* 'head') is converted into an adjective through replacing the nominal ending by an adjectival one (*hlava* 'head' → *-hlavý* 'headed', which cannot be used separately in Czech). Analogically to this example of compounding and conversion in one step, in (3) the compound is formed through compounding and derivation (i.e., the addition of the agent suffix *-ec* to the input verb). A straightforward example of composition improper is the concatenation of two nouns to a compound adverb in (4). A reversal of the ordering of the input words is permissible, resulting in the compound verb in (5).

(1) ***tmavý*** + ***modrý*** → ***tmav|-o-|modrý***
dark.ADJ blue.ADJ dark-blue.ADJ

(2) ***tvrdý*** + ***hlava*** → ***tvrd|-o-|hlavý***
hard.ADJ head.NOUN stubborn.ADJ

(3) ***černý*** + ***odít*** → ***čern|-o-|oděnec***
black.ADJ dress.VERB black dressed man.NOUN

(4) ***chvála*** ***Bohu*** → ***chvála|bohu***
praise.NOUN God.NOUN-DAT.SG thankfully.ADV

(5) ***přát*** ***blaho*** → ***blaho|přát***
wish.VERB wellness.NOUN-ACC.SG congratulate.VERB

In a recent paper on compounding in West Slavic languages, Ološtiak and Vojteková (2021) restrict themselves to non-native compounds, especially to compounds of partially or fully Greek-Latin origin (from here: *neoclassical compounds*). Four types of word-formation formants are distinguished, namely bases, baseoids, affixoids, and affixes. Bases are items that can appear freely and carry lexical meaning (*terapie* 'therapy', like in *ergoterapie* 'occupational therapy'); baseoids are items that do not appear freely, but carry lexical meaning regardless (*ergo-*, in *ergoterapie* 'occupational therapy'), and affixoids are items that are diachronically lexical, but have gradually lost their ability to appear independently and have generalized their meaning enough to effectively behave like derivational items. Three types of compounds are delimited according to the type of formants they involve. *Proper compounds*[1] are characterized as being composed of two bases (e.g. *sérum* 'serum' + *pozitivní* 'positive' → *séropozitivní* 'seropositive'). *Semi-compounds* are composed of one base and one baseoid (e.g. *krypto-* 'crypto-' + *politika* 'politics' → *kryptopolitika* 'cryptopolitics'). Finally, *quasi-compounds* are composed of two baseoids (e.g. *eko-* 'eco-' + *-logie* '-logy' → *ekologie* 'ecology').

Our conceptualization of neoclassical compounds is largely congruent with this classification, with a reduction in granularity. Everything they consider to be a *baseoid* and most of what the authors consider to be an *affixoid* is considered to be a neoclassical constituent by us. This creates a small amount of inconsistency in exchange for increased simplicity and reduced granularity. For instance, we consider the formant *-pidi-* (considered an affixoid by the authors) to be a neoclassical constituent, because it behaves almost exactly the same way as *-mini-*, with regards to both semantics and behaviour within word formation. We prefer this interpretation despite the fact that it is traced back to the Czech noun *píď* 'span' (unit of length). We also systematically interpret neoclassical constituents as identical whenever their etymology and semantics allow for it, even under circumstances where they undergo formal changes. For instance, the first element of *logografie* 'logography' (*logo-*) and the second element of *sociologie*

[1]The usage of this term by these authors is distinct from Bozděchová's proposal above.

'sociology' (*-logie*) are seen to be the same, since they both ultimately descend from the same Greek root. In our data, they are represented by the string *-log-*, cf. Section 3.2 for more details.

Štichauer (2013) presents an attempt to classify Czech compounds using three levels of categorisation, akin to the way Romance compounds are handled by Bisetto and Scalise (2005). The first level is the distinction between *coordinative*, *subordinative* and *attributive* compounds. The second level distinguishes between *exocentricity* and *endocentricity*, or *headedness* – in other words, whether or not the compound has a semantic head. The third level distinguishes between every possible combination of part-of-speech category of the input words and the part-of-speech category of the output compound in the format $[X + Y]_Z$, where X and Y stand for the input part of speech and Z stands for the part of speech of the resulting compound.

2.2 NLP approaches toward compounding

Czech has neither a static word formation data resource with a notable amount of parent-linked compounds nor a procedural tool for identifying or splitting compounds. Derivational Analyzer of Czech (Derivancze; Pala and Šmerk, 2015), as its name suggests, is limited to derivational relations in the lexicon of Czech. Another word-formation resource for the language, DeriNet, maps derivation by means of linking words to the words they are respectively derived from all the way to their roots. DeriNet, in spite of its name, is additionally equipped for handling compounding as well, in that its data format allows for a single lexeme to have multiple parents. DeriNet version 2.0 (Vidra et al., 2019) contained $33,932$ lexemes identified as compounds, out of which $1,252$ had their respective parent words identified. The work on this paper has contributed to the release of DeriNet version 2.1 by identifying the parents of $1,439$ compounds. The new version therefore contains a total of 2691 compounds with identified parents. (Vidra et al., 2021)

The situation regarding the computational handling of compounds is different in some other languages. Specifically, GermaNet (Hamp and Feldweg, 1997) contains nearly $100,000$ split nominal compounds, and CELEX (Baayen et al., 1996) includes $71,249$ split compounds for Dutch, $12,853$ split compounds for English and $19,768$ for German, all done manually.

Procedural compound splitting has been successfully demonstrated to be feasible in several languages. Henrich and Hinrichs (2011) linked German nominal compounds to their respective parents in GermaNet using an ensemble of pattern-matching models with an accuracy of 92%. Sugisaki and Tuggener (2018) achieved an F1-score of 92% for finding split-points in German compounds using an unsupervised approach, although they also restricted their efforts to noun-headed compounds only. Ma et al. (2016) achieved an accuracy of 95% using a neural approach trained on the aforementioned GermaNet. Their model performed both splitting and identification of compounds, with the accuracy being an aggregated score of both. Krotova et al. (2020) achieved an accuracy of 96% with a deep neural model trained on GermaNet data, again restricting themselves to nominal compounds.

A significant amount of research has been dedicated to the study of Sanskrit compounds. This ranges from early, relatively simple rule-and-lexicon based attempts by Huet (2005), who lists no accuracy in his study, to Hellwich and Nehrdich's (2018) deep-learning solution trained on a corpus of 560,000 Sanskrit sentences with its compound split points annotated, achieving an accuracy of 96%.

As for other languages, Clouet and Daille (2014) achieved F1-scores for finding split-points in English and Russian compounds of 80% and 63% respectively, using a corpus-based statistical approach on manually split compounds. Russian is important for this study, because it is a Slavic language like Czech and thus this result is the most comparable to the ones presented here.

3 Compilation of the data set

3.1 Challenges

What follows is a qualitative analysis of some formal difficulties that regularly appear in Czech compounding. Please note how the phenomena often accumulate within the same word, and that the list is not by any means exhaustive. None of these nor any similar difficult cases were dropped from the data.

For data-based approaches, the simplest case seem to be compounds formed by simple concatenation (cf. compounding improper in the literature discussed above). For instance, the adjective in (6) corresponds

directly to the syntactic phrase *vždy zelený* ‘always green’. In (7), neither input word undergoes any morphological change during the composition, which is characteristic for composition improper, but the output noun cannot be associated with no such phrase, which is typical of composition proper. From the perspective of algorithmic splitting, however, the two compounds are very much alike, in that the procedure of finding their parents consists merely of finding the appropriate split point.

(6) ***vždy*** ***zelený*** → ***vždy|zelený***
always.ADV green.ADJ-NOM.SG evergreen.ADJ

(7) ***garáž*** + ***mistr*** → ***garáž|mistr***
garage.NOUN master.NOUN garage supervisor.NOUN

An interfix is added between the two input words in other compounds, usually *-o-* or *-i-*. This interfix replaces the inflectional ending of any non-final parent; cf. the ending *-a* in the feminine noun *ryba* ‘fish’ in (8). Additionally, stem allomorphy often appears. It may takes the form of vowel alternation, for example /e/ → ∅, like in (9).

(8) ***ryba*** + ***lov*** → ***ryb|-o-|lov***
fish.NOUN hunt.NOUN fishery.NOUN

(9) ***krev*** + ***tok*** → ***krv|-o-|tok***
krev.NOUN flow.NOUN bloodflow.NOUN

As mentioned above, compounding and conversion (or derivation) in one step is possible, as exemplified above and here in (10). Stem alternation may take place, like in (11), where a case of stem vowel alternation (/e/ -> ∅ and /e:/ → /o/), a stem consonant alternation (/s/ → /d/), an interfix, compounding and conversion in one step all occur at the same time. Note that an alternative analysis of the compounds in (8) and (9) can be proposed that would parallel (11): *krev* ‘blood’ + *téct* ‘to flow’ → *krvotok* ‘bloodflow’, *ryba* ‘fish’ + *lovit* ‘to hunt’ → *rybolov* ‘fishery’. In the data we use in our experiments, both analyses are captured (see Section 3.2).

(10) ***modrý*** + ***oko*** → ***modr|-o-|oký***, but no ****oký***
blue.ADJ eye.NOUN blue-eyed.ADJ

(11) ***pes*** + ***vést*** → ***ps|-o-|vod***, but no ****vod***
dog.NOUN lead.VERB dog handler.NOUN

In (12), the compound is traced back to the noun phrase *chtivý holek* ‘wanting of girls’, with its original ordering switched. Additionally, there are compounds that cannot be meaningfully split into two parents; cf. the compound in (13) which is composed of a multi-word numeral expression (*dvě a půl* ‘two and a half’) and the final part which was converted from a noun (*léto* ‘year.NOUN’ → *-letý* ‘-year.ADJ’).

(12) ***chtivý*** ***holek*** → ***holek|chtivý***
wanting.ADJ girl.NOUN.GEN.PL wanting girls.ADJ

(13) ***dvě*** + ***a*** + ***půl*** + ***léto*** → ***dva|a|půl|letý***, but no ****letý***
two.NUM and.CONJ half.NUM summer.ADJ two-and-a-half-year-old.ADJ

The so-called *neoclassical compounds* constitute what Ološtiak and Vojteková (2021) consider semi-composition and quasi-composition. The noun *sociologie* ‘sociology’ in (14) is an example of quasi-composition in this framework. In a broader sense, chemical compounds satisfy the definition of semi-composition, as in (15).

(14) ***-soci-*** + ***-log-*** → ***soci|-o-|logie***, but no ****-soci-*** nor ****-log-***
-soci-.NEOCON -log-.NEOCON sociology.NOUN

(15) ***-tetra-*** + ***chlor*** + ***ethylen*** → ***tetra|chlor|ethylen***, but no ****tetra***
-tetra-.NEOCON chlorine.NOUN ethylene.NOUN tetrachlorethylene.NOUN

3.2 Manual annotation of DeriNet data

The compilation procedure began by extracting $1,500$ words from the DeriNet word-formation resource that had previously been labelled as having compound status. As their parent words had not been yet identified, so this had to be done by hand. 53 were dropped, because some had been labelled as compounds mistakenly (*levopimar*, a medicine brand name), or are derivatives of compounds (e.g. the adverb *velechytře* derived from the adjective *velechytrý* 'very clever'). After this cleanup process was done, $1,447$ compounds remained in the data set. 20% of the data set compounds was held out for the purposes of validation. The training set therefore consisted of $1,158$ hand-annotated compounds, while the *holdout data set* set consisted of 289 hand-annotated compounds. The holdout set was further split in half. The first half, the *test set*, was used to determine when to stop training *Czech Compound Splitter*. The performance of all the approaches presented here was evaluated on the other half, the *validation set*.

