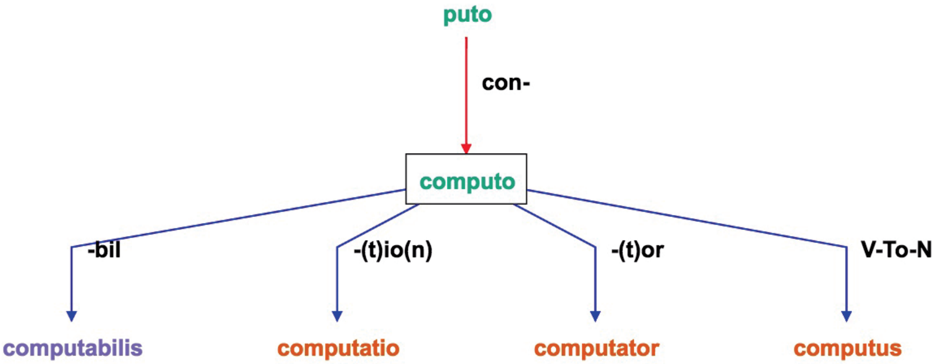
Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)

5-6 October 2017, Milano, Italy



edited by

Eleonora Litta
Marco Passarotti

EDUCatt

CIRCSE
CENTRO INTERDISCIPLINARE
DI RICERCHE
PER LA COMPUTERIZZAZIONE
DEI SEGNI DELL'ESPRESSIONE

MARIE CURIE ACTIONS

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

5 - 6 October 2017
Milano, Italy

Editors:
Eleonora Litta
Marco Passarotti

Published by
EDUCatt
Ente per il Diritto allo Studio Universitario dell'Università Cattolica
Largo Gemelli 1, 20123 Milano - tel. 02.72342235 - fax 02.80.53.215
e-mail: editoriale.dsu@unicatt.it (produzione); librario.dsu@unicatt.it (distribuzione)
web: www.unicatt.it/librario
ISBN: 978-88-9335-225-3

Cover illustration: http://wfl.marginalia.it

Preface

Recent years have seen a growing interest in research aimed at building new linguistic resources and Natural Language Processing (NLP) tools for derivational morphology. For decades, research in computational morphology was mainly focussed on its inflectional aspects and, specifically, on PoS tagging. The current increased interest in both the theoretical and applicative aspects of word formation is strictly connected to the large need for automatic semantic processing of linguistic data. Indeed, strict relations do hold between derivational morphology and semantics, as words that share the same formative elements or the same formative process also tend to share basic semantic features, which can in turn be induced automatically from those of their lexical basis.

Several lexical resources for derivational morphology have been made available for a number of languages. Among them are the lexical network for Czech DeriNet (Ševčíková and Žabokrtský [23]), the derivational lexicon for German DERIVBASE (Zeller et al., [26]) and that for Italian derIvaTario (Talamo et al., [24]). Furthermore, stemming is a technique largely used for detecting word formation processes (Goldsmith [9]), and language independent probabilistic NLP tools were developed to extract derivation information from lexical data (Baranes and Sagot [3] 2014; Virpioja et al. [25]).

Over the last decade many efforts have been invested in the creation of advanced language resources and tools for ancient languages, notably the linguistic annotation of Latin and Ancient Greek textual data through treebanks (Bamman et al. [2]; Bamman & Crane [1]; Haug & Jÿhndal [11]; Korkiakangas & Lassila [13]; Passarotti [19]). Numerous computational lexical resources for these languages have also been developed (McGillivray [16]; McGillivray & Passarotti [15]; Minozzi [17]; Passarotti et al. [21]).

In that time, what had been missing was a derivational lexicon and NLP tool for Latin. When in 2014 we decided to write a project proposal for a Marie Curie Individual Fellowship, we felt that times were ripe to address such a challenge. In our research experience before then, we had contributed to building a powerful morphological analyser for Latin (Lemlat: Passarotti et al. [22]) and to running the Index Thomisticus Treebank (Passarotti [19]) –currently the largest Latin treebank available– for more than a decade.

Derivational morphology was the missing link between inflectional morphology and syntax, so it seemed the natural next step to address. Our project proposal received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 658332-WFL. In November 2015, we began to build what later became “Word Formation Latin" (WFL: `http://wfl.marginalia.it`) (Litta et al. [14]), a word formation based lexicon and tool for Latin. The work was carried out at the CIRCSE Research Centre of Università Cattolica del Sacro Cuore in Milan (`http://centridiricerca.unicatt.it/circse_index.html`). Now, in October 2017, the project is approaching its end. To celebrate it and to consider the current status of research in the field, we organised the Workshop on Resources and Tools for Derivational Morphology (DeriMo), whose contributions are collected in these proceedings. DeriMo is not an isolated event, but it is one of the initiatives organised in the area of derivational morphology in the past few years, testifying to the growing interest in various aspects of word formation in linguistics. [1]

The Call for Papers asked for long abstracts (up to 6 pages), describing original, unpublished research, either complete or ongoing. In total, we received 14 submissions from 9 different countries in Europe and Asia. Each submission was reviewed in a double-blind fashion by three of the 28 members of the workshop's programme committee. Of the 14 submissions, 11 were accepted. The overall acceptance rate was 79%, which indicated that the average quality of the abstracts was high.

The programme opens with an invited lecture by Pius ten Hacken (University of Innsbruck, Austria) on *Computer Models and Mental Models of Derivational Morphology*. He introduces two fundamental approaches to derivational morphology in computational linguistics, *Two-Level Morphology* and *Word Manager*, evaluating how these tackle a number of issues in linguistics as well as their theoretical implications.

The workshop hosts sessions dedicated to four main themes:

1. a presentation of WFL, followed by two investigations of derivational morphology made possible thanks to the resource;

2. updates and expansion on existing resources;

3. software and algorithm development for derivational morphology that can be applied across different resources;

4. theoretical linguistics issues linked to derivation in Indo-European languages,

[1] See, for instance, the Workshop on Derivational Morphology and Spoken Language (22nd June, 2016; University of Reading, UK: `https://www.reading.ac.uk/english-language-and-applied-linguistics/News/elal_British_Academy_Workshop_June_2016.aspx`), the “First Workshop on Paradigmatic Word Formation Modeling" (ParadigMo 2017, 19th-20th June, 2017; University of Toulouse, France: `http://w3.erss.univ-tlse2.fr/ParadigMo2017/`) and the conference “The Word and the Morpheme" (22nd-24th September, 2016; Humboldt Universität Berlin, Germany: `https://www.angl.hu-berlin.de/department/staff-faculty/professors/alexiadou/workshops/workshopwordmorpheme`).

including the presentation of a digital implementation of Pāṇini's derivational morphology of Sanskrit.

Word Formation Latin (WFL) is a derivational morphology resource for Classical Latin. The contents of WFL are lexical items (represented by lemmas) analysed into their formative components. Relationships between the lexical items are established on the basis of word formation rules (WFRs). For example, the relation tying lemmas *mitto* 'to send' and *admitto* 'to send to' describes a change from a verb to another verb through the addition of a prefix that in itself bears semantic information: the prefix *ad-* generically characterises movement *towards* something.

The lexical basis for WFL is identical to that of the morphological analyser and lemmatiser for Latin Lemlat, now available in its third version (Passarotti et al. [22]). Lemlat is the result of the collation of three Latin dictionaries (Georges and Georges [7]; Glare [8]; Gradenwitz [10]), and contains 40,014 lexical entries and 43,432 lemmas (as more than one lemma can be part of the same lexical entry). Moreover, the lexical basis of Lemlat has recently been integrated with the addition of most (26,250 lemmas out of 28,178) of the Onomasticon contained in the Forcellini lexicon (Budassi & Passarotti [4]).

The WFL data is collected and organised in a MySQL relational database as follows: 1) A list of WFRs was obtained both manually and automatically; the WFRs were then identified and formalised into a table according to their type (prefixal, suffixal, compound and conversion) and to the category of transformation undergone by the lexical element in input (N-to-N, N-to-V, N-to-A, etc.). 2) A series of SQL queries is applied to the lexical data in order to pair input (origin of the derived lemma) with output (derived) lemmas according to one WFR at a time. 3) The resulting list of candidate pairs is thoroughly checked manually for coherence and amended where needed.

The WFL lexicon is now available online through a visualisation query system currently at `http://wfl.marginalia.it`. The data can be browsed according to four different perspectives implemented as four different screens, which can be accessed via a top-level menu. These represent the conceptualisation of the kinds of research questions and results that we hypothesise a user might be interested in.

The data is visualised as a list of lemmas matching a query, or as derivational (tree-like) graphs representing the derivational cluster for a specific lemma. The tree includes all the lemmas derived from the lemma selected, as well as all those words the lemma is derived from. In the cluster, lemmas are nodes and WFRs are edges.

In their paper, Budassi and Litta give an account of an experience made during the compilation of WFL. The process of inserting the multiform and rich class of suffixed *-sco* verbs highlighted some linguistic theory issues that arose from pigeonholing such verbs into the morphotactic model adopted in the resource. These issues have manifested themselves across the entire the lexical basis. Budassi and Litta propose a possible alternative perspective on derivational relationships for a series of problematic cases in the form of derivational paradigms. Their paper represents the first step towards the design of a possible 2.0 version WFL featuring a dual view of word formation families.

M. Silvia Micheli's work deals with compounding in Latin and Italian. After an overview of how compounding is treated in WFL, Micheli analyses the fate of Latin compounds in Italian from a morphological point of view, focussing on what has survived and what has been lost. Micheli shows that most Latin compounds were either lost or re-analysed as derived or simple words. This resulted in discontinuity between Latin and Italian compounding rules, and in a system reorganisation common to all Romance languages in compound word formation.

In their contribution, Namer *et alii* expand on the derivational database of French Demonette (Hathout & Namer [13]). The structure of the relational database includes properties of derivational relations connecting word pairs. The entries also specify the categorical, semantic and morpho-phonological properties of the connected words. The paper describes these morpho-phonological properties and shows how Demonette's organisation gives an original representation of these properties, together with phonological transcriptions of the word pairs and syllabic decompositions, their stems and variants.

Ševčíková *et alii* relate on the expansion of the lexical database of Czech DeriNet. This is done through a semi-automatic method of adding derivational links by identifying verbs which are derived by suffixation and constitute aspectual pairs. The contribution presents an approach toward the identification of aspectual pairs based on their extraction from the VALLEX valency dictionary, the identification of suffix substitution rules and the subsequent manual annotation. This process results in the addition of almost 6,000 derivational links to the existing DeriNet database.

Papay *et alii* describe the graph-theoretical approach they employ to evaluate and improve the German derivational lexicon DERIVBASE. The representation of derivational families, very similar to that of WFL, with labelled directed graphs in which words are nodes and relationships are directed edges, allows for a large-scale comparison of the structure of different derivational families and for the automatic identification of possible errors in the resource. A manual evaluation of this method's predictions is carried out to verify that it can successfully spot instances that are missing from DERIVBASE. This method highlights linguistic theory issues, as the predictions in this approach can be interpreted as the result of interplay among productivity constraints.

Filko & Šojat's contribution is also part of the main theme associated with the development and evaluation of existing derivational morphology resources. The authors present the expansion of the derivational database for Croatian CroDeriV, previously containing only verbs, with adjectives. Lemmas are collected from free corpora and digital dictionaries. The paper gives a good overview of major derivational processes in Croatian, and the structure of the derivational database, before discussing the methodology employed for the expansion of the database, in view of the experience gained when building the verbal category in the first phase of the project.

The paper by Vidra & Žabokrtský forms part of the session dedicated to the development of software or resources for the treatment of derivational morphology. The authors present two studies on tools developed for searching and viewing lexical derivational databases containing large amounts of clusters with many nodes. The first study describes a query language created specifically for searching databases of lexical derivations and shows its implementation in the DeriSearch online application. The second study discusses experiments carried out with visualisations of large derivational trees.

Shafaei *et alii* highlight how current derivational lexicons, although fundamental for the development of computational linguistics resources, lack in comparability. The authors present an algorithm that extracts these lexicons from the German morphological layer of CELEX, a lexical database available for English, Dutch, and German, making a step towards the creation of more comparable derivational lexicons for these languages. An evaluation is performed on the resulting DErivCelex against DERIVBASE, a large derivational lexicon of German created semi-automatically.

In the session dedicated to linguistic theory of derivational morphology in Indo-European languages, Panocova argues that Slovak international nouns ending in *-ácia* serve as the basis for verb formations. This paper investigates the direction of motivation in pairs of mostly Latin origin, such as *diverzifikovat'* 'diversify' > *diverzifikácia* 'diversification'. In the Slovak linguistic tradition, these pairs were analogically modelled as derivations from verbs to nouns. This paper discusses two types of evidence, which suggest that the direction of motivation is actually the opposite. One type is based on frequency, the other on the meaning of the two members of the pair.

Pultrová postulates that the diachronic distinction between inherited versus non-inherited has important implications for the synchronic semantic and formal analysis of word formative types. The paper also illustrates that the distinction between inherited and non-inherited (hence analogical) formations often plays a crucial role in the description of the phonological system of a language.