Neoclassical constituents, as they do not have an agreed-upon citation form, are labelled with hyphens on both sides, maintaining the original Greek stem as bare as possible. In order to reflect the consistency described in Section 2.1, we label both the second constituent of *sociologie* 'sociology' and the first constituent of *logografie* 'logography' as *-log-*. We keep the 'o', if it resolves an otherwise arising ambiguity. For example, we label the first element of *bigamie* 'bigamy' as *-bi-* and the first element of *biologie* 'biology' as *-bio-*, preferring this slight annotation inconsistency over label ambiguity. Greek orthography is respected as much as possible, so we respect the distinction between τ and θ, so the first element of *teologie* 'theology' is labeled as *-theo-* (not *-the-*, as that would be ambiguous with the root of *teorie* 'theory'). Zero ablaut forms are preferred as labels of neoclassical compounds, unless this would result in an asyllabic label. Thus, both the first element of *gastronomie* 'gastronomy' and the second element of *melanogaster* (the epithet of the fruitfly *Drosophila melanogaster*) is labeled as *-gastr-*, but the first element of *gonokok* 'gonococcus' and the second element of *mutagen* 'mutagen' are labelled as *-gen-*.

The first two of the three algorithmic solutions require a lexicon to find potential parent-candidates in. DeriNet was used as a basis for this lexicon, but it restricts itself to nouns, verbs, adjectives, and adverbs. Lexemes of other part-of-speech categories were extracted from a Czech inflectional dictionary (MorfFlex; Hajič et al. 2020) and added into the lexicon. Finally, all neoclassical constituents identified during the manual annotation were added into the lexicon.

3.3 Generation of synthetic data

Because the hand-annotated data set of compounds obtained from DeriNet is too small to reliably train a deep learning model, we simulated various compound formation procedure that take place in Czech. For example, in (16) we see the process of taking a random adjective stripped of its ending and concatenating it with an *-o-* interfix and with another random adjective. The output is usually nonsensical, like in the example, but formally correctly formed.

(16) **Adjective 1** + ***-o-*** + **Adjective 2** → **Compound Adjective**
důležitý + ***-o-*** + ***neomylný*** → ***důležitoneomylný***
important.ADJ infallible.ADJ important-infallible.ADJ

For the purposes of training *Czech Compound Splitter*, we simulated a number of such compound formation procedures in Python using randomly selected lexemes from DeriNet, creating a data set of about $280,000$ synthetic compounds. The compound part of the training data set therefore consisted of this synthetic data set combined with all of the hand-annotated compounds apart from the holdout described in Section 3.4.

3.4 Evaluation methodology

For about 38% of the hand-annotated compounds in our dataset, there was ambiguity as to which parents they should be linked to. For instance, *monoprogramový* 'having a single programme' may be considered to be either composed of the neoclassical constituent *-mon-* and the adjective *programový*,

or it alternatively may be composed of *-mon-* and the noun *program*, which would be derivation and compounding in one step. For the purposes of evaluation, both were considered to be correct splittings.

Additionally, a more relaxed metric was proposed which considers a predicted parent-candidate to be correct if it belongs to the same morphological family as the annotated parent. This metric is referred to as *root accuracy*, because all items of a morphological family are represented as a tree structure with the unmotivated word as the root node in DeriNet. DeriNet data are used to determine whether or not the predicted parent-candidate shares the same morphological family as the annotated parent. The solutions described in the following section exhibit different weaknesses and strengths.

4 Experiments

4.1 Baseline solution

This is a naive algorithm only intended as a baseline to help provide context for the performances of the other solutions. This solution assumes the given compound has two parents. It attempts to find an 'o' grapheme in the middle third of the input word. If it finds one, it splits the word on this 'o', creating two subwords. If no 'o' is found, it does the same with 'i'. If no 'i' is found, it simply splits the input in the middle, if the number of graphemes in the graphical word is even, the left subword ends up being the longer one. Between each subword and every word in the lexicon, Levenshtein (1966) distance is calculated, and the word with the smallest distance from the subword is selected. Please refer to Table 4 to see its performance.

4.2 The $IML()$-based heuristic algorithm

The second attempt to split compounds is based on a phonological similarity measurement function developed specifically for this purpose. We developed a function that takes two words as input and returns a rational number representing the total degree of phonological similarity between the two words. We then attempted to find pairs of words which, when concatenated, exhibited a low degree of $IML()$ similarity with the compound in question. $IML()$ cannot perform compound identification, because the method already assumes the input word has exactly two parents.

4.2.1 The $IML()$ matrix function

We began by manually defining a phonemic correspondence weight by hand for each possible pair of phonemes in Czech. The minimum weight is 0, which is the correspondence weight strictly between a phoneme and itself, and the maximum weight is 1, which is the correspondence weight between a phoneme and a phoneme it never alternates with, like between /a/ and /t/. Note that this relationship is asymmetric by design, because we estimated that, for example, /h/ $\rightarrow$ /z/ is much more common than /z/ $\rightarrow$ /h/. From this, it directly follows that the ordering of the words that are input into the $IML()$ function matters. There are 32 phonemes in the Czech language, so it follows that the total amount of phonemic correspondences equals $32^2 = 1024$. This can be described by a square matrix, where each column and row corresponds to one of the Czech phonemes and each element describes the correspondence weight between the Czech phonemes. This is what we call a *correspondence matrix*. Part of the matrix used in this study is shown in Table 1. Note that the diagonal is composed entirely of zeroes, and that the matrix is not symmetric with respect to said diagonal, which reflects the asymmetric nature of Czech phoneme alternation described in the previous paragraph.

The $IML()$ similarity measurement function takes two words, transcribes both of them phonologically, and uses the values found in the *correspondence matrix* to build a separate matrix of correspondence weights between every single pair of phonemes from the two input graphical words. The cheapest path through it is found, beginning in the top left corner of the matrix, and ending in the bottom right corner. We used the $A*$ algorithm, an extension of Dijsktra's algorithm, to find the shortest path (Hart et al., 1968).

The lower the output value, the higher the similarity, with $IML(word_1, word_2)$ being equal to 0 if and only if $word_1 = word_2$, because the correspondence weight between a pair of phonemes is zero if and

	/t/	/n/	/r/	/s/	/z/	/ts/	...
/t/	0	0.8	1	0.8	0.9	0.8	...
/n/	0.7	0	0.9	1	1	1	...
/r/	0.7	0.9	0	0.9	1	1	...
/s/	0.6	1	0.7	0	0.2	0.6	...
/z/	0.8	0.3	0.9	0.2	0	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Table 1: *S*ample of the referential matrix of correspondence weights between pairs of Czech phonemes.

IML(*černomodrý*,*černý* + *černý*) = 5.8
IML(*černomodrý*,*černý* + *červený*) = 5.0
IML(***černomodrý***,***černý* + *modrý***)= **0.6**
IML(*černomodrý*,*červený* + *černý*) = 5.7
IML(*černomodrý*,*červený* + *zelený*) = 6.9
IML(*černomodrý*,*červený* + *modrý*) = 2.6
IML(*černomodrý*,*modrý* + *černý*) = 9.6
IML(*černomodrý*,*modrý* + *zelený*) = 9.8
IML(*černomodrý*,*modrý* + *modrý*) = 10.6

Table 2: Sample of the algorithm's functioning, without the heuristic filter.

only if the two phonemes in the pair are identical – which is why the diagonal of the *referential matrix* is composed of zeros, and in that the only zeros in the *referential matrix* are located on the diagonal.

4.2.2 The heuristic

Based on this similarity function, we were able to find the pair of words from the lexicon mentioned previously which, when concatenated, exhibited the highest similarity with the compound word in question. The algorithm therefore requires the compound in question and a lexicon to find its parents in. A visual demonstration of the idea behind the algorithm with the word *černomodrý* 'black and blue' and a toy lexicon can be viewed in Table 2. The table shows the outputs of the $IML(compound, word_1 + word_2)$ calculations for each word pair from the {*černý*, *červený*, *modrý*} {'black', 'red', 'blue'} lexicon. The algorithm generates all pairs of lexemes from a given lexicon, concatenates them and calculates $IML()$ for each pair. It then (correctly in this case) selects the pair with the smallest value. The problem is that the size of our lexicon ultimately exceeded $800,000$ lexemes, meaning that every time a compound is split, over $800,000^2 = 6.4 \times 10^{11}$ interlexical matrices need to be built and run through the $A*$ pathfinding algorithm.

A heuristic filter was therefore added. For this purpose, a variant of the $IML()$ function, the $IML_{sub}()$ function, was defined. The two functions are similar with two key differences. First, in the case of the $IML_{sub}()$, the cheapest path does not have to reach the bottom right corner of the matrix. Instead, the path's total cost is calculated whenever it reaches either the right or bottom edge of the *interlexical matrix*. $IML_{sub}(word_1, word_2)$ returns the *degree* to which $word_2$ is a fuzzy substring of $word_1$, with respect to their phonological similarity. Second, the pathfinding algorithm used in $IML_{sub}()$ is not $A*$, but a best-first solution. This makes $IML_{sub}()$ significantly faster than $IML()$, because the whole *interlexical matrix* need not be constructed beforehand. Only word pairs $(lexeme_1, lexeme_2)$ which satisfied the following conditions were selected:

1. $First2Chars(lexeme_1) = First2Chars(compound)$,
2. $CountSyl(lexeme_1 + lexeme_2) \geq CountSyl(compound)$,
3. $IML_{sub}(lexeme_1) \leq 2.2 \ \ \& \ \ IML_{sub}(lexeme_2) \leq 2.2$,

where $First2Chars()$ is a function which returns the first two graphical characters of a given graphical word, $CountSyl()$ counts the syllables of the given graphical word (assuming it is a Czech word) and

Compound	English translation	*CCS* (incorrect) splitting	Correct splitting
dlouhohořící	'long-burning'	*dlouhohořící*	*dlouho* + *hořící*
CCS returns the original string, performing no splitting.			
osmiramenný	'eight-armed'	*osm* + *ramenný*	*osm* + *rameno*
CCS returns a non-existing derivative of an existing word.			
petrogeneze	'petrogenesis'	*-petro-* + *geneze*	*-petr-* + *geneze*
CCS includes the interfix in one of the parents.			

Table 3: A sample of the errors *Czech Compound Splitter* (CCS) typically makes.

$compound$ is the input compound being split.

4.2.3 Output evaluation

This pair of words then constituted the predicted parents. The performance of this method in compound splitting can be found in Table 4. The application of the algorithm seems to be much less practical than that of *Czech Compound Splitter*, because it takes about ten to fifteen minutes to split a single compound on a single processor given a lexicon of our size, despite the fact that the algorithm's asymptotic time complexity (even without the heuristic) is $O(n) = n^2$, where n refers to the size of the lexicon. The matrix building step takes $|word1| \times |word2|$ correspondence matrix lookup operations, but because the step occurs exactly once for each parent-candidate pair, it constitutes a constant, and is therefore by convention omitted when assessing asymptotic time complexity. It is additionally of interest that the root accuracy of this method was higher by 11 percentage points than its raw accuracy. Error analysis revealed that this increase is primarily caused a common error where a substring of a compound is homonymous to a noun derived from an adjective, while that adjective is the parent. For example, *bíločerný* 'black and white' is split into the noun *bílo* 'whiteness' and the adjective *černý* 'black', while two adjectives (*bílý* 'white' and *černý* 'black') are the correct parents.

4.3 Czech Compound Splitter

Because the performance and practicality of the $IML()$-based heuristic algorithm was deemed unsatisfactory, a neural compound splitting tool we named *Czech Compound Splitter* was created. It decides if an graphical word is a compound and if so, it returns its predicted parent words, all in one step. If the graphical word is identified as a compound, it returns its parents separated by spaces. The estimated number of parents is thus the number of spaces in the output $+1$, and the status of a compound is determined if this number is greater than 1.

The tool was created by using the Marian machine translation framework developed by Microsoft (Junczys-Dowmunt et al., 2018) to build a model and train it. This was done by feeding the model a parallel corpus of input and output data, where the model is trained to take an element of the input data, which was a Czech word, and the output was either the single derivational parent of that word if it was a non-compound, or all of the parents of that word separated by spaces if it was a compound. For example, *Czech Compound Splitter* was trained to return *kov* 'metal' upon being given the graphical word *kovový* 'made of metal', and to return *uhlík vodík* 'carbon hydrogen' upon being given the graphical word *uhlovodík* 'carbohydrate'. The non-compounds and their parents were taken from DeriNet.

The total training data set for *Czech Compound Splitter* consisted of:

- $1,164$ genuine compounds, with their splittings
- $280,000$ synthetic compounds, with their splittings
- the near entirety of DeriNet's non-compounds, with their derivational parents

The rest of DeriNet's non-compounds, totalling 144 lexemes, was held-out in order to test the performance of *Czech Compound Splitter* in compound identification.

Method	1st parent accuracy	2nd parent accuracy	Overall accuracy	Overall root accuracy
Baseline	22%	42%	11%	16%
$IML()$	42%	66%	24%	39%
Czech Compound Splitter	61%	66%	54%	61%

Table 4: Overall performances the three solutions exhibited.

4.3.1 Model evaluation

In compound identification, *Czech Compound Splitter* achieved an accuracy of 92% and an F1-score of 91%. Its performance in compound splitting can be found in Table 4. We see that root accuracy is just barely higher than accuracy. Error analysis reveals that this was due to the fact that a large proportion of the mistakes *Czech Compound Splitter* made because it often did not recognize the input as a compound. Similarly, it frequently returned a nonsensical string that is not a Czech word; see a sample of errors in Table 3.

It is worth noting that *Czech Compound Splitter* made only a single false positive error, meaning that it almost never labelled a non-compound as a compound. This suggests that it primarily recognizes compound status by detecting lexical-seeming substructures, as opposed to focusing on surface-level criteria like character length or the presence of an *-o-* interfix. *Czech Compound Splitter* run on a single GPU takes about 0.2 seconds to perform a single identification and splitting. The entire compiled model is about 300 MB in size, making its distribution as a Python package feasible, especially since it can be compiled to run on CPUs as well.

5 Conclusions

We present the results of the first attempts to automatically identify and split Czech compounds. While there has been a lot of attention invested into automatic compound splitting in languages such as German or Sanskrit, in the Slavic languages, the topic has largely, though not completely, been overlooked. We have attempted to tackle the problem using three approaches – one that uses simple heuristics, another based on an asymmetric word similarity metric based on finding the shortest path through a matrix of elements representing phonological similarity, and another utilizing a deep learning model partially trained on synthetic data. Despite a high degree of irregularity in Czech compounding, the *Czech Compound Splitter* tool achieved an accuracy of 54% in the task of compound splitting and an accuracy of 92% in compound identification. The work on this study has contributed to the creation of DeriNet version 2.1 by manually identifying the parents of 1,439 compounds, totalling 2,691 compounds with identified parents.

Acknowledgments

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101), and by the Grant No. START/HUM/010 of Grant schemes at Charles University (Reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935).

References

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. The CELEX lexical database (CD-ROM). *Linguistic Data Consortium* .

Laurie Bauer. 2003. *Introducing linguistic morphology*. Edinburgh University Press.

Antonietta Bisetto and Sergio Scalise. 2005. The classification of compounds. *Lingue e linguaggio* 4(2):319–332.

Ivana Bozděchová. 1997. *Tvoření slov skládáním*. Institut sociálních vztahů.

Elizaveta L. Clouet and Béatrice Daille. 2014. Splitting of compound terms in non-prototypical compounding languages. In *Workshop on Computational Approaches to Compound Analysis*. pages 11–19.

Jan Hajič, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, and Barbora Štěpánková. 2020. MorfFlex CZ 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3186.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a lexical-semantic net for German. In *Automatic information extraction and building of lexical semantic resources for NLP applications*. pages 9–15.

Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics* 4(2):100–107. https://doi.org/10.1109/TSSC.1968.300136.

Oliver Hellwig and Sebastian Nehrdich. 2018. Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pages 2754–2763.

Verena Henrich and Erhard Hinrichs. 2011. Determining immediate constituents of compounds in GermaNet. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing 2011*. pages 420–426.

Gérard Huet. 2005. A functional toolkit for morphological and phonological processing, application to a Sanskrit tagger. *Journal of Functional Programming* 15(4):573–614.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*. pages 116–121. https://arxiv.org/abs/1804.00344.

Irina Krotova, Sergey Aksenov, and Ekaterina Artemova. 2020. A joint approach to compound splitting and idiomatic compound detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. pages 4410–4417. https://aclanthology.org/2020.lrec-1.543.

Vladimir I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics-Doklady* 10(8):707–710.

Jianqiang Ma, Verena Henrich, and Erhard Hinrichs. 2016. Letter sequence labeling for compound splitting. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. pages 76–81.

Martin Ološtiak and Marta Vojteková. 2021. Kompozitnosť a kompozícia: príspevok k charakteristike zložených slov na materiáli západoslovanskỳch jazykov. *Slovo a slovesnost* 82(2):95–117.

Karel Pala and Pavel Šmerk. 2015. Derivancze—derivational analyzer of Czech. In *International Conference on Text, Speech, and Dialogue*. pages 515–523.

Pavel Štichauer. 2013. Je možná nová klasifikace českỳch kompozit? *Časopis pro moderní filologii* 95(2):113–128.

Kyoko Sugisaki and Don Tuggener. 2018. German compound splitting using the compound productivity of morphemes. In *14th Conference on Natural Language Processing*. pages 141–147.

Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, and Šárka Dohnalová. 2019. DeriNet 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2995.

Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, Šárka Dohnalová, Emil Svoboda, and Jan Bodnár. 2021. DeriNet 2.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3765.

Transferring Word-Formation Networks Between Languages

Jonáš Vidra
Charles University,
Faculty of Mathematics and Physics,
vidra@ufal.mff.cuni.cz

Zdeněk Žabokrtský
Charles University,
Faculty of Mathematics and Physics,
zabokrtsky@ufal.mff.cuni.cz

Abstract

In this article, we present a proof-of-concept method for creating word-formation networks by transferring information from another language. The proposed algorithm utilizes an existing word-formation network and parallel texts and creates a low-precision and moderate-recall network in a language, for which no manual annotations need to be available. We then extend the coverage of the resulting network by using it to train a machine-learning method and applying the resulting model to a larger lexicon, obtaining a moderate-precision and high-recall result. The approach is evaluated on French, German and Czech against existing word-formation networks in those languages.

1 Introduction

A word-formation network is a dataset capturing information about how are lexemes created using derivation, compounding, conversion and other types of relations. Such networks can be created using various degrees of automatization. On one end of the spectrum are networks created by manually annotating the individual relations, resulting in a dataset that is highly precise, but either expensive to create or small in size.

In this article, we explore a method from the other, unsupervised, part of the scale: a method which does not require any human input or in-language annotations of word-formation relations. Instead, it transfers an existing word-formation network from another language using parallel texts and off-the-shelf tools for tokenization and lemmatization. Parallel texts are significantly more abundant and easier to obtain than word-formation annotations and they are available for more languages – compare the OPUS collection (Tiedemann, 2012), where just the OpenSubtitles corpus is available for 65 languages, to a survey of available word-formation networks listing only 63 resources for 22 languages (Kyjánek, 2018).

As a result, our method should allow for a cheap and rapid creation of word-formation networks for many languages, although at a cost of lower quality. We hope that it is possible to emulate the successes of transfer learning methods used for other similar tasks in natural language processing, such as syntactic parsing (McDonald et al., 2011), part-of-speech tagging (Zhang et al., 2016) or lemmatization (Rosa and Žabokrtský, 2019).

The main idea behind our methods is that translation of text between languages is supposed to preserve the pragmatic meaning of texts and it usually preserves also the semantic meaning of individual sentences and words. Since word-formational relations connect words with similar semantics and orthography, multiple possible target-language translations of a single source-language word are word-formationally related with a higher probability than randomly selected words. Moreover, many types of word-formational relations have parallels across languages. For example, actor nouns are typically derived from verbs – and if we take two such nouns from two languages, which are translations of one another, chances are that their predecessor verbs will also be translation equivalents (e.g. the Czech and English relations *opravit* ("to repair") → *opravář* ("repairman") are parallel, even though one uses derivation and the other one compounding). Therefore, we believe that some information about word-formation relations can be shared across languages.

By further filtering the transferred relations by orthographic distance, we obtain a moderate-precision and low-recall word-formation network. The recall can be improved by extracting the discovered string-wise word-formation patterns using a statistical machine-learning method and finding more examples of them across the lexicon.

The pilot experiments presented in this paper focus on one-to-one relations between lexemes. We omit compounding altogether and simplify the task of creating a word-formation network to a task of assigning each lexeme a single *parent* lexeme, or deciding that it is unmotivated and should function as a root of the morphological family.

Moreover, although we aim to produce algorithms and models which would be able to create word-formation networks for any language with mostly concatenative morphology and written in an alphabetic script, we currently focus on French, German and Czech, because these are among the few languages for which a large, high-quality word-formation network already exists. The existing networks, Démonette (Hathout and Namer, 2014), DErivBase (Zeller et al., 2013) and DeriNet (Žabokrtský et al., 2016), serve a dual role as data for transfer on the source side, and evaluation datasets on the target side of each of the six possible independent translation pairs.

2 Related work

Several unsupervised methods of creating word-formation networks have been proposed before. Baranes and Sagot (2014) created a method that infers derivational relations from inflectional paradigms and reported a very high precision (80-98% depending on the language). The relations are detected by first extracting a list of possible prefixal and suffixal changes and then pattern-matching pairs of words against it. The inflectional paradigms are used for reducing problems with suppletion and allomorphy within stems, which would otherwise cause the prefix- and suffix pattern matching to fail – e.g. if we know that *worse* is a comparative form of the lemma *bad*, we can link the lexeme *worsen* to *bad* using the rule *X-e* → *X-en*.

A different solution to the problem of allomorphy is proposed by Lango et al. (2021), who use a pattern-mining method to detect rules of allomorphy jointly with affixation. The patterns are extracted automatically in an unsupervised fashion and the potential relations are ranked by a machine-learning model trained on a small manually annotated word-formation network.

Batsuren et al. (2019) deal with cognate detection (i.e. linking words of common origin, identical meaning and similar spelling in different languages) using a multilingual approach. The multilingual data they use is a specialized linguistic resource containing information about etymological ancestry, which means that their methods are not directly applicable in our semi-supervised setting.

Cognates can also be used as a clue for aligning parallel corpora and several methods for detecting cognate pairs were developed with the alignment task in mind, but these methods need not be very precise – e.g. Church (1993) uses identical character 4-grams and Simard et al. (1992) use pairs of words with identical first four characters; both methods are too imprecise to recognize exact word-formational relations.

A method utilizing cosine distance between neural-network word embeddings was used by Üstün and Can (2016) to construct an implicit word-formation network as an intermediate step in morphological segmentation. Word embeddings are also used by Musil et al. (2019), who show that words created through similar word-formation processes have similar embedding differences; however, they do not use these results to actually construct a network out of word-embedding data.

3 Transfer algorithm

To transfer a word-formation network from a source to a target language, we view the network as a list of parent-child derivational relations and attempt to find the best parent for each target-side lexeme using a word-translation model together with target-side formal similarity metrics. Conceptually, the source lexeme C is first backtranslated into the source language as C', a suitable parent P' of the translation is found in the source word-formation network and this parent is translated into the target language as P.