Finally, Scharf gives an overview of the efforts made to produce an XML formalisation of Pāṇini's linguistic system. Pāṇini's linguistic system consists of a set of about 4,000 rules, that classify semantic objects, add affixes to basic roots and nominal bases under semantic and co-occurrence conditions, and make morphophonemic and phonetic modifications to reconstruct utterances of the language. Each rule organises a set of regular expressions and attributes into a tree consisting of XML elements. The aim is to produce a comprehensive lexicon of Sanskrit hierarchically categorised under the verbal roots, and indexed according to semantic, syntactic, morphological, and inflectional factors as well as rules applied in the course of derivation.

Overall, we think that the four themes provide an extensive overview of the theoretical, methodological and practical questions related to resources and tools for derivational morphology.

We hope you will enjoy the workshop and the proceedings. We wish to thank all of the authors who submitted papers, the members of the programme committee, Pius ten Hacken, who agreed to give the invited talk, Savina Raynaud (the Director of CIRCSE) and our colleagues, PhD candidates and students who helped us organise DeriMo.

The Co-chairs of DeriMo:
Eleonora Litta and Marco Passarotti

References

[1] Bamman, David and Crane, Gregory. The design and use of a Latin dependency treebank. In J. Nivre and J. Hajic (Eds.), Proceedings of the Fifth Workshop on Treebank and Linguistic Theories (TLT2006). Prague, Czech Republic: ÚFAL, pp. 67–78, 2006.

[2] Bamman, David, Mambrini, Francesco and Crane, Gregory. An ownership model of annotation: The Ancient Greek Dependency Treebank. In *Proceedings of TLT 8*, pp. 5–15. Milano: EDUCatt, 2009.

[3] Baranes, Marion and Sagot, Benoît. A Language-Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). ELRA, Reykjavik, Iceland, pp. 2793–2799, 2014.

[4] Budassi, Marco and Passarotti, Marco. Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon,*Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), Berlin, Germany*, The Association for Computational Linguistics, pp. 90–94, 2016.

[5] Forcellini, Egidio. *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Corradini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos. Perin . Typis Seminarii*, Padova, 1940.

[6] Fruyt, Michèle. Word Formation in Classical Latin. In James Clackson (ed.) *A companion to the Latin language*, pp. 157–175. Chichester: John Wiley & Sons, 2012.

[7] Georges, Karl Ernst. *Ausführliches lateinisch-deutsches und deutsch-lateinisches Handwörterbuch.* Vol. 2. Hahn'sche Verlags-buchhandlung, 1880.

[8] Glare, Peter GW. *Oxford latin dictionary*. Oxford: Oxford University Press, 1982.

[9] Goldsmith, John. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2): pp. 153–198, 2001.

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

[11] Haug, Doug and Jÿhndal, M. Creating a Parallel Treebank of the Old Indo-European Bible Translations, in K. Ribarov, C. Sporleder (eds.), *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008)*, ELRA, Marrakech, pp. 27-34, 2008.

[12] Jenks, Paul Rockwell. *A Manual of Latin Word Formation for Secondary Schools*. Boston/NewYork/Chicago: DC Heath & Company, 1911.

[13] Korkiakangas, Timo and Lassila, Matti. Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material. In Francesco Mambrini, Marco Passarotti & Caroline Sporleder (eds.), *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities*, pp. 61–72. Sofia: Bulgarian Academy of Sciences, 2013.

[14] Eleonora Litta, Passarotti, Marco and Culy, Chris. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*. Napoli, aAccademia University Press, pp. 185–189, 2016.

[15] McGillivray, Barbara and Passarotti, Marco. The Development of the Index Thomisticus Treebank Valency Lexicon. In Proceedings of LaTeCH-SHELT&R Workshop 2009. Athens, Greece: ACL, pp. 43–50, 2009.

[16] McGillivray, Barbara. *Methods in Latin Computational Linguistics*. Leiden: Brill, 2013.

[17] Minozzi, Stefano. The Latin WordNet project. In P. Anreiter and M. Kienpointner (Eds.), *Latin Linguistics Today. Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik*. Innsbruck: Innsbrucker Beiträge zur Sprachwissenschaft, pp. 707–716, 2010.

[18] Oniga, Renato. *I composti nominali latini: una morfologia generativa*. Bologna: Pàtron, 1988.

[19] Passarotti, Marco. Language Resources. The State of the Art of Latin and the Index Thomisticus Treebank Project. M.S. Ortola (ed.), *Corpus anciens et bases de données*. ALIENTO N.2. Presses universitaires de Nancy, Nancy, pp. 301–320, 2011.

[20] Passarotti, Marco and Mambrini, Francesco. First Steps towards the Semi-automatic Development of a Wordformation-based Lexicon of Latin, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012). May 23-25, 2012, Istanbul, Turkey*, 852-859, ELRA, 2012.

[21] Passarotti, Marco, González Saavedra, Berta and Onambele, Christophe. Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Grobelnik, Marko, Maegaard, Bente, Mariani, Joseph , Moreno, Asuncion, Odijk, Jan, Piperidis, Stelios (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). May 23-28, 2016,* Portoroz, Slovenia: European Language Resources Association (ELRA), pp. 2599–2606, 2016.

[22] Passarotti Marco, Budassi, Marco, Litta, Eleonora and Ruffolo, Paolo. The Lemlat 3.0 Package for Morphological Analysis of Latin, in Bouma, Gerlof and Adesam, Yvonne (eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. 22nd May 2017 Gothenburg*, Northern European Association for Language Technology (NEALT) Proceedings Series, Vol. 32. Linköpings Universitet: Linköping University Electronic Press, pp. 24–31, 2017.

[23] Ševčíková, Magda and Žabokrtský, Zdeněk. Word-Formation Network for Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland:ELRA, pp. 1087–1093, 2014.

[24] Talamo, Luigi, Celata, Chiara and Bertinetto, Pier Marco. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, 9(1): pp. 72–102, 2016.

[25] Sami Virpioja, Smit, Peter, Grönroos, Stig-Arne and Kurimo, Mikko. *Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline*. Helsinki: Aalto University, 2013.

[26] Zeller, Britta D., Snajder, Jan and Padó, Sebastian. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria: ACL, pp. 1201–1211, 2013.

Programme Committee

Chairs:
Eleonora Litta (Università Cattolica del Sacro Cuore, Milano, Italy)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milano, Italy)

Members:
Mark Aronoff (USA)
Piermarco Bertinetto (Italy)
Jim Blevins (UK)
Nabil Hathout (France)
Dag Haug (Norway)
Gerd Haverling (Sweden)
Andrew Hippisley (USA)
Claudio Iacobini (Italy)
Sandra Kübler (USA)
Rochelle Lieber (USA)
Silvia Luraghi (Italy)
Cerstin Mahlow (Germany)
Francesco Mambrini (Germany)
Fiammetta Namer (France)
Renato Oniga (Italy)
Sebastian Padó (Germany)
Renáta Panocová (Slovakia)
Vito Pirrelli (Italy)
Lucie Pultrová (Czech Republic)
Jan Radimský (Czech Republic)
Savina Raynaud (Italy)
Benoît Sagot (France)
Magda Ševčíková (Czech Republic)
Andrew Spencer (UK)
Pavel Štichauer (Czech Republic)
Zdeněk Žabokrtský (Czech Republic)

Contents

Computer Models and Mental Models of Derivational Morphology

Pius ten Hacken

Leopold-Franzens-Universität Innsbruck
E-mail: pius.ten-hacken@uibk.ac.at

Abstract

Models of derivational morphology have been developed both in computational linguistics and in theoretical linguistics. In both domains, the aim is to express relations between lexical entries. In theoretical linguistics, these relations are meant to correlate with the organization of the mental lexicon. For computer models, the main concern is that they serve a system solving a particular problem of computational linguistics. Here I present two examples of basic approaches to derivational morphology in computational linguistics, Two-Level Morphology and Word Manager, and consider how they stand to some issues in linguistic theory and how they give insights that can be interpreted in theoretical linguistics.

In a database of morphology, it is necessary to design a model of morphological rules, lexical units, and the relationship between them. Here, I will focus on the modelling of derivational morphology, but it will not be possible to exclude considerations pertaining to inflection and compounding, because of the interaction between them. Models of morphology have also been developed in linguistic theory. The main criterion for such linguistic models is that they contribute to an explanatory model of the mental realization of language. The question to be studied here is to what extent computer-oriented models and linguistic models can inform each other. First, section 1 gives some background from the computational perspective. Then, I turn to some central issues in the linguistic modelling of morphology in section 2 and the nature of the notion of *word* in section 3. On this basis, section 4 investigates how central linguistic issues are treated in a computational context and section 5 how computational modelling can be used in linguistic theorizing.

1 Models of morphology in computational linguistics

In computational linguistics, it is often possible to observe a certain tension between two types of purpose. On one hand, there is the computational emphasis on the development of applications that perform a particular task or solve a particular problem. On the other hand, there is the linguistic emphasis on developing models of language that can be interpreted as theories of linguistics or used to select a theory among competitors. In the case of Machine Translation (MT), the second goal was assigned an overwhelming importance in the rule-based approach of the 1970s and 1980s. As I argue in ten Hacken [21], the current generation of more performing MT systems was only possible because this goal was abandoned in favour of supporting actual translators.

In the case of derivational morphology as a component of computational linguistics, we are dealing with a field of a rather different type than MT. In computational morphology, the result of development is not an application, but a component. The idea of developing reusable components for a wide range of systems of computational linguistics emerged in the late 1980s, when the so-called lexical bottleneck was discovered. The image of a bottleneck was used to visualize how systems of computational linguistics were not able to realize their full potential in practice, because their lexicon was very small. This inspired research of the type collected in Atkins & Zampolli [4] and Walker et al. [35].[1] Much of this research focused on the reusability of existing dictionaries for applications of computational linguistics.

It is in this context that morphology emerged as a computational problem. Text words are not always dictionary words. The main task of a morphological component is to bridge this gap. Not all words in a text that are not in a dictionary are linked to dictionary words by morphological rules, but only for the ones for which there is such a link is there a rule-based approach to covering them. As shown in Sproat's overview [34], Two-Level Morphology, devised by Koskenniemi [28] and Karttunen [26], was the dominant approach at least until the early 1990s. In its pure form, this approach adopts a model with only a set of formatives and a set of rules, as in Fig.1.

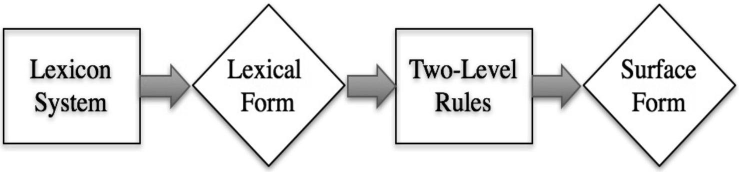


Figure 1: The model of Two-Level Morphology

In Fig.1, morphology is divided into a concatenative component in the Lexicon System and a non-concatenative component in the Two-Level Rules. Formatives in the Lexicon System have a continuation class, which specifies which formatives may follow. They are grouped into sublexicons. A continuation class is a set of sublexicons. Thus, English regular verbal endings are a sublexicon and all regular

[1] For the chronology of events, it is important to note that Atkins & Zampolli [4] is derived from a summer school in Pisa in 1988, Walker et al. [35] from a workshop in Grosseto in 1986.

verb stems have a continuation class linking to it. Two-Level Rules are responsible for regular non-concatenative changes, such as the transformation of *try+s* into *tries*.

As a reaction to perceived shortcomings of the Two-Level approach, database systems for morphological dictionaries were developed. An early example is Domenig's *Word Manager* [13]. Ten Hacken [18] gives an overview of the system in its later stages and of databases and applications developed with Word Manager. The model of Word Manager can be represented as in Fig.2.[2]

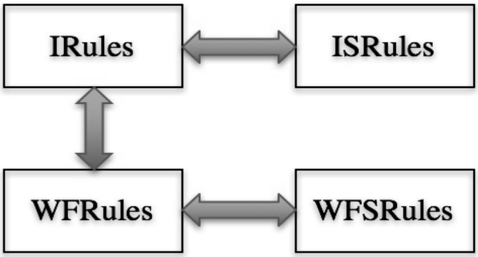


Figure 2: The model of Word Manager

In Fig.2, IRules and WFRules model inflection and word formation, respectively. SRules are spelling rules, corresponding in their purpose to the Two-Level Rules in Fig.1. The main difference with the model in Fig.1 is the role of lexemes. Whereas Two-Level Morphology is a system for generating text words from formatives, Word Manager generates lexemes, organized lists of connected word forms. The difference between IRules and WFRules is that the application of an IRule represents a lexeme and the application of a WFRule produces a new lexeme and assigns it to the appropriate IRule. The centrality of the lexeme is at the basis of the so-called *Bow Tie Model* in Fig.3.