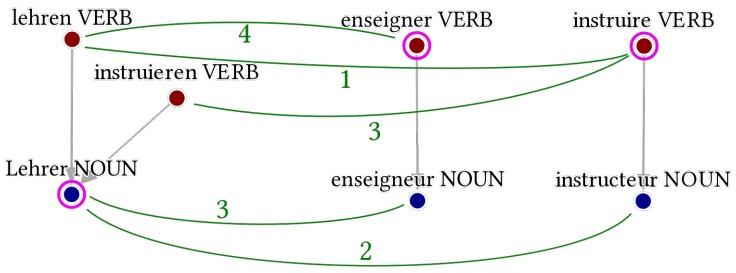


Figure 1: An example of finding a parent for the German lexeme *Lehrer* ("teacher") by transferring information from a French word-formation network, with word-formation relations in grey and alignments in green. *Lehrer* is aligned to *enseigneur* ⅗ times, which has *enseigner* available through 1 relation, to which *lehren* is aligned ¼ times. *Lehrer* is also aligned to *instructeur* ⅖ times, which has *instruire* available through 1 relation, to which *lehren* is aligned ¼ times and *instruieren* ¾ times. The translation score of *lehren* → *Lehrer*, calculated according to Equation 1 below, is therefore $\frac{3}{5} \cdot \frac{1}{2} \cdot \frac{4}{4} + \frac{2}{5} \cdot \frac{1}{2} \cdot \frac{1}{4} = 0.35$ while the score of *instruieren* → *Lehrer* is $\frac{2}{5} \cdot \frac{1}{2} \cdot \frac{3}{4} = 0.15$. The relative edit distance is ⅔ for *lehren* → *Lehrer*, and ⁸⁄₁₁ for *instruieren* → *Lehrer*. Therefore, the final score of *lehren* → *Lehrer* is $\frac{0.35+5\cdot(1-2/6)}{6} = 0.336$ and the score of *instruieren* → *Lehrer* is $\frac{0.15+5\cdot(1-8/11)}{6} = 0.252$.

The translations and backtranslations are found using a probabilistic word translation lexicon induced from word-aligned data obtained by running FastAlign (Dyer et al., 2013) on a lemmatized parallel corpus. Since the present article does not consider compounding, univerbation or other word-formation relations connecting more than two lexemes, we count each pair of aligned lexemes separately, regardless of whether one of the lexemes has other alignments in that parallel sentence pair. As a result, a lexeme aligned to a multi-word phrase is considered to be equally translated from each member lexeme of that phrase.

Since there may be multiple possible translations of each lexeme, and because the most suitable parent needn't be the direct parent of C', but rather another member of its word-formational family (e.g. the Czech lexemes *svoboda* ("freedom") → *svobodný* ("free") have the opposite derivational relation from English or German *frei* → *die Freiheit*), the process is conducted probabilistically, yielding many potential parents P for each C, each with a score. The target network is then found by finding the spanning tree of this graph of relations which maximizes the product of the scores (Chu and Liu, 1965).

The score of each potential relation is obtained as a weighted arithmetic mean of one minus the relative edit distance between C and P and their translation score. The relative edit distance is the Levenshtein distance between the lemmas of C and P divided by the maximum of their lengths, yielding a number between 0 and 1.

We define the translation score of C and P as $\mathrm{Xfer}(C, P)$ according to Equation 1 below, where $|\mathrm{align}(x, y)|$ denotes the number of alignments between lexemes x and y seen in the aligned data and $\mathrm{dist}(C', P')$ denotes the number of relations on the shortest path from C' to P' in the source network.

$$\mathrm{Xfer}(C, P) = \sum_{\forall C', P'} \frac{|\mathrm{align}(C, C')|}{\sum_{\forall x} |\mathrm{align}(C, x)|} \cdot 0.5^{\mathrm{dist}(C', P')} \cdot \frac{|\mathrm{align}(P', P)|}{\sum_{\forall x} |\mathrm{align}(P', x)|} \quad (1)$$

Therefore, the translation score is the product of the conditional probability of obtaining the backtranslated lexeme C' given the lexeme C and the conditional probability of obtaining the translated parent lexeme P given P', halved for each relation that has to be traversed between C' and P'. If there are multiple possible choices of C' and P' for the given C and P, their translation scores are summed.

To prevent relations with low scores from being selected in the case where there are no better candidates, a relation is only considered for inclusion if its score is higher than a threshold.

An illustration of the translation score calculation is given in Figure 1.

The transfer algorithm is parametrized by the weights used for calculating the weighted mean of the translation and edit distance scores, and by the threshold. Since we intend to use the transfer algorithm in

an unsupervised setting, it is necessary to obtain the weights without training them using e.g. grid search or gradient descent on in-language annotations. We have, however, found that although the algorithm is moderately sensitive to the setting of the weights and the threshold, the optimal settings in all tested languages are nearly identical. This allows us to train the hyperparameters on one language pair in a supervised manner and use them on other pairs without further training. Therefore, we set the weight of the edit distance to 5, the weight of the translation to 1 and the threshold to 0.8 using grid search on the Czech → German transfer pair and use these hyperparameters on all pairs.

4 Expansion through machine learning

The word-formation network obtained via cross-lingual transfer covers only lexemes with alignments, i.e. high-frequency ones. Therefore, it is desirable to increase coverage of lower-frequency parts of the lexicon and lexemes not seen in the parallel data. We perform this by extracting affixal patterns from the transferred network and applying them across the data.

The affixal pattern of a (proposed) word-formational relation is an unsupervised approximation of the morpheme difference between the related lexemes. We obtain it as the leftover substrings to the left and right of the longest common contiguous substring shared by lowercased lemmas of the lexemes. For example, the relation *Kampf* ("a fight") → *kämpfen* ("to fight") has the longest common contiguous substring *mpf* and affixal pattern *ka-* → *kä-* + *-en*.

We use the transferred network as a seed to train a machine learning method to predict derivational relations by classifying pairs of lexemes as either directly derived or non-derived from one another. The output network is obtained by finding the maximum spanning tree of the graph of predictions (Chu and Liu, 1965). The features used for classification are the one-hot-encoded part-of-speech categories of both lexemes, their edit distance, the difference of their lengths, whether each of them starts with a capital letter and the frequency of their affixal pattern as seen in the training dataset.

Since classifying all pairs of lexemes found in the dataset is too computationally expensive, we only sample pairs of lexemes that are near one another when the dataset is lexicographically sorted by lemma, in both prograde and retrograde fashions. The prograde-sorted list puts lemmas with common beginnings near each other, meaning that pairs of words differing only in short suffixes will be selected for classification. The retrograde-sorted one does the same with lemmas differing only in a short prefix.

We perform the lexicographic sorting on uppercased lemmas stripped of accent marks so that e.g. the German word *Wunsch* ("a wish") sorts close to *wünschen* ("to wish") despite the differences in case and the presence or absence of the umlaut.

This method of obtaining relation candidates depends on the linguistic properties of the languages under consideration, namely Czech, French and German. All three derive words predominantly by affixation, with limited allomorphy in the stem and only rare examples of circumfixation, apophony or suppletive relations, which this method generally doesn't detect as possible relations. Therefore, looking at a window of ±5 lexemes catches 85 % of all possible derivational relations in DErivBase and ±10 catches 90 %. On Démonette, 96 % of derivations are within ±5 and 98 % are within a ±10 window. In DeriNet, a window of ±5 contains 85 % of all relations and ±10 contains 90 %. The method would perform poorly on languages with more frequent circumfixation or nonconcatenative morphology, such as transfixation or templatic morphology found in e.g. Hebrew or Arabic.

A possible systematic fix for detecting words derived by circumfixation would be to use a more complex measure of morphological similarity. A method we tried is the orthographic part of the model from Proxinette (Hathout, 2008), which approximates morphological relatedness by counting common n-grams of varying length, probabilistically weighting them by rarity in the corpus. Its construction allows enumerating lexemes most similar to an input lexeme in a computationally-tractable way, without considering all pairs. However, it produces inferior results on the three datasets we use, we therefore don't use it in our experiments.

We evaluated multiple classification methods implemented in the scikit-learn package (Pedregosa et al., 2011), namely SVC, LogisticRegression, AdaBoostClassifier, KNeighborsClassifier, DecisionTreeClassifier, BernoulliNB and Perceptron and selected logistic regression for consistent evaluation performance.

```
for gold_child in gold.lexemes:
  if not gold_child.parent:
    true_negative++
  else:
    for t_child in translations(gold_child):
      for t_parent in family(t_child):
        for parent in backtranslations(t_parent, gold_child):
          if parent = gold_child.parent:
            true_positive++
            continue_line 1
    false_negative++
  accuracy := ((true_positive + true_negative) / (true_positive +
   ↪ true_negative + false_negative))
  recall := true_positive / (true_positive + false_negative)
```

Listing 1: Pseudocode for calculating oracle accuracy and recall of the transfer algorithm. The backtranslation function returns all backtranslations of `t_parent`, except those that translate to `gold_child`.

5 Evaluation Method

We evaluate the performance of our systems by measuring precision, recall and accuracy in the task of assigning a parent to a lexeme. We define precision as the ratio of correctly predicted relations to all predicted relations, recall as the ratio of correctly predicted relations to all gold relations and accuracy as the ratio of correctly assigned parents or correctly recognized unmotivated lexemes to all gold lexemes. Therefore, the precision and recall don't take into account unmotivated lexemes, while the accuracy does. The gold-standard data is taken from the existing word-formation network for the target language.

Because the set of lexemes captured in the transferred network differs from the one used in the gold-standard data, we calculate the metrics in two ways, which differ in their treatment of missing lexemes. "External" measures consider all gold-standard relations of lexemes missing from the evaluated network to be false negatives, while the "internal" measures ignore them instead measures and only measure scores on the intersection of the two lexicons. Precision is the same for both methods, but recall and accuracy differ. The baseline measures and the networks obtained by machine learning are created from the set of lexemes found in the gold-standard network, which makes the internal and external measures identical.

5.1 Baselines

To establish a lower bound of reasonably achievable scores, we created two baselines: one trivial, called "empty", and one inspired by the purely left- or right-branching parse, the standard baseline in syntactic parsing, called "closest-shorter".

The empty baseline for a given lexicon is calculated as the scores of an empty word-formation network created over that lexicon, i.e. a network without any relations. The lexemes from gold-standard data which have no assigned parent are therefore evaluated as correct, while all lexemes with parents are incorrect, resulting in unmeasurable (zero) precision, zero recall and moderate-to-high accuracy.

The closest-shorter baseline gives each lexeme four options for its parent and selects the one which has a shorter lemma and the closest orthographic distance, as measured by the ratio of the length of the longest common contiguous substring to the sum of lengths of the two lemmas. The options to choose from are the previous and next lexemes in prograde sorting of the lexicon, and the previous and next lexemes in retrograde sorting. The lemma length criterion means that lexemes surrounded by longer neighbors in both prograde and retrograde sorting of the lexicon remain unmotivated. We have already observed that both ends of most derivational relations lie within a small window on a sorted lexicon, making this baseline rather strong in terms of both precision and recall.

Lang pair	Sentences	Tokens on left	Tokens on right
de — cs	15 237 340	48 320 109	45 922 280
fr — cs	25 838 124	83 108 504	87 983 667
fr — de	14 779 572	44 135 610	48 440 995

Table 1: Sizes of parallel data for each language pair after part-of-speech category filtering.

5.2 Oracle Score

As an additional measure of the potential quality of the transfer approach, we measured the oracle score of obtaining the gold-standard parent through any combination of back- and forward-translations of gold-standard child lexemes. Under this measure, unmotivated lexemes are always considered to be correct, and a derived lexeme is considered to be correctly connected to its parent if it can be backtranslated to a member of a word-formational family, which contains a member that can be translated to the correct parent. The pseudocode of this algorithm is present in Listing 1. The recall and accuracy obtained using this algorithm represent the maximum scores achievable with the transfer method, if it selected the gold parent for each lexeme every time it is available.

Any error in the recall can be broken down into three categories: first, where we cannot translate the child to the language of the transferring network; (no `t_child` on line 5 of Listing 1); second, where there are no translations of any members of the translated lexeme's family (no `parent` on line 7) and third, where no possible parent matches the gold one (predicate on line 8 is always false).

5.3 Experimental setting

For the purposes of this paper, we conducted experiments on Czech, French and German, which are all languages with existing word-formation networks suitable for transfer – DeriNet 2.0 (Žabokrtský et al., 2016) with 809 282 relations, Démonette 1.2 (Hathout and Namer, 2014) with 13 808 relations and DErivBase 2.0 (Zeller et al., 2013) with 43 368 relations, respectively. For ease of use, we used their versions available in the UDer 1.0 collection (Kyjánek et al., 2019), which have been converted to a common format at a slight loss of information. We transferred each network into both other languages and compared the result to the existing network for that language.

The transfer was realized using word dictionaries obtained from word alignments of parallel data. We used the OpenSubtitles dataset from the OPUS collection (Tiedemann, 2012) for all language pairs, lemmatizing them with UDPipe 1.2 (Straka and Straková, 2017) and extracting only words tagged as adjectives, adverbs, nouns and verbs. The lemmatizer uses pretrained models trained on treebanks from Universal Dependencies (Nivre et al., 2016). The lemmatized corpora are then aligned using FastAlign (Dyer et al., 2013). The data sizes are listed in Table 1.

6 Evaluation Results

As can be seen in Table 2, the networks created by the transfer algorithm are rather small in size. Within the constructed network, precision and recall are moderate for most language pairs, but when compared to the gold standard data, recall is nearly zero for all of them.

The performance of the transfer method depends a lot on the size of the transferred network. Since the Czech DeriNet is an order of magnitude larger than the other networks, the gold scores for networks created by using it as a base are the highest ones, but even these don't match more than 2.5% of relations from the gold-standard data.

The precision of the constructed networks is also influenced by the translation quality. The alignment data trained on the de—fr pair (in both directions) has many incorrect alignments. This doesn't affect the oracle score, since the correct translations will generally be found, but the wide distribution of the probability mass hurts the actual algorithm, which is unable to distinguish plausible and implausible translations.

The machine learning method provides a way of generalizing the output of the transfer method, as it

		Size		Internal scores [%]				Gold scores [%]		
Alg	Lang pair	Lex	Rel	Prec.	Recall	F1	Acc.	Recall	F1	Acc.
Xfer	de → cs	18 118	5 971	39.66	33.11	36.09	53.71	0.29	0.58	1.19
	fr → cs	20 225	7 045	42.46	36.11	39.03	53.79	0.37	0.73	1.33
	cs → de	13 803	3 847	27.06	35.36	30.66	65.88	2.45	4.50	17.07
	fr → de	2 938	600	14.33	14.14	14.24	64.74	0.20	0.39	4.19
	cs → fr	2 769	1 219	23.54	30.50	26.57	42.72	2.10	3.86	7.65
	de → fr	439	144	3.47	11.36	5.32	59.45	0.04	0.07	1.84
ML	de → cs	1 026 036	743 469	45.70	73.81	56.45	48.90	73.81	56.45	48.90
	fr → cs	1 026 036	742 784	39.60	70.00	50.58	43.99	70.00	50.58	43.99
	cs → de	280 454	68 154	35.02	67.73	46.17	80.15	67.73	46.17	80.15
	fr → de	280 454	34 809	44.25	39.35	41.66	84.62	39.35	41.66	84.62
	cs → fr	21 288	15 136	60.33	88.64	71.79	66.30	88.64	71.79	66.30
	de → fr	21 288	4 700	35.57	13.79	19.88	36.69	13.79	19.88	36.69
closest-shorter baseline	cs	1 026 036	808 933	21.03	53.54	30.20	23.35	53.54	30.20	23.35
	de	280 454	225 092	5.22	56.51	9.55	20.70	56.51	9.55	20.70
	fr	21 288	17 451	31.65	82.71	45.79	38.55	82.71	45.79	38.55
empty baseline	cs	1 026 036	0	N/A	0.00	0.00	21.14	0.00	0.00	21.14
	de	280 454	0	N/A	0.00	0.00	84.62	0.00	0.00	84.62
	fr	21 288	0	N/A	0.00	0.00	35.15	0.00	0.00	35.15

Table 2: Evaluation scores of the results and baselines for each language pair. Internal scores are measured on the set of lexemes in the generated network, gold scores on the set of lexemes from gold data. Precision is identical for both. For the machine learning and baseline algorithms, the distinction between internal and gold scores does not matter, since the lexicon used for prediction is taken from the gold-standard data as is.

	Scores [%]		Error cause [%]			WFN rel count	
Lang pair	Recall	Accuracy	No child trans	No parent trans	No match	Xferred	Gold
de → cs	5.10	29.14	91.05	0.08	3.77	43 368	809 282
fr → cs	6.75	31.74	89.62	0.05	3.59	13 808	809 282
cs → de	34.47	89.82	52.08	0.23	13.22	809 282	43 368
fr → de	26.24	92.69	50.60	0.02	22.14	13 808	43 368
cs → fr	34.67	80.11	56.81	0.20	8.33	809 282	13 808
de → fr	22.26	64.01	61.89	0.07	15.78	43 368	13 808

Table 3: Transfer oracle scores for each language pair. Precision is 100% in all cases. The error causes list percentage of cases where the lexeme cannot be translated to the language of the transferring network, where no possible parents can be translated back, and when none of the translated parents match the gold one, respectively. The error percentage points are relative to the total relation count, i.e. they sum up to 100 together with recall. The last two columns list sizes of the transferred and gold-standard word-formation networks, measured in relations.

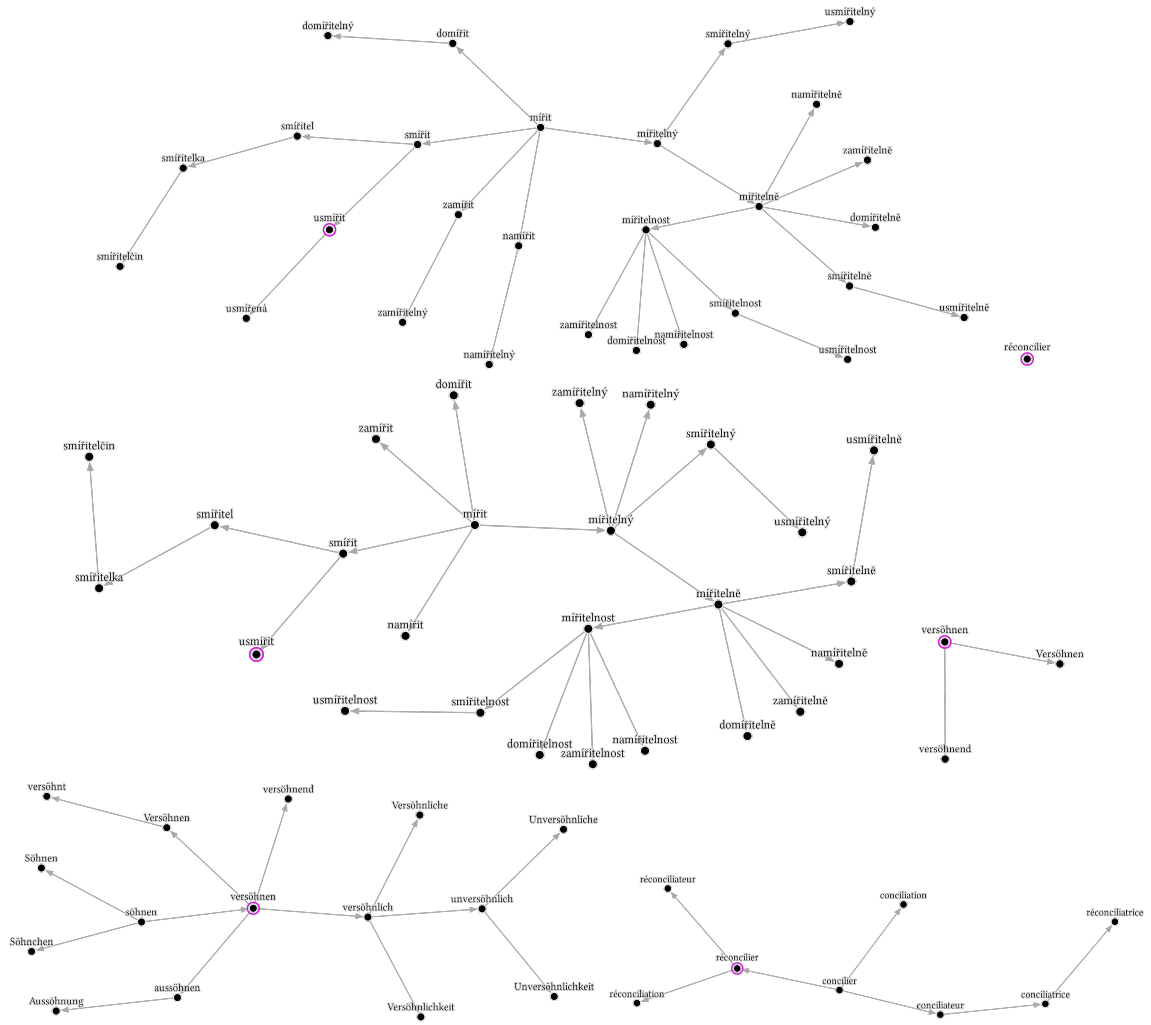


Figure 2: Word-formation networks generated by the machine learning expansion of the transferred networks, showing the family of lexeme *to reconcile* circled in violet for each of the six language pairs. Clockwise from top left: de-cs, de-fr (single lexeme), fr-de, cs-fr, cs-de, fr-cs.

learns frequent affixal patterns from the transferred data and applies them to a larger lexicon, omitting infrequent (often spurious) patterns. As seen in the second part of Table 2, this results in increased precision on the networks transferred to French and German, where the gold standard data consists of relatively few selected paradigms and therefore skews towards fewer, more productive patterns. The results on the Czech data, which is more varied, still reach precision comparable to the transferred networks we train on. Recall increases in all cases, even when compared to the "internal" scores, which are more favorable to the transferred networks. Due to this large increase, F1-score also increases. Sample outputs of the machine learning method can be seen in Figure 2.

The oracle scores are in Table 3. The scores are influenced by the ratio of sizes of the word-formation networks used for transfer and evaluation; transferring a large network and evaluating on a smaller one gives an advantage in recall in comparison to the opposite scenario, simply because a larger source network offers more options to select from after transfer. The error causes listed in the table correspond to the sources of error in recall as categorized in Section 5.2.

For all language pairs, most of the errors are attributable to the first cause, where the gold data contains untranslatable lexemes. For the pairs that translate to Czech, this is again explainable by the size and composition of its DeriNet network, which contains many unattested lexemes – finding rare lexemes such as *přeskočitelnost* ("skippability") in the parallel data is unlikely.

Additionally, transfers of networks to German have higher accuracy than transfers to French, even

though the recall is comparable. This is because the German network, DErivBase, contains many compounds, which don't have their parents annotated and are listed as unmotivated. These are counted in the accuracy scores (the definition of oracle score above considers missing relations to be always correctly recognized) but do not contribute to recall of relations. The unmotivated words are also the reason behind the fact that the fr-de pair has higher accuracy than cs-de, despite having lower recall – fewer relations are translated, resulting in more unmotivated words being correct.

The oracle scores show that the main bottleneck is the word translation dictionary – the "No child trans" category accounts for 50-90% of all errors. This is also the reason why the networks obtained through the machine learning expansion have better scores than the oracle of the transfer algorithm. The transfer lexicon is limited to the lexemes found in the parallel data, whose source-side alignments are found in the source word-formation network, and for evaluation purposes, we further limit the lexicon to lexemes from the gold-standard data. The machine-learning pipeline uses the gold-standard lexicon directly, eliminating the "No child trans" class of errors entirely.