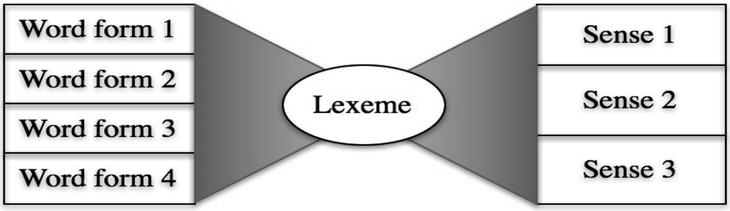


Figure 3: The Bow Tie Model

Fig.3 represents two mappings, on one hand between lexeme and word forms

[2] Fig.2 only represents the Word Manager core, as presented in Domenig & ten Hacken [14], not the extensions for clitics and multi-word units in Phrase Manager, as presented originally in Pedrazzini [33].

and on the other between lexeme and senses. These mappings are independent of each other. The idea is that Word Manager covers the entire mapping between word forms and lexemes without distinguishing senses. In this way, applications in computational linguistics can start from lexemes rather than from text words.

Two-Level Morphology and Word Manager are both among the earlier computational systems for morphology, but they are still of interest because they make choices in the treatment of problems that newer systems also have to deal with. Therefore, I will take these models as a basis for the discussion.

2 Models of morphology in linguistics

In linguistics, the aim is to develop a theory of language. Such a theory specifies first of all what language is and then proposes a description and an explanation of selected parts of this object. In line with mainstream assumptions of Chomskyan linguistics, I will assume here that the primary manifestation of language is the speaker's competence.[3] For morphology, this means that the central question is how it is implemented in a speaker's mental language system.

Some of the basic design issues in theories of morphology are listed in (1). They are generally older than Chomskyan linguistics.

1. (a) What is the position of morphology in the architecture of grammar?

   (b) Should morphology be divided into inflection and word formation?

   (c) Do morphological rules arrange morphemes or apply processes?

   (d) Should morphemes be divided into stems and affixes or into free and bound morphemes?

The questions in (1) are not all at the same level. Some of them imply specific answers to others. (1a) includes the question of the autonomy of morphology. Halle & Marantz's Distributed Morphology (DM) [22] and Jackendoff's Parallel Architecture (PA) [25] both give a negative answer to this question, though from very different starting points. Other aspects of (1a) involve the interaction with the lexicon and with (different components) of syntax. (1b) is a classical issue discussed already by Bloomfield [6]. As described in ten Hacken [20], there is a strong sceptical tradition that tends to deny the possibility of a systematic division, but from a terminological perspective, the question is only whether we want to impose a boundary between inflection and word formation or not. If desired, a definition can be formulated and applied. For derivational morphology, (1a-b) determine whether it belongs to a specialized component or not. If not, we can still use the term descriptively.

[3] Chomskyan linguistics is much broader than the theories of Noam Chomsky. As described in ten Hacken [17], the central place of competence is also assumed in related research programmes such as Lexical-Functional Grammar and Jackendoff's Parallel Architecture.

In (1c) we turn to the nature of the rules of morphology. The opposition between Item & Arrangement (IA) and Item & Process (IP) was originally formulated by Hockett [23]. Although it is possible to reformulate any rule from IA into IP and the reverse, there is still the question of how competence is actually organized. DM and PA both adopt IA as a rule format. Anderson [2] argues for an IP model on the basis of the importance of processes changing the form of their input (e.g. *sing* > *song*). In IP, there are no morphemes, which explains Anderson's name of *a-morphous morphology* for his model. Question (1d) only arises in IA models. What is at issue is whether morphemes should only be specified for their distributional properties (free or bound) or should be divided into separate classes of stems and affixes on the basis of other, additional criteria.

The questions in (1) have been the subject of much debate. This is because it is not merely a choice of convenience. It may be difficult to find convincing evidence for one answer or another, but unless we deny the existence of a speaker's competence, there is an empirical reality that the questions are about. The right answer is not just the most convenient one, but the one that corresponds to a speaker's competence.

3 The notion of *word*

Before we can turn to a comparison of computational and mentalist perspectives on morphology, we have to consider the notion of *word* in a bit more detail. Morphology is the study of the structure of words. However, different concepts have been designated by *word*. Various parameters can be used to characterize their relationships. An obvious parameter to start with is the distinction between competence and performance. A word in competence is a combination of form and meaning. In performance, whether written or spoken, only the form is realized. Meaning is only assigned when competence is used to interpret the form. There is also the sense of a word of a language, for instance as realized in a dictionary. As I show in ten Hacken [19], this is not an empirical notion, but one that is derived from the interpretation of a collection of performance and competence data.

One sense of *word* is what Di Sciullo & Williams [12] call a *listeme*. They consider a listeme the minimal unit that must be listed in the mental lexicon of a speaker, because it displays properties that cannot be derived by rules. Chomsky's Lexicalist Hypothesis [11] uses a similar reasoning to argue that nominalization (and by extension, derivational morphology) has to be in the lexicon and not in a kind of extension to the syntax. Jackendoff [25] turns this argument on its head. He shows that there is a fluent transition between words, lexicalized phrases, idioms, and syntax rules. From this he concludes that words and rules are listemes of basically the same type.

Another understanding of *word* is as a domain for a particular type of rule application. The nature of the domain is defined by the class of rules. Thus, we can see the word as a phonological domain, when we say, for instance, that Polish has stress on the penultimate syllable of a polysyllabic word, or as an orthographic domain, when we say that a word is written together. Note that such domains are in principle independent of the distinction between competence and performance. We can observe words in spoken or written performance, but they reflect units of competence. Domains can also be syntactic or morphological, based on the application of syntactic or morphological rules, although here the empirical basis in performance is less direct. This is even stronger for the semantic word, i.e. a form that designates one concept.

A final sense, which is particularly relevant in the discussion of derivational morphology, is the word as a lexeme. Originally proposed in the context of Matthews's Word and Paradigm (WP) [31] morphology, *lexeme* is a term that unites in it a large part of the questions related to the nature of morphology. A lexeme is an organized pattern of word forms such as that for a combination of features, the corresponding word form is listed. Matthews [31] presents WP as an alternative to both IA and IP. In his word-based morphology, Aronoff [3] uses *word* in the sense of *lexeme*, so that word formation is the formation of new lexemes. This presupposes a distinction between inflection and word formation.

4 Linguistic theory in computational models

The computational models presented in section 1 both focus on orthographic words, although in principle it would be possible to apply them to transcriptions of spoken language. From this perspective, we can then turn to the questions in (1) as a way of characterizing the computational models.

Among the questions in (1), (1c) is perhaps the best starting point. In Two-Level Morphology, the continuation classes deal with concatenative morphology and the two-level rules with non-concatenative morphology. In Word Manager, a similar division can be observed, where IRules and WFRules do the former and SRules the latter. At first sight, it may then seem that both combine IA rules for arranging morphemes with IP rules for changing their forms. However, IP is characterized by the uniform treatment of all phonological changes induced by the rule. Affixes are treated as phonological changes. This would mean using two-level rules and SRules for affixation. Such an encoding of morphology is probably possible, but it would be very unnatural. Therefore, it is safe to conclude that both models are basically IA-oriented. Two-level rules and SRules are what Aronoff [3] calls *adjustment rules*.

For the stem-affix distinction in (1d), it is important that Two-Level Morphology is based on finite-state rules, whereas Word Manager uses context-free rules. In Word Manager, stems have a clearly distinct role to affixes. Whereas stems are the base of a lexeme, affixes depend on rules. In Two-Level Morphology, Koskenniemi [28] assumes a special role for the Root Lexicon. The Root Lexicon is the starting point for concatenation and it contains the stems. However, his focus on Finnish means that all affixation is suffixation. For languages with a combination of suffixation and prefixation, including all Indo-European languages, we cannot maintain the same assumptions. Perhaps the simplest way to proceed is to restrict the scope of the system to inflection. For many Indo-European languages, prefixation occurs only in derivational morphology. Some counterexamples are listed in (2).

2. (a) sehen – gesehen 'see – seen' DE
   (b) lepszy – najlepszy 'better – best' PL

German past participles, as in (2a), and Polish superlatives, as in (2b), are formed by prefixation. They are generally considered to be instances of inflection. A more adequate scope restriction is, therefore, to state suffixation as the definition of the coverage, independently of the distinction between inflection and derivation. In order to include prefixation in the scope, one would have to abolish the correspondence of the starting lexicon with the stem-affix distinction, i.e. to change the answer to (1d).

Turning to (1b), the approach to the distinction between inflection and word formation shows another difference between the two computational models. Whereas both Two-Level Morphology and Word Manager are equipped for the mapping between text words and dictionary words, only Word Manager has a linguistically informed notion of *lexeme* as in Fig.3. This not only means that stems are distinguished from affixes, because they are the base of a lexeme, but also that IRules have an inherently different function to WFRules. IRules realize the inflectional paradigm of a lexeme, whereas WFRules are individual applications of a word formation rule to a lexeme (or two lexemes, in the case compounding) producing a new lexeme and assigning it to the appropriate IRule. No such distinction can be made in Two-Level Morphology, which does not have lexemes as units, but only performs a mapping between a surface word form and a sequence of formatives.

This leaves (1a), the position of morphology in language. Of course, it cannot be expected of a system for morphology that it has a model of language. Moreover, reusability in different applications, which may have different underlying assumptions about the organization of grammar, is a design feature of such systems. Therefore, this is not an issue that can be used to characterize them further.

5 Computational models in linguistic theory

The idea that linguistic theory can benefit from computational modelling is old and persistent. While not directly aimed at computer implementation, Chomsky's work in mathematical linguistics, summarized in Chomsky & Miller [10], Chomsky [9] and Miller & Chomsky [32], results in formal representations of grammars and language users. Bresnan & Kaplan [7] motivate their theoretical approach by the computational properties of the processing model. Barton et al. [5] calculate the computational complexity of grammar formalisms as an argument for their plausibility.

Against this background, it is interesting to consider what the formalisms of Two-Level Morphology and Word Manager may tell us about the nature of derivational morphology. Of course we cannot deduce from the fact that both formalisms are IA that IA is the correct modelling of morphology. Hockett [23] already remarked that any IA account can be transformed into an IP account. Moreover, although both computers and human brains may process language, we cannot be sure to what extent the procedures they use are similar. What we can say, however, is that they process the same data. Therefore, it is especially the problems with encoding data in the computer formalisms that are interesting to consider, because they indicate how the data restrict the choice of an adequate formalism.

In Two-Level Morphology, a design criterion that transpires is simplicity. By adopting finite-state rules throughout, a degree of computational simplicity was achieved which in the 1980s was still relevant. With the emergence of faster computers and larger amounts of memory, the kind of space and time efficiency achieved with finite-state rules lost some of its prominence. As noted in section 4, finite-state morphology is not ideally equipped to deal with prefixation, especially if it cooccurs with suffixation in the same set of rules. For derivational morphology, this is generally the case in Indo-European languages.

Kiraz [27] discusses the application of finite-state morphology to Semitic languages. In Semitic languages, consonantal stem patterns are combined with vocalic and metrical patterns, which may express inflectional and derivational information, to form words. His solution is typical of later developments in finite-state morphology. Rather than considering the formalism as a representation of the language data that is equally adequate for computational and for human processing, the finite-state nature of the rules is used for computational implementation only. Any finite set of data can be modelled into finite-state rules. Such a compilation underlies much of the modern use of finite-state rules in morphology. As a consequence, there is no representation in the computational model of categories such as stem or affix, which are used in linguistic theories. The computational code performs a task, but it is not meant to be interpreted linguistically.

Evaluating Two-Level Morphology from the perspective of theories of derivational morphology, we can conclude that the problems in formulating derivation rules provide further evidence that a finite-state model of the type in Fig.1 is not a likely candidate for modelling the human processing of derivational morphology. This is not surprising as it is in line with various other arguments against finite-state models for human language processing, starting with Chomsky ([8]: 21).

Turning to Word Manager, the aspect that is potentially of most theoretical interest is the Bow Tie Model in Fig.3. In order to understand the background of the Bow Tie Model, it is important to see the original context in which it was proposed. As elaborated in ten Hacken [15], the model was developed in opposition to the model of reusable lexical databases underlying much of the work in Atkins & Zampolli [4] and Walker et al. [35]. The aim of that work was to develop a lexical database that would be theory-neutral. The concept of theory-neutral information is highly problematic. As the study of scientific revolutions and incommensurability shows (e.g. Kuhn [29], Hoyningen-Huene [24], Anderson et al. [1]), being theory-dependent is a necessary property of all scientific concepts, but it is more prominent the further the relevant theory is removed from one's own assumptions.

The contrast between the Bow Tie Model and the model of theory-neutral lexical databases is that in the former, the lexeme functions as a hinge between the reusable morphology and the application-specific other types of information, whereas in the latter, the reusable domain is meant to be all lexical information. Because in the Bow Tie Model, a system for morphological dictionaries covers a functionally coherent domain, the interface along which information should be adapted in order to reuse it in a new application is much smaller than when all lexical items are fully specified. Word Manager covers all morphological information (entries and rules) and does not make any assumptions on other domains. In a theory-neutral lexical database, rules for all domains of processing interact with fully specified lexical entries, so that the interface for adapting the information to the new application runs through all features of each entry.