7 Conclusion

In this paper, we presented a cross-lingual method for creating word-formation networks by transferring an existing network using a word-translation lexicon induced from word alignments. The transferred small networks are then expanded by extracting paradigms using statistical machine learning and applying them to a larger set of lexemes. The resulting word-formation networks show moderately high precision and good recall on six language pairs.

Acknowledgments

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, by the Charles University Grant Agency (project No. 1176219) and by the SVV project number 260 575. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIAH CZ project (LM2015071, LM2018101).

The OpenSubtitles corpus was kindly provided by `http://www.opensubtitles.org/`.

References

Marion Baranes and Benoît Sagot. 2014. A language-independent approach to extracting derivational relations from an inflectional lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, pages 2793–2799. http://www.lrec-conf.org/proceedings/lrec2014/pdf/379_Paper.pdf.

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 3136–3145. https://doi.org/10.18653/v1/P19-1302.

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. *Science Sinica* 14:1396–1400.

Kenneth Ward Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *31st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Columbus, Ohio, USA, pages 1–8. https://doi.org/10.3115/981574.981575.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 644–648. https://www.aclweb.org/anthology/N13-1073.

Nabil Hathout. 2008. Acquisition of morphological families and derivational series from a machine readable dictionary. In *Proceedings of the 6th Décembrettes.*. Cascadilla, Bordeaux, France, Cascadilla Proceedings Project, pages 166–180. https://hal.archives-ouvertes.fr/hal-00382808.

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11:125–162.

Lukáš Kyjánek. 2018. Morphological resources of derivational word-formation relations. Technical Report ÚFAL TR-2018-61, ÚFAL MFF UK, Praha, Czechia. http://ufal.mff.cuni.cz/techrep/tr61.pdf.

Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra. 2019. Universal Derivations kickoff: A collection of harmonized derivational resources for eleven languages. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*. ÚFAL MFF UK, Praha, Czechia, pages 101–110.

Mateusz Lango, Zdeněk Žabokrtský, and Magda Ševčíková. 2021. Semi-automatic construction of word-formation networks. *Language Resources and Evaluation* 55(1):3–32.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 62–72. https://aclanthology.org/D11-1006.

Tomáš Musil, Jonáš Vidra, and David Mareček. 2019. Derivational morphological relations in word embeddings. In *The BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP at ACL 2019*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 173–180.

Joakim Nivre et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. ELRA, pages 1659–1666.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Rudolf Rosa and Zdeněk Žabokrtský. 2019. Unsupervised lemmatization as embeddings-based word clustering.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*. Montréal, Canada. https://aclanthology.org/1992.tmi-1.7.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Vancouver, Canada, pages 88–99. http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Ahmet Üstün and Burcu Can. 2016. Unsupervised morphological segmentation using neural word embeddings. In Pavel Král and Carlos Martín-Vide, editors, *Statistical Language and Speech Processing: 4th International Conference, SLSP 2016*. Springer International Publishing, Cham, Switzerland, pages 43–53. https://doi.org/10.1007/978-3-319-45925-7_4.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 1201–1211. http://www.aclweb.org/anthology/P13-1118.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 1307–1317. https://doi.org/10.18653/v1/N16-1156.

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra, and Adéla Limburská. 2016. Merging data resources for inflectional and derivational morphology in Czech. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association, Paris, France, pages 1307–1314.

Compound Splitting and Analysis for Russian

Daniil Vodolazsky
Sber; Higher School of Economics
Moscow, Russia
daniil.vodolazsky@mail.ru

Hermann Petrov
Sber
Moscow, Russia
gerpetrum@yandex.ru

Abstract

This paper presents a method for compound identification, splitting, and analysis for Russian. First, we identify whether a word is a compound with a neural network. Then we apply a rule-based approach to generate a set of linguistically possible hypotheses with analyses, including the normalization of the compound components. Finally, we score and rank the hypotheses using three techniques: word frequencies, word embeddings, neural networks. We evaluate models on a manually collected and annotated test dataset. We make the dataset and code publicly available.

1 Introduction

Due to the productivity of most languages, it is possible to generate an infinite number of different words, and the speakers can freely form and analyze them even if they have not seen them before. However, this becomes a challenge for computational models in various NLP tasks. Word-level models suffer from a lack of understanding of an internal word structure and cannot process unseen tokens.

The most common approach nowadays is the byte-pair encoding (BPE) (Sennrich et al., 2015), but its segmentation strategy significantly depends on a training corpora domain. All these methods are data-driven and language-independent. It has been shown (Sennrich and Haddow, 2015; Li et al., 2018; Hofmann et al., 2021) that prior knowledge about the language improves the models' performance.

The goal of this work is to propose a linguistically grounded approach to subword segmentation for Russian compounds.

A compound is a lexeme that consists of two or more stems. Compounding is a word-formation process that creates such lexemes. It can be highly productive in synthetic languages such as German or Russian. Many compound splitting and analyzing tools exist for German (Koehn and Knight, 2003; Schmid et al., 2004; Henrich and Hinrichs, 2011; Weller-Di Marco, 2017), and, to the best of our knowledge, no for Russian. According to Gromenko (2020), 32% of the neologisms in Russian are produced with compounding which makes it important to have a tool for their analysis.

This task is complicated for several reasons. Although the number of productive compounding patterns in Russian is relatively small (compared to the derivational ones), there is structural ambiguity in many cases. For instance, the adjective железнодорожный '*zhel'eznodorozhnyy*' (*railway-related*) matches the following three rules:

- **rule761**([adj + ITFX] + noun + н1(ый) → adj): железный(-ая), дорога '*zhel'eznyy(-aya), doroga*' (*railway, lit. iron road*);
- **rule754**([adj + ITFX] + adj → adj): железный, дорожный '*zhel'eznyy, dorozhnyy*';
- **rule776**(adv + adj → adj): железно, дорожный '*zhel'ezno, dorozhnyy*'.

The correct choice depends on the semantics of the source and produced words. As shown in the analyses above, the endings of the left items are removed and replaced with the *-o-*/*-e-* interfix[1]; the stem of the word

[1]The choice of a vowel depends on a previous consonant, e. g. *-e-* is used after palatalized consonants.

дорога '*doroga*' (*road*) is changed with a velar—sibilant alternation before the derivational suffix -н1(ый) '*-n1(yy)*'. A morphemic segmentation in all three analysis would be the same: *zhel'ez|n|o|dorozh|n|yy*— and the structural information would be lost. Therefore it is desirable to provide not just the splitting but also the analysis that includes the lemmas of the source words and (optionally) the rule ID.

In this paper, we propose an algorithm for splitting and analysis of Russian compounds. Our method is a combination of rules, finite state machines, and deep learning. We manually collected a set of 1.7K ground-truth compound analyses and used it to compare the performance of the models. We make our code and data freely available on GitHub[2].

2 Compounding in Russian

In this section, we overview the principal categories of Russian compounds. For each of them, we provide a brief description and give examples. Note that our categorization is not Russian-specific and can be applied to other languages, such as Polish (Szymanek, 2017).

Compounding is a word-formation process of concatenating two or more stems. It is often accompanied by morphophonological alternations in a place of the concatenation of stems. The results of compounding are **compounds**. In this paper, to avoid terminological ambiguity, we call them **pure compounds**. Examples of pure compounds are *doghouse*, *blackwood*, *redhead*, in Russian: северо-запад '*severo-zapad*' (*northwest*), нефте|промышленность '*nefte|promyshlennost''* (*oil industry*), цельно|металлический '*tsel'no|metallicheskiy*' (*full metal*).

There are other compounds, spelled as two words or hyphenated: рыба '*ryba*' (*fish*) + меч '*mech*' (*sword*) → рыба-меч '*ryba-mech*' (*swordfish*). This sort is getting more common because of the influx of compounds loaned (and sometimes calqued) from English. However, in this work, we classify them as pure compounds because of their structural similarity: $w_1 + w_2$.

Parasynthetic compounds are produced with both compounding and derivation with one or many affixes. The most common pattern for such words is $w_1 + w_2 + sfx$. For example, the word зелено|глазый '*zeleno|glazyy*' (*green-eyed*) is a parasynthetic compound since the lexemes *зеленоглаз(а) '*zelenoglaz(a)*' and *глазый '*glazyi*' do not exist in the language.

According to Bisetto and Melloni (2008), words are analyzed as parasynthetic compounds in the following cases:

1. either $w_1 + w_2$ and $w_2 + sfx$ are non-existent lexemes;
2. $w_1 + w_2$ is not attested and $w_2 + sfx$ is instead an independently occurring lexeme. 'However, scopal ambiguity effects systematically arise in this case, since the suffix has scope over the complex base, rather than just over the second stem, giving rise to a bracketing paradox', they say.

In contrast, 'Russkaya Grammatika'[3] (Shvedova, 1980) follows a formal criterion: if the lexeme $w_2 + sfx$ exist, then a compound is pure, not parasynthetic. Thus, for example, the word древне|египетский '*drevne|egipetskiy*' (*Ancient Egyptian*) is parasynthetic by the former definition and pure by the latter.

In this work, we follow the convention from 'Russkaya Grammatika', because it provides not only a definition but a significant number of examples for each compounding rule.

A frequent pattern in Russian and other Slavic languages is (prep + noun + sfx → adj) (cf. за '*za*' (*beyond*) + граница '*granitsa*' (*border*) + -н1(ый) '*-n1(yy)*' → за|границ|ный '*za|granich|nyy*' (*overseas*) (ru), *za|hranič|ní* (cz), *za|granicz|ny* (pl)). According to the Russian linguistic tradition, the first morpheme in such words is analyzed as a prefix, hence the morphological process is derivation, not compounding.

Classical and neoclassical compounds are composed from classical Latin or ancient Greek roots: *hydrogen*, *biology*, *democracy*, *homophobia*. Such words appear in all areas of science and technology. Moreover, classical compounding is the main source of new words in these domains (Gromenko, 2020).

Some of such roots and other international morphemes (e. g. авто- '*avto-*' (*auto-*), аква- '*akva-*' (*aqua-*), -фобия '*-fobiya*' (*-phobia*), мини- '*mini-*' (*mini-*), etc.) can attach to normal Russian words

[2] https://github.com/s231644/rucompoundsplitter
[3] http://rusgram.narod.ru/760-790.html, § 760.

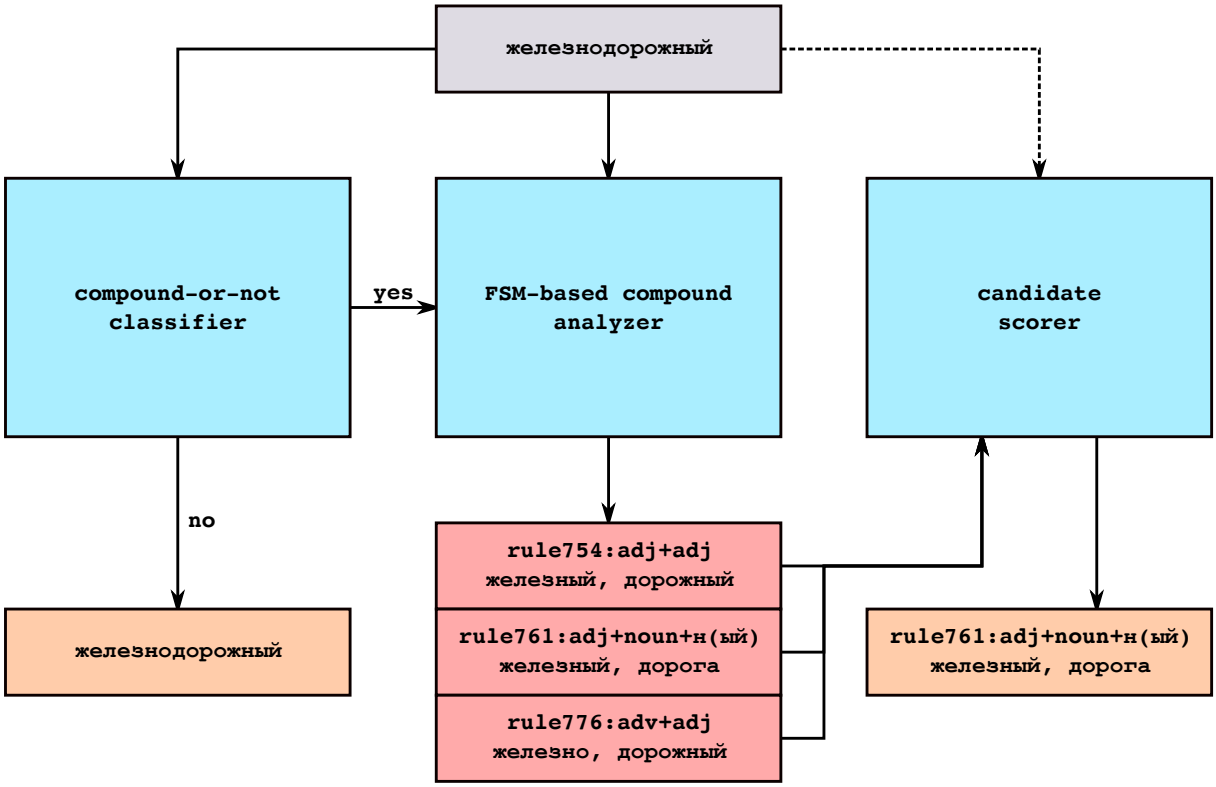


Figure 1: Diagram of the proposed pipeline. Gray: input (*zhel'eznodorozhnyy*), blue: algorithmic blocks, red: analyses generated by a rule-based model, orange: output (final analyses). The dotted line indicates that the target word is not always used in a scoring process.

acting as quasi-affixes: игро|тека '*igro|teka*' (*playroom*), cf. библио|тека '*biblio|teka*' (*library*); мега|завод '*mega|zavod*' (*megafactory*), but these words do not satisfy the definition of classical / neo-classical compounds.

Phrasal compounds are derived from the whole expressions. For example, the English word *above-mentioned* and its Russian counterpart выше|упомянутый '*vyshe|upomyanutyy*' are phrasal compounds. Such constructions may contain more than two words like in с|ума|сшедший '*s|uma|sshedshiy*' (*mad, crazy, lit. gone out of mind*), or even affixes: с|ума|сшествие '*s|uma|sshestvie*' (*madness*) (note that *сшествие '*sshestvie*' cannot be used as a separate lexeme), ничего|не|делание '*nichego|ne|delanie*' (*doing nothing*), *what|about|ism*.

The only productive phrasal compounding pattern in Russian is (adv + adj/part → adj)[4].

3 Models

In real-world applications, we do not know if a particular word is a compound. For this reason, we separate our pipeline into three stages: compound-or-not classification, rule-based splitting and analysis, and hypotheses (candidates) scoring. Figure 1 illustrates the complete process of word analysis.

3.1 Compound-or-Not Classification

First, we classify whether a word is a compound. We solve this task as a binary classification problem with a character-level bidirectional LSTM (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) with attention (Bahdanau et al., 2014). More formally, given a sequence of input character embeddings (including the special characters for the beginning and ending of the word) $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we

[4]Cf. § 778 in 'Russkaya Grammatika'.

compute

$$\begin{aligned}
\boldsymbol{h}_1^F, \ldots, \boldsymbol{h}_n^F, \boldsymbol{c}^F &= \text{LSTM}^F(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n); \\
\boldsymbol{h}_1^B, \ldots, \boldsymbol{h}_n^B, \boldsymbol{c}^B &= \text{LSTM}^B(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n); \\
\boldsymbol{h}_t &= [\boldsymbol{h}_t^F; \boldsymbol{h}_t^B], t = 1, \ldots, n; \\
\boldsymbol{q} &= \boldsymbol{Q}[\boldsymbol{c}^F; \boldsymbol{c}^B]; \\
\boldsymbol{a} &= \text{MultiheadSelfAttention}(\boldsymbol{q}, (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n)); \\
\boldsymbol{y} &= \sigma(\boldsymbol{W}\boldsymbol{a}).
\end{aligned}$$

We used the A. N. Tikhonov dictionary with morpheme segmentation[5] from (Sorokin and Kravtsova, 2018) in this experiment. Having the morpheme segmentation for all words, we computed the number of roots in them, so the target variable in the classification task was $[\#(roots) > 1]$.

3.2 Splitting and Analysis

Next, the rule-based model is applied to produce a set of hypotheses for each word.

We use the FSM concatenation technique to find all possible valid compound partitions. Most of the Russian compounds can be described with the regular expression $l(i)r(s)$, where l is the left word, r is the right word, i is the interfix applied to the left stem (optional), and s is a suffix (optional)[6]. Thus, for each rule, we use the left FSM to analyze the left part and the right one to analyze the right part. Since the FSMs are the same for many rules (e. g. noun + interfix in the left part), we pre-compute all possible FSMs for the left and right parts independently and then combine the proper two for each rule.

We constructed the FSMs using the Wiktionary vocabulary. We implemented the rules for interfixes and derivational suffixes using the `DerivBase.Ru` framework (Vodolazsky, 2020). Many derivational suffixes that are used in parasynthetic compounding are the same as for pure derivation, so we reused the existing implementation of such rules. Then we applied all rules to all words of the corresponding part of speech and stored the produced outputs in FSMs (one FSM for one rule). Thus, for instance, for the rule **rule761**([num + ITFX] + noun + н1(ый) → adj) the left FSM has ID **ruleINTERFIX**(num) and contains all numerals with interfixes, and the right FSM with ID **rule619***(noun + н1(ый) → adj) contains all utterances obtained after adding the suffix -н1(ый) '*-n1(yy)*' to all nouns from Wiktionary. See Figure 2 with the illustration of such FSMs. The IDs of the rules correspond to the paragraphs in 'Russkaya Grammatika' (Shvedova, 1980).

If a word is recognized by an FSM, then we consider such analysis as possible.

3.3 Hypotheses Scoring

Unfortunately, the described procedure is not enough, as FSMs return multiple analyses. Given a word and a set of analyses produced by a rule-based model, we score them with a model F and (optionally) select the top-k with the highest scores.

The model F scores each analysis given the left l and right r parts of the compound c.

3.3.1 Baselines

We study the following baselines:

1. Random score from a uniform distribution: $F \sim \mathcal{U}(0,1)$.
2. Frequency addition on lemmas: $F(l,r) = \#(l) + \#(r)$.
3. Frequency multiplication on lemmas: $F(l,r) = \max(1, \#(l)) \cdot \max(1, \#(r))$.
4. PMI-based score on paradigms:

$$F(l,r) = \log_2 \left(\frac{\max(1, \sum_{l_f \in C_l} \sum_{r_f \in C_r} \#(l_f, r_f) + \#(r_f, l_f))}{\max(1, \sum_{l_f \in C_l} \#(l_f)) \cdot \max(1, \sum_{r_f \in C_r} \#(r_f))} \right).$$

[5] https://github.com/AlexeySorokin/NeuralMorphemeSegmentation

[6] Note that the interfix and the suffix can modify the stems of the corresponding words

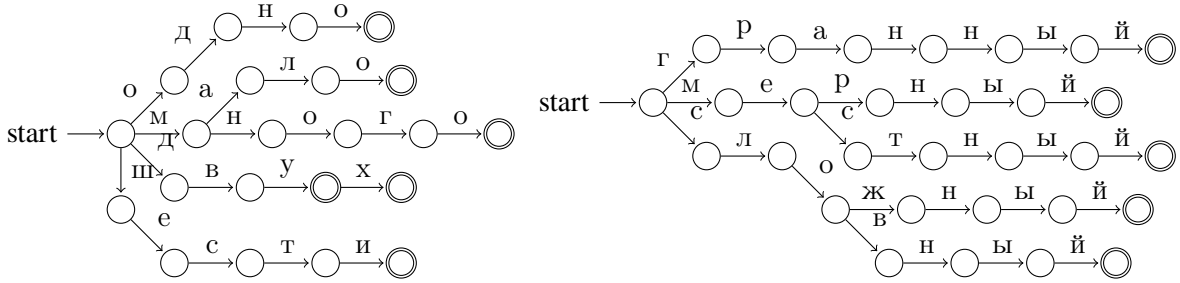


Figure 2: Left: the FSM for numerals with interfix (один '*odin*' (*one*) → одно '*odno*', два '*dva*' (*two*) → дву/двух '*dvu/dvukh*', шесть '*shest'*' (*six*) → шести '*shesti*', мало '*malo*' (*few*) and много '*mnogo*' (*many*) remain unchanged). Right: the FSM for adjectives derived from nouns грань '*gran'*' (*face (of a figure)*), мера '*mera*' (*measure*), место '*mesto*' (*place, seat*), слог '*slog*' (*syllable*), слово '*slovo*' (*word*) with the suffix -н1(ый) '*-n1(yy)*'. The concatenated FSM with ID **rule761**([num + ITFX] + noun + н1(ый) → adj) can recognize words многомерный '*mnogomernyy*' (*multidimensional*), однословный '*odnoslovnyy*' (*one-word*), etc. The corresponding analyses would be represented in the form (**rule761**([num + ITFX] + noun + н1(ый) → adj), один, слово).

5. Cosine of two lemmas based on word embeddings:

$$F(l,r) = \cos\left(\text{emb}(l), \text{emb}(r)\right);$$

6. Cosine of three lemmas based on word embeddings (Cordeiro et al., 2016):

$$F(l,r,c) = \cos\left(\text{emb}(c), \frac{\text{emb}(l)}{\|\text{emb}(l)\|} + \frac{\text{emb}(r)}{\|\text{emb}(r)\|}\right).$$

We took the unigram and bigram counts from the Russian National Corpora *n*-gram statistics[7]. To get the counts for a paradigm, we simply sum the counts for each wordform in the paradigm. Wordforms for the given lemmas were generated with `pymorphy2` (Korobov, 2015). The pre-trained `ruwikiruscorpora_upos_skipgram_300_2_2019`[8] model was used for the baselines on word embeddings.

All described baselines have at least two shortcomings. Firstly, they are completely zero-shot in terms of hypotheses scoring task, but it is desirable to fit them on a ground-truth training data. On the other hand, such approach can be applied when there is no such data. Secondly, many morphemes such as Greek roots or international quasi-affixes do not typically appear as separate tokens in corpora, therefore it is difficult to correctly estimate their frequencies and impossible to retrieve their embeddings.

3.3.2 Neural Model

In contrast to the baseline models, a neural model can fit on training data. As our training dataset is relatively small, we do not expect the model to learn the necessary linguistic patterns during training. Thus, we designed the model's architecture in a way that allows loading weights pre-trained on another, high-resource task, such as compound-or-not classification (cf. Section 3.1).

The adapted to the hypotheses scoring task model takes a compound word c and N hypotheses in the form (l, r, R_l, R_r, R_c), where l and r are as defined above, R_l and R_r are the IDs of rules applied to each of parts of the compound, R_c is the ID of the whole word-formation pattern. To process such rule IDs, we associate each of them with a trainable embedding.

Additionally, we define three special tokens p_l, p_r, p_c with the corresponding embeddings to give a model the information about words positions in the analysis. For instance, the model will learn that word био- '*bio-*' (*bio-*) is usually seen with the p_l token that corresponds to the beginning of compounds.