It is obvious that the kind of reusability that motivates the Bow Tie Model is specific to computational linguistics. It is nevertheless interesting to consider the problems that it raises when it is transplanted into a linguistic theory. In the project described by ten Hacken [16], the Bow Tie Model was applied rigorously in the sense that what counts as a lexeme was determined only by morphological properties. The problems that emerge can be illustrated by the case of *inflame*. There can only be one lexeme *inflame*, because there is no distinction in any inflectional form. It underlies the lexemes in (3).

3. (a) inflammable
   (b) inflammation
   (c) inflammatory

Each of (3) illustrates a regular derivation rule applied to *inflame*, but with different senses as input. These senses are not in the domain of Word Manager, so all of (3a-c) are in the same word formation family. A more striking case involves the words in (4).

4. (a) Angle$_N$ 'member of an ancient Germanic people'
   (b) anglicize 'make English in form or character'
   (c) angle$_N$ 'space between two intersecting lines or surfaces'
   (d) angledozer 'a bulldozer with a blade at an oblique angle'
   (e) angle$_V$ 'to present so as to reflect a particular viewpoint'
   (f) angle$_V$ 'to fish with a hook and bait'
   (g) angler 'a person who fishes with a rod and line'

The words in (4) are accompanied by indicative sense descriptions, which in the Bow Tie Model belong to the right-hand side. For Word Manager, (4a) and (4c) can only be one lexeme, because they have no inflectional differences. Similarly, (4e) and (4f) are one lexeme. Therefore, if (4b) is derived from (4a), it is equally related to (4c), because (4a) and (4c) are the same lexeme. Moreover, if (4e) is derived from (4c), also (4f) is analysed as derived from (4c). This means that all words in (4) are related, even (4g) to (4a), which is clearly not the correct analysis for the mental lexicon.

Examples such as (3) and (4) show the limitations of the Bow Tie Model in the same way as the ones in (2) are problematic for Two-Level Morpology. In such cases, we can either accept that the computational model is not correct for the mental lexicon or find a patch. In Word Manager, arbitrary features were introduced to distinguish lexemes that do not behave as the strict interpretation of the Bow Tie Model would have it.

The interest of interpreting computational models as a model of the mental lexicon resides in the counterintuitive examples we find. The problems with prefixation in Two-Level Morphology show that (at least for Indo-European languages) morphological processing is not strictly linear. The problems with lexemes in Word Manager show that a lexeme is not a purely morphological category.

References

[1] Anderson, Hannel, Barker, Peter and Chen, Xiang. *The Cognitive Structure of Scientific Revolutions*. Cambridge: Cambridge University Press, 2006.

[2] Anderson, Stephen R. *A-Morphous Morphology*. Cambridge: Cambridge University Press, 1992.

[3] Aronoff, Mark H. *Word Formation in Generative Grammar*. Cambridge (Mass.): MIT Press, 1976.

[4] Atkins, B.T.Sue and Zampolli, Antonio (eds.).*Computational Approaches to the Lexicon*, Oxford: Clarendon, 1994.

[5] Barton Jr., G. Edward, Berwick, Robert C., and Ristad, Eric Sven. *Computational Complexity and Natural Language*. Cambridge (Mass.): MIT Press, 1987.

[6] Bloomfield, Leonard. *Language*. London: Allen & Unwin, 1933.

[7] Bresnan, Joan and Kaplan, Ronald M. Introduction: Grammars as Mental Representations of Language, in Bresnan, Joan (ed.), *The Mental Representation of Grammatical Relations*. Cambridge (Mass.): MIT Press, pp.xvi–lii, 1982.

[8] Chomsky, Noam. *Syntactic Structures*. Den Haag: Mouton, 1957.

[9] Chomsky, Noam. Formal properties of grammars. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter (eds.) *Handbook of mathematical psychology*, vol. 2. New York: John Wiley & Sons. pp. 323–418, 1963.

[10] Chomsky, Noam & Miller, George A. Introduction to the Formal Analysis of Natural Languages. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter (eds.) *Handbook of mathematical psychology*, vol. 2. New York: John Wiley & Sons. pp.269–321, 1963.

[11] Chomsky, Noam. Remarks on Nominalization. In Jacobs, Roderick A. & Rosenbaum, Peter S. (eds.), *Readings in English Transformational Grammar*, Waltham (Mass.): Ginn, pp.184-221, 1970.

[12] Di Sciullo, Anna Maria & Williams, Edwin. *On the Definition of Word*. Cambridge (Mass.): MIT Press, 1987.

[13] Domenig, Marc. *Word Manager: A system for the Specification, Use, and Maintenance of Morphological Knowledge*. Habilitationsschrift, Universität Zürich, 1989.

[14] Domenig, Marc & ten Hacken, Pius. *Word Manager: A System for Morphological Dictionaries*. Hildesheim: Olms, 1992.

[15] ten Hacken, Pius. Two Perspectives on the Reusability of Lexical Resources, *McGill Working Papers in Linguistics*, 14, 39–49, 1999.

[16] ten Hacken, Pius. *Word Formation and the Validation of Lexical Resources*. In González Rodríguez, Manuel & Paz Suárez Araujo, Carmen (eds.), *LREC 2002: Third International Conference on Language Resources and Evaluation*, pp.935–942, 2002.

[17] ten Hacken, Pius. *Chomskyan Linguistics and its Competitors*. London: Equinox, 2007.

[18] ten Hacken, Pius. Word Manager. In Mahlow, Cerstin & Piotrowski, Michael (eds.), *State of the Art in Computational Morphology*. Berlin: Springer, pp.88–107, 2009.

[19] ten Hacken, Pius. What is a Dictionary? A View from Chomskyan Linguistics. *International Journal of Lexicography* 22, pp. 399–421, 2009b.

[20] ten Hacken, Pius. Delineating Derivation and Inflection. In Lieber, Rochelle & Štekauer, Pavol (eds.) *The Oxford Handbook of Derivational Morphology*. Oxford: Oxford University Press, pp.10–25, 2014.

[21] ten Hacken, Pius. Translation, Theory, and the History of Machine Translation. In Zybatow, Lew N.; Stauder, Andy & Ustaszewski, Michael (eds.) *Translation Studies and Translation Practice: Proceedings of the 2nd*

International TRANSLATA Conference, 2014, Part I, Frankfurt/M: Lang, pp.13–26, 2017.

[22] Halle, Morris & Marantz, Alec. Distributed Morphology and the Pieces of Inflection. In Hale, Kenneth & Keyser, Samuel J. (eds.) *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. Cambridge (Mass.): MIT Press, pp.111–176, 1993.

[23] Hockett, Charles F. Two Models of Grammatical Description, *Word* 10, pp. 210–231, 1954.

[24] Hoyningen-Huene, Paul. *Die Wissenschaftsphilosophie Thomas S. Kuhns*. Braunschweig: Vieweg, 1989.

[25] Jackendoff, Ray. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press, 2002.

[26] Karttunen, Lauri (ed.). KIMMO: A Two Level Morphological Analyzer. *Texas Linguistic Forum* 22, Department of Linguistics, University of Texas, Austin, pp.163–278, 1983.

[27] Kiraz, George A. *Computational Nonlinear Morphology With Emphasis on Semitic Languages*. Cambridge: Cambridge University Press, 2001.

[28] Koskenniemi, Kimmo. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki, Department of General Linguistics Publications No. 11, 1983.

[29] Kuhn, Thomas S. *The Structure of Scientific Revolutions*, Second Edition, Enlarged. Chicago: University of Chicago Press (orig. 1962), 1970.

[30] Luce, R. Duncan, Bush, Robert R. & Galanter, Eugene (eds.). *Handbook of Mathematical Psychology*. Volume II, New York: Wiley, 1963.

[31] Matthews, Peter H. *Morphology: An Introduction to the Theory of Word Structure*. Cambridge: Cambridge University Press (8th impr. 1989), 1974.

[32] Miller, George A. and Chomsky, Noam. Finitary Models of Language Users. In Luce, R. Duncan, Bush, Robert R. & Galanter, Eugene (eds.) (1963) *Handbook of Mathematical Psychology*. Volume II, New York: Wiley, pp.419–491, 1963.

[33] Pedrazzini, Sandro. *Phrase Manager: A system for Phrasal and Idiomatic Dictionaries*. Hildesheim: Olms, 1994.

[34] Sproat, Richard W. *Morphology and Computation*. Cambridge (Mass.): MIT Press, 1992.

[35] Walker, Donald E., Zampolli, Antonio and Calzolari, Nicoletta (eds.). *Automating the Lexicon: Research and Practice in a Multilingual Environment*. Oxford: Oxford University Press, 1995.

In Trouble with the Rules. Theoretical Issues Raised by the Insertion of -*sc*- Verbs into Word Formation Latin

Marco Budassi and Eleonora Litta
University of Bergamo/Pavia and Università Cattolica del Sacro Cuore, Milan
E-mail: marcobudassi@hotmail.it and e.littamodignani@gmail.com

Abstract

Word Formation Latin (WFL) is a derivational lexicon of Classical Latin that connects lexical items on the basis of word-formation rules (WFRs). This paper describes the process of inserting the class of Latin -*sc*- verbs as a test case for discussing a number of linguistic theory issues arising from pigeonholing such a multiform class of verbs into a model regulated by a strict morphotactic approach. Additionally it discusses the first steps towards the design of a Word and Paradigm model for the representation of derivational families in Latin.

1 Introduction

Word Formation Latin (WFL) is a language resource for Classical Latin that connects lexical items on the basis of word-formation rules (WFRs).[1] The scope of WFL is to assign a WFR to each morphologically-complex lexeme (i.e. one word morphologically derived from another word) and to link each complex lexeme to its ancestor. All those lexemes that share a common (not derived) ancestor belong to the same "word formation family" (Litta et al. [14]). For instance, the noun *bellatrix* 'she who wages war', the verb *rebello* 'to revolt, rebel', and the adjective *bellicosus* 'fond of war' all belong to the word formation family whose ancestor is noun *bellum* 'war'.

The semi-automatic insertion of lemmas into the WFL database establishes input-output relations for a set of lemmas matching the features that characterise each WFR. Occasionally, however, the directional input-output morphotactic approach does not fit certain word formation processes, so alternative solutions or tweaks must be employed.

[1] It is the outcome of a project that has received funding from the European Union's Horizon 2020 Research and Innovation programme under the Marie Sklodowska-Curie grant agreement No 658332-WFL.

The aim of this paper is threefold: a) to account for the insertion of -*sc*- verbs as three different derivation processes (Section 4); b) to point out some problematic cases that have emerged from accommodating a multiform class of verbs to the resource's strict morphotactic approach, and show how these have been dealt with in WFL (Section 5); c) to show how a paradigmatic approach to derivational morphology could solve the conundrums typically raised by the current methodology (Section 6 and Conclusions).

2 Word Formation Latin

WFL records all word formation processes acting in Classical Latin: derivation, which consists in affixation (prefixation/suffixation: e.g. *re-ferio* 'to hit back', *cruciatio* 'torture') and conversion (e.g. *aureus* 'made of gold', adjective > *aureus* 'gold coin', noun), and compounding (e.g. *damnum* + *cupidus* = *damnicupidus* 'harm-loving').

Applying WFRs to lexical data requires that each morphologically-derived lemma be assigned a WFR and paired with its base lemma. WFRs are modelled as directed one-to-many relations between lemmas. These relations are implemented within a relational database and they are enhanced with their attributes (e.g. type of WFR, PoS, affix, etc.). WFL uses a morphotactic approach, where one word formation process is treated individually, and the output of a WFR is usually richer (containing more morphemes) than the input (with the exception of conversion, which only involves a change of PoS). Each output lexeme can only have one source, except in the case of compounds, where it is possible to have two input lexemes for one output lexeme (*nox* + *color* = *nocticolor* 'night-coloured').

From a theoretical point of view, WFL is based on the assumption that WFRs are conceived according to the Item-and-Arrangement (IA) model, which considers word forms either as simple (non-derived) morphemes or as a sequence of morphemes meeting the following conditions: 1) Baudoin's assumption that both base and affixes are lexical elements (i.e. they are both morphemes); 2) they are dualistic: they have both form and meaning (Bloomfield's "sign-base" morpheme theory); 3) they both exist in a lexicon (Bloomfield's "lexical morpheme" theory, see Hockett [12].

IA was chosen as a basic theoretical model for two main reasons: first, because it emphasises the semantic significance of affixal elements as they are found in the lexicon (see, for example, the *Oxford Latin Dictionary* [9]); secondly, IA is the model adopted by other existing derivational lexica (*Word Manager*, Domenig & ten Hacken [6]) after which WFL was designed.