[7]https://ruscorpora.ru/new/corpora-freq.html
[8]https://rusvectores.org/ru/models/

Parameter	Value	Hyperparameter	Value
number of epochs	30	embedding size	128
batch size	128	embedding dropout	0.1
optimizer	Adam	body	BiLSTM
learning rate	1e-4	body hidden size	256
scheduler	constant	body dropout	0.25
objective	binary cross-entropy	number of body layers	2

Table 1: Training parameters (left) and model hyperparameters (right) for the compound-or-not classification task.

For each hypothesis the model independently processes (c, R_c, p_c), (l, R_l, p_l), (r, R_r, p_r) through a shared BiLSTM. Next, the attention is applied to the united BiLSTM outputs. A query vector is combined from the last cell states. The result of the attention is a vector that is finally fed into the classification head. Although we work here with a scoring task, we use a binary cross-entropy objective to classify whether an analysis is correct or not. During prediction, we select a candidate with the highest model's confidence.

4 Training, Evaluation, and Error Analysis

4.1 Compound Identification

We trained the model with a binary cross-entropy objective. The model hyperparameters and the training parameters are shown in Table 1.

We used 10% of the training data for validation and got 64831, 7203, 24012 samples for training, validation, and test, respectively. We selected the best checkpoint according to a validation F1 score. The resulting model achieved precision 0.9404, recall 0.9256 and F1-measure 0.9329 (4354 true positives, 276 false positives, 350 false negatives) on a test set.

As the numbers of false positives and false negatives were relatively small, we could interpret the model's errors. We found that many of them can be grouped into several categories. In addition to the obvious causes of errors, such as the incorrect gold annotation, there are other reasons for them. For example, false positives could be also caused by coincidence between parts of test words and common parts of the training compounds. False negatives could be caused by language-specific phrasal abbreviations or loanwords. See Table 2 for details.

Error Type	Examples
incorrect gold annotation	ломонос, невоеннообязанный, автономный
hyphened prefixes	по-буднишнему, по-военному, экс-король
compound-like beginning	столыпинщина, семафор
compound-like ending	задубелый, внеочередной
incorrect gold annotation	враскачку, ботвинья
loanwords	ватерпольный, ватержакетный, миастения
phrasal abbreviations	комбикорм, помдиректора, торгпредство

Table 2: Main error types in the compound-or-not classification task. Top: false positives, bottom: false negatives.

4.2 Hypotheses Generation and Scoring

We manually collected 1729 compounds with their gold-standard analyses. All compounds are taken from 'Russkaya Grammatika' (Shvedova, 1980). Then we split them into training (1143), validation (160), and test sets (364) preserving nearly equal distributions of rule IDs in the three sets. The gold dataset is stored as a table with the four columns: 'rule_id', 'compound', 'left', 'right'. Some entries of

rule_id	compound	left	right
rule579([noun + ITFX] + verb + 0m2 → noun)	водопад	вода	падать
rule754([num + ITFX] + adj → adj)	стопроцентный	сто	процентный
rule767([noun + ITFX] + verb + н1(ый) → adj)	травоядный	трава	есть
rule961([adj + ITFX] + verb → verb)	взаимодействовать	взаимный	действовать

Table 3: Samples from the evaluation dataset: водопад '*vodopad*' (*waterfall*), стопроцентный '*stoprotsentnyy*' (*one-hundred-percent*), травоядный '*travoyadnyy*' (*herbivorous*), взаимодействовать '*vzaimodeystvovat''* (*interact*).

the dataset can be found in Table 3.

For each compound, we check if a rule-based model produced the correct analysis. Given the total number of produced hypotheses, we are able to compute precision, recall and F1-measure of the rule-based model according to the following formulas:

$$P = \frac{\#(\text{correct analyses})}{\#(\text{total analyses})}; R = \frac{\#(\text{correct analyses})}{\#(\text{total examples})}; F1 = \frac{2PR}{P+R}.$$

The model achieved precision 0.0748, recall 0.7796, and F1-measure 0.1366 on a test set. The high recall score indicates that the proposed method works well on diverse data. The low precision score means that the model suffers from overgeneration and must be used only in combination with a scoring or filtering model. Since we do have a scoring model in this work, the most important metric for the analyzer is recall, as without producing a correct analysis, a scoring model will not be able to make a correct decision. We studied the errors of the analyzer model (false negatives) and found the main error sources listed below.

- Some of the examples from the evaluation dataset did not match the pattern $l(i)r(s)$, whereas some specific types of compounding involve prefixes (e. g. в|пол|голоса '*v|pol|golosa*' (*in an undertone*)), so additional modifications needed to make it possible to analyze them with our FSMs. Another case is having more than two stems, e. g. as in газо|нефте|хранилище '*gazo|nefte|khranilische*' (*oil and gas storage*). Our analysis algorithms could be modified easily, but in sake of simplicity, we did not make it in this work.

- The other source of false negatives is having the proper names and rare wordforms in one of the compound parts, e. g. чеховед '*chekhoved*' (*an expert in Chehkov's literary works*), троеборец '*troeborets*' (*triathlete*) (the reference left part is трое, while the typical form of the numeral три '*tri*' (*three*) is трёх '*tryokh*'), славянофил '*slavyanofil*' (*slavophile, lit. slavs lover*), метеоусловия '*meteousloviya*' (*weather conditions*) (славяне and условия are plural forms, therefore they did not appear in the lemmasets used for constructing FSMs).

- One more reason is small mistakes in the `DerivBase.Ru` rules, i. e. the words with zero suffixes желтощёк '*zheltoschyok*' (*yellowcheek*) and шелкопряд '*shelkopryad*' (*silkworm, lit. silk spinner*) were not correctly processed by the corresponding rules.

- Finally, some words in the evaluation dataset do not belong to any productive word-formation pattern and we did not implement rules for such occasional words: боеготовный '*boegotovnyy*' (*combat-ready*), первогодок '*pervogodok*' (*first-year*). Phrasal compounds different from (adv + adj/part → adj) also were not covered by the rules: шапкозакидательский '*shapkozakidatel'skiy*' (*cocksure, lit. related to throwing caps*).

To evaluate scoring models, we use the accuracy metric, i. e. the percentage of matches of top-1 best candidates and ground-truth analyses. The best model—neural net with pre-trained weights—achieved 60.3 accuracy. The results are presented in Table 4.

Model	Accuracy
Random (100 runs)	24.89 ± 1.62
Freq. additive (lemmas)	41.05
Freq. multiplicative (lemmas)	42.70
PMI-based (paradigms)	22.59
Cosine, two lemmas	42.42
Cosine, three lemmas	27.54
neural net, trained from scratch (30 epochs, batch size 8)	57.30
neural net, pretrained, zero-shot	31.96
neural net, pretrained, fine-tuned (30 epochs, batch size 8)	**60.33**

Table 4: Results of the scoring models.

5 Conclusion and Future Work

In this work, we presented our approach to compound identification, splitting, and analysis for Russian. We solved this task using the combination of a rule-based method, finite state machines, and neural networks. The models we used achieved the high F1 score on a compound classification task and the high recall score on a hypotheses generation task. We evaluated and compared the five unsupervised scoring functions based on word frequencies and distributional semantics (word embeddings). The dataset used as the gold standard was manually collected and annotated by us and will be publicly available on our GitHub.

One of the directions of our future work is making an end-to-end pipeline. Also, we aim to try different model architectures such as convolutional neural networks or transformers instead of BiLSTMs. We hope that this will improve the quality of compound analysis.

The second direction is an application of the proposed algorithm to a large-scale vocabulary and integration of compounds into the Russian part of Universal Derivations (Kyjánek et al., 2019).

Finally, we plan to collect a larger dataset for training and evaluation of compound analysis models.

Acknowledgments

We thank Igor Galitskiy for providing his computing server and the anonymous reviewers for their comments and suggestions.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Antonietta Bisetto and Chiara Melloni. 2008. Parasynthetic compounding. *Lingue e linguaggio* 7(2):233–260.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1986–1997. https://doi.org/10.18653/v1/P16-1187.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks* 18(5-6):602–610.

ES Gromenko. 2020. Special'naja leksika xx veka kak ob'ekt neografii: K istorii voprosa (na materiale slovarja-spravoČnika po materialam pressy i literatury 60-h gg.). *Vestnik Nižegorodskogo universiteta im. NI Lobačevskogo* (3).

Verena Henrich and Erhard Hinrichs. 2011. Determining immediate constituents of compounds in GermaNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Association for Computational Linguistics, Hissar, Bulgaria, pages 420–426. https://www.aclweb.org/anthology/R11-1058.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves bert's interpretation of complex words.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *10th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Budapest, Hungary. https://www.aclweb.org/anthology/E03-1076.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts*, Springer International Publishing, volume 542 of *Communications in Computer and Information Science*, pages 320–332.

Lukáš Kyjánek, Zdeněk Žabokrtskỳ, Magda Ševčíková, and Jonáš Vidra. 2019. Universal derivations kickoff: A collection of harmonized derivational resources for eleven languages. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. pages 101–110.

Haonan Li, Zhisong Zhang, Yuqi Ju, and Hai Zhao. 2018. Neural character-level dependency parsing for Chinese. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA), Lisbon, Portugal. http://www.lrec-conf.org/proceedings/lrec2004/pdf/468.pdf.

Rico Sennrich and Barry Haddow. 2015. A joint dependency model of morphological and syntactic structure for statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 2081–2087.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* .

Natalija Shvedova. 1980. *Russkaja grammatika*. Number t. 1 in Russkaja grammatika. Izd-vo Nauka. https://books.google.ru/books?id=7BNgAAAAMAAJ.

Alexey Sorokin and Anastasia Kravtsova. 2018. Deep convolutional networks for supervised morpheme segmentation of russian language. In Dmitry Ustalov, Andrey Filchenkov, Lidia Pivovarova, and Jan Žižka, editors, *Artificial Intelligence and Natural Language*. Springer International Publishing, Cham, pages 3–10.

Bogdan Szymanek. 2017. Compounding in Polish and the absence of phrasal compounding. *Further investigations into the nature of phrasal compounding* 1:49.

Daniil Vodolazsky. 2020. DerivBase.Ru: a derivational morphology resource for Russian. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 3937–3943. https://www.aclweb.org/anthology/2020.lrec-1.485.

Marion Weller-Di Marco. 2017. Simple compound splitting for German. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, Valencia, Spain, pages 161–166. https://doi.org/10.18653/v1/W17-1722.

Index

Bobkova, Natalia, 13, 31
Bonami, Olivier, 14, 42

Dal, Georgette, 13, 52
Duboisdindien, Guillaume, 13, 52

Filko, Matea, 13, 120

Generalova, Valeria, 13, 61

Hathout, Nabil, 13, 14, 70, 114
Huguin, Mathilde, 13, 76
Huyghe, Richard, 14, 21

Knittel, Marie-Laurence, 14, 86

Laks, Lior, 13, 95
Litta, Eleonora, 13, 105

Mambrini, Francesco, 13, 105
Marín, Rafael, 14, 86
Moretti, Giovanni, 13, 105

Namer, Fiammetta, 13, 14, 70, 95, 114

Padó, Sebastian, 13, 20
Passarotti, Marco, 13, 105
Pellegrini, Matteo, 13, 105
Petrov, Hermann, 14, 149

Sanacore, Daniele, 13, 114
Svoboda, Emil, 14, 129

Tribout, Delphine, 14, 42

Vidra, Jonáš, 13, 139
Vodolazsky, Daniil, 14, 149

Ševčíková, Magda, 14, 129
Šojat, Krešimir, 13, 120
Štefanec, Vanja, 13, 120
Žabokrtský, Zdeněk, 13, 139

Published by:

ATILF (CNRS & UNIVERSITÉ DE LORRAINE)
44, avenue de la Libération
BP 30687
54036 Nancy Cedex
France

ISBN: 978-2-9580006-0-8