The lexical basis used to compile the resource is that of the morphological analyser for Latin Lemlat, which brings together lemmas from three Classical Latin dictionaries (Georges & Georges [8]; Oxford Latin Dictionary [9]; Gradenwitz [10]) as well as the Onomasticon of Forcellini's ([7]) 5th edition of *Lexicon Totius Latinitatis* (Budassi & Passarotti [2]). It counts 40,014 lexical entries and 43,432 lemmas (as more than one lemma can be included in the same lexical entry), and 26,250 lemmas driven from Forcellini's Onomasticon (Passarotti et al. [15]).

Lemmas are added to WFL in a semi-automatic manner. First, *ad hoc* SQL queries pair candidates that might have undergone a single word formation process (e.g. nouns ending in *-tio* whose base matches that of a corresponding verb); next, a thorough manual check rectifies false positives and duplicates generated by the high number of homographs and identical bases. Candidates that are not found with SQL queries are manually identified.

The WFL lexicon can be accessed online through a visualisation query system at `http://wfl.marginalia.it`. The data is visualised in two different means, resulting from four different ways of browsing the WFRs: a) lists of lemmas matching a query, and b) tree-like graphs representing the derivation cluster of a single lemma. Should the chosen lemma be the root lemma, the derivation cluster corresponds to its word formation family. In the tree-graphs, lemmas are nodes and WFRs are edges.

3 The class of *-sc-* verbs in Latin

The class of *-sc-* verbs is quantitatively broad in Latin, and its semantic history has been much debated (Haverling [11]): traditionally, they are considered dynamic/intransitive counterparts of stative/transitive base verb forms (e.g. *augesco*, 'I grow', intransitive < *augeo*, 'I increase', transitive) (Da Tos [5]), or at least describe the beginning of a situation (e.g. *calesco* = *calere incipio*, 'to become warm') (Viti [17]).

As a matter of fact, Latin *-sc-* verbs are much less homogeneous. In Early as well as Classical Latin, (unprefixed) *-sc-* verbs had a dynamic but non-terminative meaning. They occurred in constructions with expressions of duration (e.g. *duos menses*, 'for two months'), with *dum* when such conjunction meant 'while/for the time that', as well as in constructions with verbs meaning 'to begin' and 'to stop' (e.g. *dum haec silescunt turbae*, 'while these troubles are calming down', 29a: Ter. *Ad*. 785-786). Furthermore, a considerable number of *-sc-* verbs were formed from other transitive or rather frequently stative verbs. In these cases, the *-sc-* suffix had the clear function of indicating transitivity or a dynamic action. In addition to these categories, a small group of *-sc-* verbs were formed from dynamic intransitive verbs; in such cases, the *-sc-* suffix expressed a gradual process (e.g. *aboriscor*, 'I gradually disappear' < *aborior*, 'I die') and the ongoing nature of the action (e.g. *tremisco*, 'I tremble in front of something or someone' < *tremo*, 'I tremble').

All of these semantic distinctions were however almost lost in Late Latin (Haverling [11]), so that the original dynamic value of *-sc-* disappeared and new *-sc-* verbs were formed with a stative meaning (e.g. *lippesco*, 'I have red eyes'; *delitesco*, 'I am hiding'). Moreover, in the last evolutionary stages of Latin, *-sc-* verbs formed on the basis of dynamic verbs did not even differ from these semantically (e.g. *fumo* - *fumesco*, 'I emit smoke').

Prefixes affected the form of action expressed by *-sc-* verbs in several ways. Compare, for example, the verb forms *aresco*, *inaresco* and *exaresco*. *Aresco* is a dynamic atelic verb form ('I become dryer'). The prefixed forms *inaresco* ('I start becoming dry') and *exaresco* ('I dry out'), instead, display an ingressive and a completive meaning respectively, stressing «the initial and the final phrase [...] of the situation» (Viti [17]: 174). Nevertheless, over time this detailed semantic differentiation decayed, to the point that it was basically lost in Late Latin.[2]

To conclude, it is worth-noting how even from a morphological point of view *-sc-* verbs are not uniform, as they are not all the result of a deverbal derivation. In certain cases the ending *-sc-* is already an integral part of old verb roots (e.g. *pasco*, 'I feed', which describes a non-terminative process of eating); in others, *-sc-* verbs are back-formed from participles (e.g. *nascor*, 'I am born' < *natus* 'born'). Furthermore, *-sc-* verbs do not only derive from other verbs, but also from nouns (e.g. *puellasco*, 'I become like a girl' < *puella*, 'girl'), as well as adjectives (e.g. *iuvenesco*, 'to become young' < *iuvenis*, 'young').

4 Inserting *-sc-* verbs into WFL

There are 688 verbs ending in *-sco(r)* in the Lemlat lexical basis. Among these, a group of verbs such as *(g)nosco* 'to get to know', *pasco* 'to feed', *disco* 'to learn' have not been inserted in WFL as a result of a WFR (although they are *-sc-* verbs) since their *-sc-* derivation goes back to pre-attested phases of Latin or in certain cases even to Proto-Indo-European (viz. Haverling [11]). Owing to the fact that they do not have a corresponding base verb to be related to in the resource, they are considered as "roots".

A number of *-sco* ending verbs (327) have been inserted into WFL as the result of prefixation rather than *-sco* suffixation (e.g. *seneo* > *senesco* > *assenesco*, *desenesco*, *insenesco*, etc). This point will be taken into deeper consideration in the next Section.

326 verbs are considered the result of suffixation with *-sco*: 261 derive from other verbs, 38 derive from nouns and 27 derive from adjectives. These numbers include 7 fictional verbs that have been created to account for parasynthetic formations into the WFL morphotactic system: 3 are deverbal, 3 denominal and 1 deadjectival (see below). Their morphological distribution is shown in detail in Table 1,[3] together with an overview of all the derivation patterns of *-sc-* verbs in WFL.

[2] For instance, the difference once existing between non-terminative unprefixed *-sc-* verbs and ingressive/completive prefixed *-sc-* verbs became blurred. If we take into account e.g. *nosco*, *agnosco* and *cognosco*, they can replace one another as they show the stative meaning of *novi* ('I know').

[3] In Table 1, V stands for "verb", N for "noun", and A for "adjective"; numbering indicates conjugation, declension, or adjectival class numbers, DT stands for Derivation Type. V5 are *e/i* verbs ending in *-io* (e.g. *cupio* 'to desire'), and VA are "anomalous verbs" (e.g. *esco* < *sum*, used with future sense).

V-to-V		**N-to-V**		**A-to-V**	
DT	**total**	**DT**	**total**	**DT**	**total**
V1→V3	99	N1→V3	6	A1→V3	17
V2→V3	114	N2→V3	13	A2→V3	10
V3→V3	15	N3→V3	17		
V4→V3	24	N4→V3	1		
V5→V3	8				
	260 (+1 VA = 261)		**37**		**27**

Table 1: Morphological distribution of -*sc*- derivations.

5 Theoretical issues unravelled by -*sc*- verbs

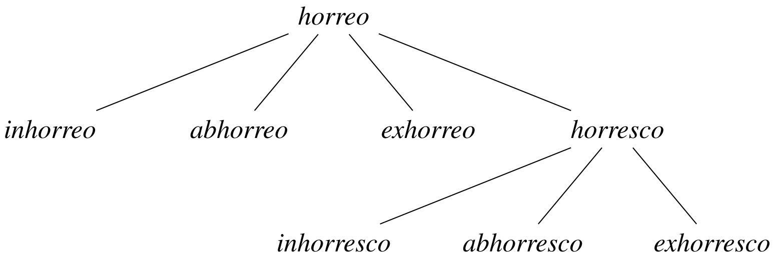
Representing (as well as analysing) word formation processes via directed tree-graphs raises some significant questions, and -*sc*- verbs demonstrate to be a reliable testing ground for investigating such theoretical issues.

The first and most evident problem that comes glaringly to light is how to fit prefixation and/or suffixation within the same derivation process (Crocco Galèas & Iacobini [3]). The derivational family starting from the verb *horreo* (‘to be stiff, to have a dreadful aspect’) perfectly exemplifies this issue. On the basis of this verb, a considerable number of other verbs are derived via prefixation or suffixation (e.g. *exhorreo*, *ex-* + *horreo*, ‘to shudder’; *horresco*, *horreo* + -*sc*-, ‘to bristle’; etc.). In addition, there is a relevant number of verbs that result from the attachment of both a prefix and a suffix to the base verb *horreo* (e.g. *inhorresco*, *in-* + *horreo* + -*sc*-, ‘to stand on hand’).

Although it is clear that e.g. *inhorreo* < *in* + *horreo*, it is uncertain whether e.g. *inhorresco* derives from *in* + *horresco* or rather from *inhorreo* + -*sc*- (that is, *inhorresco* < [*in* + [*horr* + *esc*]] or *inhorresco* < [[*in* + *horr*] + *esc*]). In this and similar cases it is necessary to understand whether prefixation “has happened” before suffixation, or vice versa.

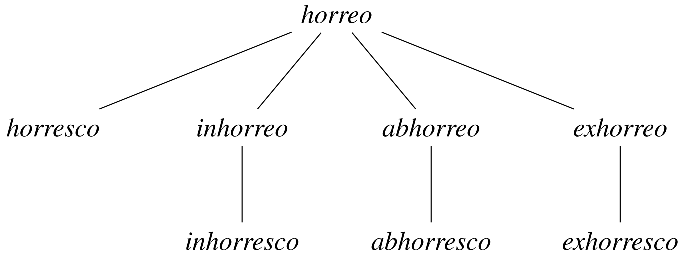
This issue leads to two different scenarios. Let us consider a sub-tree involving *horreo* (‘to be stiff’), *exhorreo* (‘to shudder’), *inhorreo* (‘to stand on hand’), *abhorreo* (‘to shrink back from’), *horresco* (‘to bristle’), *exhorresco* (‘to shudder with fear’), *inhorresco* (‘to become stiffly erect’) and *abhorresco* (‘to become disgusted’). On the one hand (Tree 1),[4] if we prioritise prefixation, the resulting pattern of derivation of verbs with both a prefix and a suffix will pass through the node with the suffixed verb (*horresco*) only, from which such verbs are derived.

[4] Derivational patterns represented in Tree 1 and Tree 2 are not intended to be correct. For example, Haverling [11] states quite clearly that on the basis of both style and textual criticism the derivation *exhorreo*> *exhorresco* is implausible. What Tree 1 and Tree 2 illustrate serves only the purpose of highlighting the problematics of the WFL morphotactic model.



Tree 1: *horreo* morphological sub-family with prioritised prefixation.

On the other hand (Tree 2), if we prioritise suffixation, the resulting pattern of derivation of verbs with both a prefix and a suffix will pass through several nodes with each of these verbs with the prefix only.



Tree 2: *horreo* morphological sub-family with prioritised suffixation.

Additionally, this approach raises a serious problem of coherence: there are cases in which all verbs belonging to the derivational patterns shown above are attested (e.g. *horreo*, *exhorreo*, *exhorresco*), but also cases in which they are not (e.g. *horreo*, *cohorresco*, but **cohorreo*). Similar (frequent) scenarios pose arduous limits to the treatment of derivational patterns via tree-graphs. Indeed, if should one choose to prioritise e.g. suffixation as in Tree 2, so that the derivational pattern of a verb like *abhorresco* would result in [[*ab* + *horr*] + *esc*], instances of verbs that occur with *both* a prefix and a suffix – rather than a prefix only (e.g. *cohorresco*, but **cohorreo*) – would not be coherent with the prioritised suffixation analysis. Remarkably, moving beyond the individual case of -*sc*- verbs, this situation is quite widespread in Latin.

For this reason, one is forced to take a number of decisions and commit to them as formal work policy. The first methodology used in decision making is the assumption that, when in doubt, the *Oxford Latin Dictionary* [9] must be followed as the ultimate reference source. This measure alone can ensure consistency throughout the lexicon.

Another phenomenon highlighting problems with WFL's approach is the presence of a few -*sc*- verbs that resemble parasynthetic formations, as these seem to have been formed through the simultaneous addition of the -*sc*- suffix and a prefix. See for example *decaulesco* 'to form a stem', which, according to the *Oxford Latin Dictionary*, is formed as *de* + *caulis* + *e-sco*. The project's morphotactic approach imposes the creation of a fictional lexeme **caulesco* in the lexical basis of WFL in order to fill a gap in the derivation tree. **Caulesco* is not expected to have ever existed, but only acts as a *trait d'union* between *caulis* 'stem', and *decaulesco*, hence accounting for two formative processes instead of one: *decaulesco* < [*de* + [*caul* + *esc*]]. The same method has been applied to fulfill other parasynthetic derivations in WFL.

Backformation is also worth a mention. Backformation is a derivation process that involves the removal of an affix (or a supposed one), so that a new word is created by analogy with similar looking existing ones (see for example the English *addict* from *addicted*, or *to diagnose* from *diagnosis*).

As far as -*sc*- verbs are concerned, it is commonly thought (Haverling [11] and Oxford Latin Dictionary [9]), for instance, that *irascor* 'to be angry' derives from the adjective *iratus* 'angry' (*ira-tus* > *ira-scor*). The same has happened with *nascor* 'to be born (begin life)' (< *natus*), *proficiscor* 'to set out' (< *profectus*). At the time of writing, these and other backformation processes are not marked in WFL, but are portrayed incoherently as follows: *iratus* > *irascor* (A-To-V -sc-); *(g)nascor* = root verb (i.e. not derived), origin of all other derivations;[5] *proficio* > *proficiscor*.[6]

If many backformation processes could be represented in WFL by marking a directional edge so that it appears in a specific color, or indeed shows a different direction, the above examples seem to be of difficult representation: the participles (or adjectives resembling them), from which the -*sc*- verb is supposed to have originated by analogy, are not included in the WFL lexical basis. The corresponding verb would in any case be shown as the base of the derivation.

In addition to all these points, the automatic process employed to insert input-output relationships in WFL cannot account for morphotactically-obscure derivations, which means that the morphotactic and/or semantic relationship between base and derived lexemes is not evident. This means that these relationships have to be established manually: for example, *opulesco* 'to grow richer', is reported by the *Oxford Latin Dictionary* as coming from *opulens* (or *opulentus*) by inferring the segment *opul*- and attaching -*e-sco* supposedly by analogy with similar formations. It has been inserted in WFL as the result of deadjectival formation from *opulens*; in the same way, *seresco* 'to grow dry', which is seen both by the *Oxford Latin Dictionary* and Haverling [11] to be derived from the adjective *serenus* 'clear (of the sky)' by attaching the -*e-sco* to the segment *ser*-, has also been inserted manually thus confirming this hypothesis.

[5] Notice here that the Lemlat lexical basis does not include participles used as adjectives, as they are considered part of the verbal paradigm. There would be, in any case, no adjective *natus* to link to *(g)nascor* in either direction.

[6] Haverling [11] derives *proficiscor* from the adjective *profectus* and points out how *proficio* is attested from later times. In WFL however, seeing as there is no adjective *profectus* for the reasons explained in the previous note, *proficiscor* is marked as deriving from *proficio* through suffixation with -*sc*-.

6 Derivational paradigms

Given the examples above, it seems that the current model employed in WFL does not always serve well to represent all derivation processes coherently. The recent emergence of interest in the application of Word and Paradigm (WP) models to derivational morphology led us to explore their potential in explaining those processes. The main advantage offered by paradigmatic models is that relations between words are not limited to base-derivative pairs, and that they may be oriented both ways or have an unspecified direction (Jackendoff [13]).

According to Štekauer, «potentiality is a crucial term for the concept of derivational paradigm», which means that in those cases where there is a gap to fill (e.g. compare *equalis* > *equalitas* > *equaliter* with Ø > Ø > *totaliter*, Cuzzolin [4]), a paradigmatic approach to word formation «guarantees a high level of predictability and regularity [...] in the sense that existing gaps in the system can be filled anytime with actual words» (Štekauer [16]: 369), or not. Following these concepts, all the examples shown above involving prefixation, parasynthesis, backformation and obscure morphotactics can be explained as having occurred independently within their own derivational family and without the need for directionality. However, WP applications to derivational morphology are still in a gestational stage and concrete models applicable to all languages have yet to be proposed. The following is an attempt to rework the problematic examples outlined above into derivational paradigms.[7]

In Section 5 we saw how the issue of prefixation vs. suffixation is difficult to represent in directed tree-graphs. The necessity of a strict morphotactic approach, and coherence in taking decisions cause noticeably often incorrect representations of Latin derivations (e.g. *exhorreo* > *excorresco*, which is philologically inaccurate, is required by the fact that in other cases we have e.g. *abhorreo* > *abhorresco*). The WP model seems more fitting to describe this and similar cases.

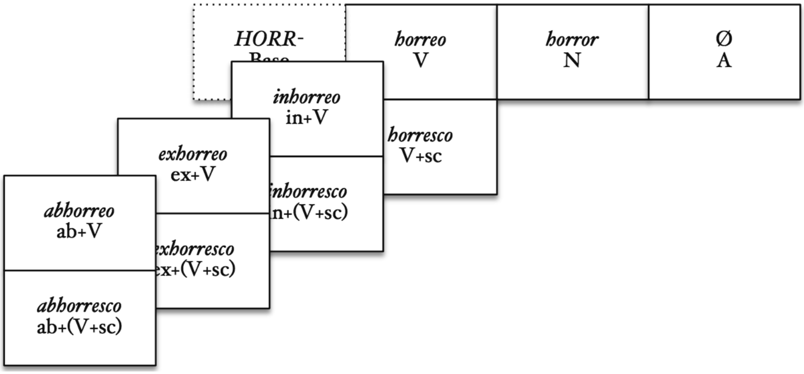


Figure 1: WP representation of the *HORR*- morphological sub-family.

[7] The following figures are inspired by a research paper by Beecher and al. [1]. They do not include yet, at this preliminary stage, any reference to a meaning/form correlation.

Figure 1 proposes a depiction of part of the *horreo* derivational family. The cell labelled ‘Base’ (dotted line) displays the fundamental lexical morpheme which does not change throughout the paradigm. Horizontally, the paradigm relying on the pure stem of *horreo* is reported. Vertically, the base verb *horreo* holds a relation with *horresco*, which is built by adding the -*sc*- suffix to the base verb form. Prefixed forms of *horreo* are placed perpendicularly to *horreo*, prefixed forms of *horresco* are perpendicular to *horresco*. This three-dimensional orientation allows to display prefixed forms with no assumption on which form was created first: the relationship existing between *horreo* and e.g. *inhorreo* or *exhorreo* on the one hand, *horresco* and e.g. *inhorresco* or *exhorresco* on the other, is indeed paradigmatic in nature.

This perspective allows one to better adhere to historical evidence, avoiding the restrictions given by morphotactics or coherence, and cases such as the later attestation of *exhorreo*, rather than *exhorresco*, are better sustained by this paradigmatic view of derivation. If we imagine a stage of Latin in which the box in Figure 1 containing *exhorreo* were blank, the presence of other prefixed -*sc*- verbs such as *inhorresco* or *abhorresco* would justify the creation of *exhorresco* without the need for an input *exhorreo*.[8] The fact that *exhorreo*, then, started to be used, filling an empty box in the paradigm (consider e.g. the pairs *inhorreo* - *inhorresco* or *abhorreo* - *abhorresco*), goes in the same direction.

In a similar way, WP can explain cases of parasynthesis. Figure 2 offers a paradigmatic representation of *decaulesco*. The paradigm of *caulis* simply features blank spaces in ‘V’ and ‘V+sc’ boxes, indicating that *decaulesco* was created naturally by analogy with other prefixed -*sc*- verbs. Once again, the paradigm avoids misunderstandings and wrong analyses of derivational patterns due to oriented tree-graphs and fictional lemmas.

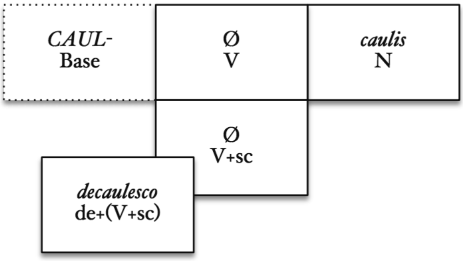


Figure 2: WP representation of *CAUL*-.

Figure 3 offers a WP view of part of the morphological family of *ira*. As stressed in Section 5, *irascor* is generally considered to come from *iratus* via

[8] A purely mechanical perspective on derivational morphology is nevertheless not sufficient. Semantics plays a primary role and must not be neglected. In our example, for instance, some blank boxes in paradigms may be due to semantic constraints: in Latin we have *cohorresco* but we do not have **cohorreo*. This is perhaps due to the perfective value of the prefix *co*-. Since *horreo* has stative meaning, it is not compatible with the perfectivising prefix *co*-. On the contrary, *horresco*, which is not stative, can be prefixed with *co*-.

backformation. This is mostly proved by diachronic and literary evidence. As far as WFL is concerned, a derivation like *iratus* > *irascor* is not entirely satisfying due to the lack of a clear morphotactic pattern. With WP, even without positing the existence of a verb **iro*, or even the non-existence but the necessity of a fictional **iro*, the attestation of *irascor* does not require forced directionality.

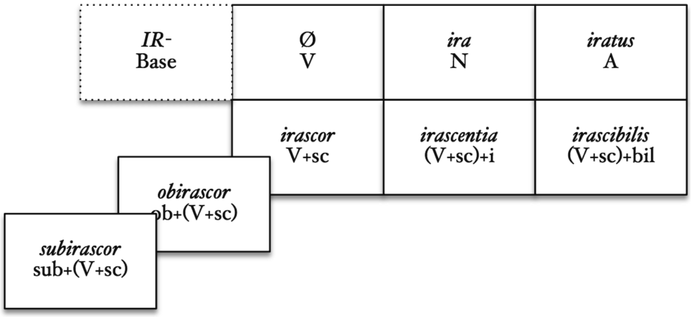


Figure 3: WP representation of *IR*- morphological sub-family.

Cases of obscure morphotactics can equally be better rendered through WP (Figure 4). For instance, the pattern of *opulens* > *opulesco*, was chosen in WFL because the *Oxford Latin Dictionary* proposes *opulesco* < **opul*- < *opulens*. Noticeably, what the *Oxford Latin Dictionary* posits is very close to a paradigmatic approach to the derivational morphology of these forms. The *Oxford Latin Dictionary* goes back to a form **opul*-, to which *e-sco* would have been added. This **opul*-, however, is exactly the so-called ‘Base’, which we consider the basis of the main paradigmatic derivations of a morphological family. The positing of such a lexical morpheme can explain why we have some forms and not others (which would be mandatory with a different, linear and oriented perspective on derivational morphology). The fact that even in the *Oxford Latin Dictionary*, which does not have any theoretical pretension but is based on pure philological, diachronic as well as phonological data, a base form for paradigmatic derivation comes to light, deserves attention.

7 Conclusions and Future Work

In this paper, we delved into some theoretical issues behind Word Formation Latin, bound by a strict morphotactic rule based on the Item and Arrangement model of grammatical description. The morphotactic approach, which considers word formation as a transformation from an input into an output word, has the clear merit of allowing one to compute a series of large-scale studies on productivity that would not be so easily performed otherwise. Nevertheless, considering word formation as exclusively made of input-output relations,[9] puts us at risk of misunderstanding the

[9] Although it is made clear that the WFL derivational trees do not contain any diachronic statement, feedback from users continues to confirm that an exclusively tree-based view is misunderstood for a step by step transformation within a time-frame.

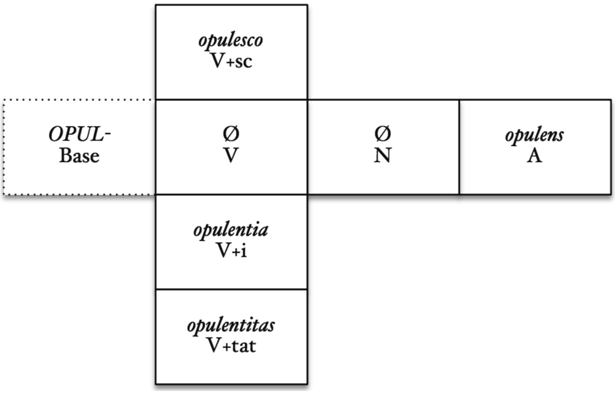


Figure 4: WP representation of *OPUL*- morphological sub-family.

role that analogy plays in the creation of new words. In this sense, a paradigmatic analysis of derivational families could offer a better view on word formation, which exemplifies that not all words have been created in a linear process. To conclude, we expect to evaluate the possibility of creating a second version of WFL to display all word formation families in a paradigmatic way, starting from the preliminary attempts shown above. This would, on the one hand, allow for a more coherent and unvarying analysis of word formation patterns, and, on the other, give way to a new generation of studies on Latin derivational morphology.

References

[1] Beecher, Henry, Ackerman, Farrell, Rose, Sharon, and Barker, Chris. Derivational paradigm in word formation. Research Paper, 2004.

[2] Budassi, Marco and Passarotti, Marco. Nomen omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, pp. 90–94, Berlin: The Association for Computational Linguistics, 2016.

[3] Crocco Galèas, Grazia and Iacobini, Claudio. Lo sviluppo del tipo verbale parasintetico in latino: i prefissi *ad*-, *in*-, *ex*-. *Quaderni Patavini di Linguistica*, pp. 31–68, Padova: Unipress UP, 1993.

[4] Cuzzolin, Pierluigi. L'espressione della totalità in latino. In Alberto Manco (ed.) *L'espressione linguistica della totalità*. Quaderni di AION, Napoli, 2014.

[5] Da Tos, Martina. The Italian finire-type verbs: a case of morphemic attraction. In Silvio Cruschina, Martin Maiden and John Charles Smith (eds.) *The Boundaries of Pure Morphology*. Oxford University Press, pp. 45–67, 2013.

[6] Domenig, Marc & ten Hacken, Pius. Word Manager: A System for Morphological Dictionaries, Hildesheim: Olms, 1992.

[7] Forcellini, Egidio *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Corradini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos*. Perin. Typis Seminarii, Padova, 1940.

[8] Georges, Karl Ernst and Georges, Heinrich. *Ausführliches Lateinisch-Deutsches Handwörterbuch*. Hannover: Hahn, 1913-1918.

[9] Glare, Peter G.W. *Oxford Latin Dictionary*. Oxford University Press, 1982.

[10] Gradenwitz, Otto *Laterali Vocum Latinarum*. Leipzig: Hirzel, 1904.

[11] Haverling, Gerd. *On sco-verbs, prefixes and semantic functions. A study in the development of prefixed and unprefixed verbs from Early to Late Latin*. Göteborg University Press, 2000.

[12] Hockett, Charles F. Two Models of Grammatical Description. *Words*, 10: pp. 210–231, 1954.

[13] Jackendoff, Ray. Morphological and Semantic Regularities in the Lexicon. *Language*, 51:3, pp. 639–671, 1975.

[14] Litta, Eleonora, Passarotti, Marco and Culy, Chris. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC–it 2016)*. Napoli, aAccademia University Press, pp. 185–189, 2016.

[15] Passarotti Marco, Budassi, Marco, Litta, Eleonora and Ruffolo, Paolo. The Lemlat 3.0 Package for Morphological Analysis of Latin, in Bouma, Gerlof and Adesam, Yvonne (eds.), Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. 22nd May 2017 Gothenburg, Northern European Association for Language Technology (NEALT) Proceedings Series, Vol. 32. Linköpings Universitet: Linköping University Electronic Press, pp. 24–31, 2017.

[16] Štekauer, Pavol. Derivational Paradigms, in Lieber, Rochelle and Štekauer, Pavol (eds.) *The Oxford Handbook of Derivational Morphology*. Oxford: OUP, 2014.

[17] Viti, Carlotta. The use of frequentative verbs in Early Latin. In Haverling, Gerd (ed.) *Latin linguistics in the early 21st century: acts of the 16th International Colloquium on Latin Linguistics*. Uppsala, Uppsala Universitet: pp. 170–182, 2015.

Expansion of the Derivational Database for Croatian

Matea Filko and Krešimir Šojat

Department of Linguistics
University of Zagreb
E-mail: matea.filko@ffzg.hr, ksojat@ffzg.hr

Expansion of the Derivational Database for Croatian

Abstract

In this paper we present CroDeriV – a large morphological database for Croatian. Croatian is a Slavic language with rich morphology and numerous derivational processes. A derivational database consisting of morphologically analyzed lexemes which are connected into derivational families via shared roots is an essential language resource. So far, the derivational database of Croatian consisted solely of verbs. Here we will present its expansion with adjectives.

1 Introduction

This paper deals with strategies for building and organizing a large morphological database for Croatian. It is a South Slavic language with very rich morphology, both in terms of inflection and derivation. As in many Slavic languages, allomorphy of lexical and derivational morphemes at the morpheme boundaries, mainly triggered by phonological rules, is very frequent.

Computational processing of Croatian morphology was so far oriented predominantly towards inflection, since it plays an important role in NLP tasks such as lemmatization, POS and morphosyntactic (MSD) tagging etc. There are two publicly available large inflectional lexica for Croatian – Croatian Morphological Lexicon (CML) [10] and hrLex [2]. CML contains complete inflectional data for ca 125.000 lemmas, and can be used as a lemmatizer and a word-form generator. hrLex consists of ca 100.000 lemmas with almost 5 million token, lemma, MSD triples and it is used for MSD tagging. However, large-scale processing of derivational phenomena has not taken place until recently.

In this paper we present the current shape of the CroDeriV, the derivational database for Croatian, and its expansion to other POS, namely adjectives. The paper is structured as follows: In Section 2 we briefly describe major derivational processes in

Croatian. Section 3 presents the structure of the derivational database for Croatian and gives an overview of underlying principles. In Section 4 the expansion of the database is discussed. The paper concludes with an outline of the future work.

2 Croatian derivational morphology

Croatian morphology can be divided into inflectional [1] and word-formational morphology. Major word-formation processes in Croatian comprise derivation and compounding. The main difference between them is that derivatives and base words have one identical root, whereas compounds include two, or in some cases even six roots.[2] Further in this paper we focus on derivation.

The most productive derivational processes are 1) **suffixation**, especially for the formation of nouns and adjectives, 2) **prefixation**, particularly important in verbal derivation and 3) relatively rare **simultaneous suffixation and prefixation** [3]. Derivational processes also include **conversion** (e.g. derivation of nouns from adjectives without affixation) and **back-formation** (removing of suffixes). Derivational affixes in Croatian can be POS-changing (*voziti* 'to drive' + *-ač* > *vozač* 'driver') and POS-maintaining (*pre-* + *voziti* 'to drive' > *prevoziti* 'to transport').

Derivatives are derived from stems, which can be either bare roots or already derived forms. Generally, the group of derivatives from bare roots is significantly smaller than the group of derivatives with more complex stems. Derivationally connected words constitute derivational families, in some cases comprising more than 170 different lexemes (cf. [9]). Members of the derivational families belong to all major parts of speech. Derivational family structured around the root *glas* 'voice, vote' is shown in Figure 1.

Babić ([1]: 18) lists 771 suffixes and 77 prefixes used in Croatian word-formation processes. As numbers indicate, suffixation is the most productive process. In more detail, there are 526 nominal, 160 adjectival, 61 verbal, 24 adverbial suffixes. Every affix is treated individually. In other words, these are all different suffixes,

[1] Inflectional data is already covered by HML and hrLex (cf. Introduction), and we do not deal with paradigms and inflectional patterns in our database. The only inflectional data we deal with is the final affix which is interchangeable with other inflectional endings as required by morphosyntactic structure of sentences. For example, infinitives are always marked by an ending *-ti* (e.g. *čitati* 'to read'), whereas e.g. 1st person singular present indicative is usually marked by the ending *-m* (e.g. *čitam* 'I read').

[2] In Croatian, suppletion exists only within the inflectional paradigm (and even there is very rare, cf. [5]), there is no suppletion between base and its derivative and there are no suppletive affixes (cf. [6] : 74, for more extensive account of suppletion in Croatian cf. [4]). However, one morph can have several allomorphs. One allomorph is taken to be representative for all other allomorphs. In our database, all allomorphs are connected to the representative morph.

[3] Simultaneous suffixation and prefixation is not as same as the circumfixation in the traditional sense, where circumfix is one affix consisting of two parts. Suffixes and prefixes in the simultaneous suffixation and prefixation are independent units since they can be attached to stems individually, preserving the same meaning. E.g. *bez-* + *nad(a)* 'hope' + *-an* > *beznadan* 'hopeless' (**nadan* or **beznad* are not words in Croatian) vs. 1) *osjećaj* 'feeling' + *-an* > *osjećajan* 'sensitive'; 2) *bez-* + *osjećajan* 'insensitive'.

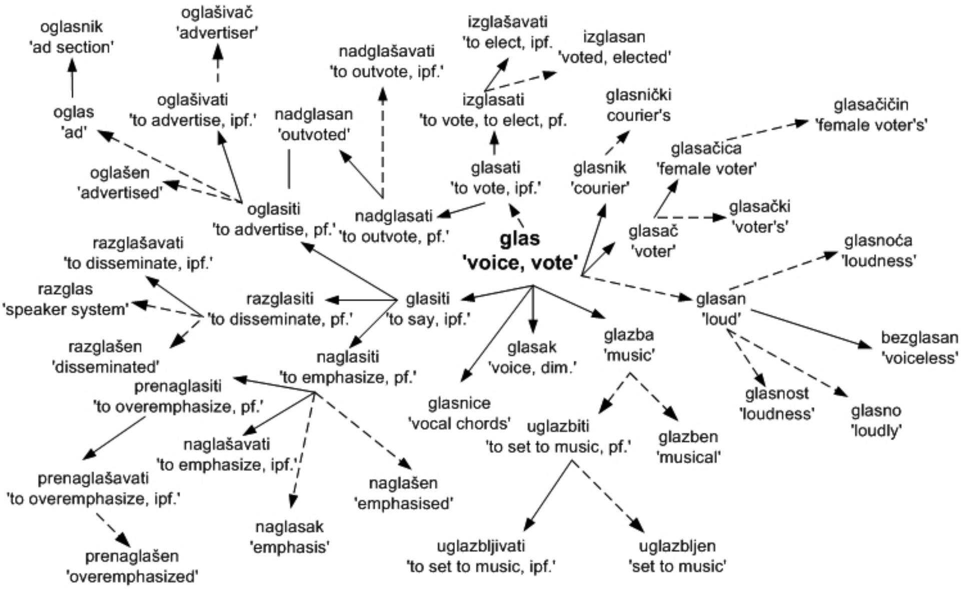


Figure 1: Derivational family of the root *glas**, not all members are included. Dashed line = POS-changing process; full line = POS-maintaining process.

although their meanings or functions can overlap (e.g., there are several suffixes used for the formation of possessive adjectives).

Such a diversity of affixes and their complex combinations raise several questions concerning the building and the design of a derivational lexicon. In the next section we discuss these issues.

3 CroDeriV - derivational lexicon for Croatian

As mentioned, CroDeriV [14][4] is the only language resource for Croatian dealing with derivational morphology. It currently consists of approximately 14.500 Croatian verbs, both simple and derived. Lemmas were collected from free corpora and digital dictionaries. Simple verbs are those that consist of one root, one suffix used for the formation of verbs and infinitive ending *-ti*. This group of verbs is derived from bare roots (e.g. *pis-a-ti* 'to write, imperfective'). The other group refers to those derived either from other verbs (e.g. *pisati – na-pisati* 'to write, perfective') or from other parts of speech (e.g. *ribar* 'fisherman' – *ribariti* 'to fish').

The morphological analysis of data in CroDeriV consisted of two steps. In the first step all verbs, i.e. the verbs from both groups, were fully segmented into morphemes (e.g. *rib-ar-i-ti* is analyzed as *rib* – root, *ar* – masculine nomen agentis, *i* – verbal suffix, *ti* - infinitive ending). In the second step stems used for formation of various derivatives were marked (e.g. *ribar* is used as the stem for the derivation

[4] Search interface is available at croderiv.ffzg.hr.

of the verb *ribar-iti*, the adjective *ribar-ski* 'fisherman's' etc.).

As stated above, Croatian verbs are derived from other POS, e.g. from nouns or adjectives, by suffixation, or from other verbs by prefixation and suffixation. In majority of cases, verbs are derived from verbal stems by prefixation. In terms of verbal aspect, prefixation almost exclusively results in perfective verbs (e.g. *pisati* 'to write, imperfective' > *do-* + *pisati* > *dopisati* 'to add by writing, perfective'). Prefixation is frequently followed by imperfectivizing suffixation (e.g. *dopis(ati)* + *-ivati* > *dopisivati* 'to add by writing, perfective'). Verbal derivation is recursive and verbal derivatives frequently occur with two, three and even four prefixes.

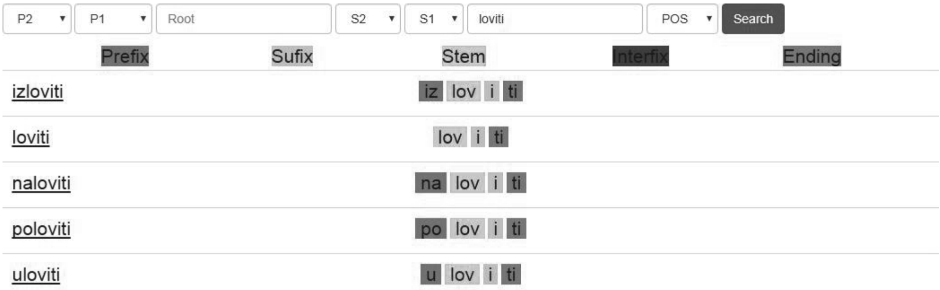
We chose this POS in the initial building stages primarily for two reasons: we wanted to test 1) to what extent an automatic rule-based approach can be used for the recognition of derivationally connected lemmas, and 2) to what extent a derivational process, i.e. the affixes that are used, can be recognized. Since the results were rather unsatisfactory due to extensive allomorphy and homography of roots as well as affixes, all the results were manually checked. Simultaneously, each lemma was segmented into lexical and derivational morphemes, and all allomorphs were linked to a single representative morpheme. During this process we payed attention to semantic disambiguation of homographic forms. Although both verbs *ribariti* 'to fish' and *ribati* 'to scrub' contain the root *rib*, we distinguish between such homographic forms and mark them as *rib1*, *rib2* etc. All derived forms are linked accordingly. In the future development of CroDeriV we intend to provide a short meaning definition for each lemma in the database.

All derivationally connected words were mutually linked via shared roots. Such a procedure enables the detection of full derivational spans of verbs derived from other verbs. At the same time it provides full information about affixes used in verbal derivation, their combinations and productivity. By productivity we consider the number of lemmas formed by means of the same affix. In other words, the greater the number of lemmas in which particular affix occurs, the greater the productivity of this affix.[5]

The search interface of this publicly available database enables a wide range of queries (cf. Figure 2). It can be searched for particular roots, derivational morphemes, suffixal and/or prefixal combinations etc.

From the theoretical point of view, the complete morphological analysis of verbs in infinitive form enabled us to establish the general morphological structure of Croatian verbs. In our approach, it contains four types of slots for different morphemes: (1) the **prefixal part** consists of four slots for derivational prefixes, (2) the **lexical part** includes three slots. In the majority of cases only one is filled, whereas the additional two are provided for verbal compounds of two roots and an interfix, (3) the **suffixal part** contains three slots for derivational suffixes and, finally, (4) an **inflectional slot** for an infinitive ending. Generally, each Croatian verb consists of at least one lexical morpheme, two derivational suffixes and an inflectional ending. The morphological structure of 11 morpheme slots can accommodate all

[5] For the exact numbers of affixal productivity defined in this sense cf. [14].



izloviti	iz lov i ti
loviti	lov i ti
naloviti	na lov i ti
poloviti	po lov i ti
uloviti	u lov i ti

Figure 2: CroDeriV search interface

the recorded combinations of morphemes for Croatian verbs. The lexical entry for each verb also includes the information about verbal aspect and reflexivity. All 11 slots are actually never simultaneously occupied, i.e. such morphologically complex verbs do not exist in Croatian. However, the provided structure is flexible enough for all the verbs recorded in the database. This structure of the database also enables further expansion with other POS.

Next objectives in this project are: 1) to expand the existing derivational families with other POS, and 2) to introduce new derivational families, i.e. families built around roots yet not recorded. Currently, the derivational database is being expanded with nouns and adjectives. Further in this paper we focus on adjectives.

4 Expansion of the database

4.1 Derivation of adjectives

The expansion of the database in each step requires preprocessing and analysis of data. Based on their semantics, Croatian adjectives are generally divided into descriptive (qualitative) and relational adjectives.[6] In terms of morphological features, only descriptives can be compared and distinguished as definite and indefinite. Both groups are marked by typical suffixes: *-an, -at, -iv, -ast* for descriptives; *-ski, -ov, -in, -ji, -ni* for possessives.

In the first part of our experiment we have extracted a list of 1.000 most frequent adjectives from the Croatian frequency dictionary [8] in order to get an initial overview of their morphological and derivational properties. The analysis of this initial set of adjectives reveals that there are 164 adjectives that appear as bare roots, out of which 134 are used as stems for the derivation of verbs.[7] The small

[6] For a more elaborated description of Croatian adjectives cf. [3]. Although modern approaches (e.g. [1], [12], [11]) divide Croatian adjectives into two groups: descriptive and relational adjectives (possessive being the subgroup of relational adjectives), more traditional approaches divide them into three groups: descriptive, possessive and constructive (e.g. *drven* ’wooden’), for more information about different categorizations of Croatian adjectives cf. [7].

[7] Adjectives that are used as stems for the derivation of verbs are already morphologically analyzed and recorded in our database (otherwise the complete morphological analysis of verbs would

subgroup of 21 adjectives appears in the form of inseparable units, i.e. synchronically, they cannot be further segmented into morphemes, e.g. *širok* 'broad' cannot be divided into *šir* as root and *-ok* as suffix[8]. However, *šir-* is used as a stem in the derivation of the verb *širiti* (*šir-iti* 'to make broad') and the noun *širina* (*šir-ina* 'width').

The second group consists of 159 adjectives derived from verbs.

This analysis revealed that adjectives are predominately derived from nouns. The most productively used process is suffixation. Among 1.000 most frequent adjectives, 456 of them are derived from nominal stems, the suffix *-ski* and its allomorphs being the most frequent. Twelve adjectives in this group are derived through simultaneous prefixation and suffixation, e.g. *privremen* (temporary) (*pri-vrem-en*: *pri-* = prefix, *vrem** = root, *-en* = suffix).

48 adjectives are derived from other adjectives. The most productive affix is the prefix *ne-*, used for the formation of antonyms. 55 adjectives are of foreign origin (e.g. *aktivan* 'active'), whereas 31 adjectives are compounds.

Generally, the numbers indicate that adjectives are predominantly derived from nouns through suffixation and their semantics can be mostly predicted on the basis of their suffixes. The number of adjectives used as stems for the verbal derivation is relatively small when compared to verbal stems used for the derivation of adjectives. Adjectival suffixes could be detached from stems in a rule based procedure. The recognition of adjectives as descriptive or possessive can be relatively straightforward if typical suffixes and their allomorphs are previously recognized. As far as the complete morphological analysis is concerned, i.e. the segmentation into all morphemes that make up a particular adjective, the situation is more complex. We deal with this issue in the next section.

4.2 Methodology

Croatian lexicon consists of extensive derivational families. Derivational families can vary from those with two or three members to those composed of several dozens of members of different POS (cf. Figure 1). The general purpose of the derivational database described here is to enable the detection of full derivational families in Croatian, along with the word-formation path from the root to derivatives in final stages of derivational processes. A complete morphological analysis of each lexeme in the database, as described above for verbs, should enable the detection of a generalized morphological structures for all major POS in Croatian. As mentioned, verbs derived from other POS are already included in CroDeriV. Nominal, adjectival, adverbial etc. parts of such verbs are morphologically ana-

not be possible), but our search interface does not support queries via POS of the stem. However, we plan to facilitate such queries, as well as of tracking complete word-formational paths of derived words, e.g. *oglas* 'ad' < *oglasiti* 'to ad' < *glasiti* 'to sound' < *glas* 'voice'.

[8] Suffix *-ok* is an old Slavic suffix which is no longer used in the word-formation of Croatian adjectives. Still, it exists in several very frequent Croatian adjectives, e.g. *širok* 'wide', *dubok* 'deep', *žestok* 'fierce'. The same holds for the suffix *-ak* (e.g. *kratak* 'short', *plitak* 'shallow').

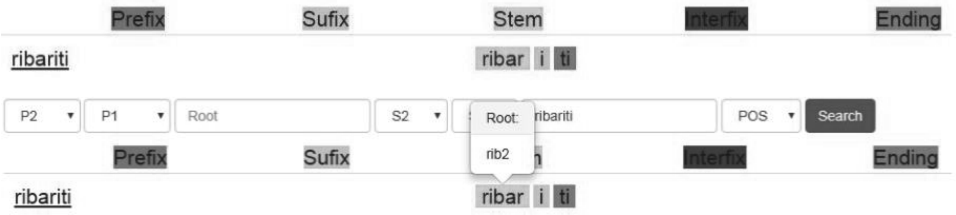


Figure 3: CroDeriV search interface – *ribariti* ’to fish’

lyzed, but this information is yet not displayed publicly. The representation of the verb *ribariti* ’to fish’ in the search interface, containing the whole stem *ribar*, but also the information about the root *rib2*, is given in Figure 3. The complete word-formational path of the verb *ribariti* ’to fish’ is as follows: *rib(a)* ’fish’ + *-ar* > *ribar* ’fisherman’ + *-iti* > *ribariti* ’to fish’. Full morphological analysis of the stem, currently encoded in the background, will be used in the expansion of the database with other POS. The design of the search interface will be therfore adjusted in order to a) fit the generalized morphological structure of other POS, b) include the link to previous steps in word-formation paths. For example, *ribariti* ’to fish’ will be linked to *ribar* ’fisherman’ and *ribar* will be in turn linked to *riba* ’fish’, whereas *ribarski* ’fisherman’s will be linked to *ribar* ’fisherman’.

The expansion of the derivational database with adjectives consisted of several steps. In the first step we have collected adjectival lemmas from available corpora and on-line dictionaries for Croatian. We have thus obtained more than 10.500 adjectives for morphological analysis. Then, we have randomly extracted 1.000 adjectives for manual morphological analysis. All analyzed adjectives were in Nominative singular, masculine gender. The goal of this analysis was to determine whether it is possible to establish a general morphological structure that could accommodate all adjectives in the database and whether it is possible to speed up this process with a rule-based automatic procedure. The analysis takes into account:
1) complete morphological analysis
e.g. *istražiteljski* ’investigative’ is_traž*i+telj+sk+i?
where _ = prefix, * = root, + = derivational suffix, ? = inflectional ending[9]
2) word-formation pattern
e.g. *istražiteljski* ’investigative’ < *istražitelj* ’investigator’ + *-ski*
3) root = *traž**
4) corresponding root in the derivational lexicon (if it exists) = *traž**; this allows us to link the adjective to the existing verbal derivational families
5) allomorph and morph of the stem
e.g. *kritički* (**krit*ič+**k+i?) ’critical’ = **kritik**(a) ’critique’ + *-ski*; allomorph of the stem = *kritič*; morph of the stem = *kritik*

[9] CML, the inflectional lexicon for Croatian [10] can generate all possible inflectional forms of a particular lemma, thus we decided to use lemmas as entries in our database.

6) for compounds only, words that served as the basis for compounding[10]
e.g. *krvožilni* 'circulatory, cardiovascular' < *krv* 'blood'; *žila* 'blood vessel'
7) POS of the base word in the respective word-formation process
e.g. *istražiteljski* 'investigative' = *istražitelj* 'investigator' + *-ski*; POS = N (noun).
This line of morphological processing should enable us to detect adjectives of the same root and, at the same time, adjectives that share roots with the verbs from the existing derivational lexicon. It should also help us expand the derivational families in the derivational database and track all derivational steps in the word-formation path from the root to the final lexeme. The obtained patterns used in the word-formation of adjectives are discussed in the following subsection.

4.3 Adjectival patterns

4.3.1 Word-formation analysis

In the material analyzed so far (1.000 adjectives), the most frequent suffixes are *-ski* (ca 36 %), *-ni* (ca 18 %), *-an* (ca 10 %), and *-ički* (ca 5 %).[11]
Suffix *-ski* undergoes different phonological changes at the morpheme boundaries. It has several possible allomorphs: *-ski* (*škol**ski*** 'school'), *-ški* (*vite**ški*** 'knightly'), *-ki* (*mučeni**čki*** 'martyr's'), which makes the automatic morphological processing difficult and inaccurate. Suffix *-ski* has the most general adjectival meaning of all the suffixes used in the derivation of adjectives. Its meaning can be paraphrased as 'belonging to, related to' and it can be attached to a wide range of nominal bases. Adjectives formed with suffix *-ski* denote relations to plurality or to indefinite individuals, e.g. *sestrin**ski*** 'sisterly'. They differ from the adjectives formed with suffixes *-in, -ov, -ev,* which generally denote the relationship to the specific individual, e. g. *sestr**in*** 'sister's' [7]. Suffix *-ski* is usually attached to animate basis, while the second most frequent suffix *-ni* is generally attached to inanimate basis, cf. [12].
Suffixes *-ni* and *-an* have the same meaning, differing only in the grammatical category of (in)definiteness. The suffix *-an* is used to form indefinite adjectives and the suffix *-ni* to form definite ones. Their meaning can be roughly paraphrased as 'having the property of [meaning of the base]', and they are attached to bases of different POS (e.g. *cvijet* 'flower' + *-an* > *cvjetan* 'floral', *isprav(iti)* 'to correct' + *-an* > *ispravan* 'correct', *izmjenice* 'alternately' + *-ni* > *izmjenični* 'alternating').
Suffix *-ički* bears the similar meaning to the suffix *-ski*, but it is attached solely to the stems of foreign origin (e.g. *artist* 'artist' + *-ički* > *artistički* 'artistic').
Apart from the most frequent suffixes discussed above, we came across another ca 15 suffixes occuring in less than 20 examples and having more specialized meanings, e.g. *-ast* (*kruška* 'pear' + *-ast* > *kruškast* 'resembling to a pear').

[10] Other steps of the analysis, when formalized into a table, require single input, i.e. single root, allomorph and morph of the stem, and this field will enable us to determine roots, allomorphs and morphs of the stems for compounds as well.

[11] The statistical analysis in this subsection takes into account only the final step of the word-formation, i.e. the final suffix, and not all suffixes in the morphological structure of particular words.

4.3.2 Morphological analysis

Based on the material analyzed by now, the preliminary generalized morphological structure of the Croatian adjective is as follows:
1) prefixal part: two slots (**iz_ne**_nad*an+ 'sudden')
2) lexical part: three slots (**ran***o#**sred***n+j+o#**vjek***ov+n+i? 'early-medi-eval'; -o- = interfix)
3) suffixal part: three/four slots[12] (is_traž***iv+ač+k+**i? 'investigative')
4) inflectional ending (is_traž*iv+ač+k+**i?** 'investigative'). However, at least two things have to be pointed out regarding this structure and the design of the database. Firstly, some derivational affixes are already used in previous derivational stages. More precisely, they are actually used for the derivation of adjectival stems, and not for the derivation of adjectives themselves. We can illustrate this with the following derivational process: *iz-* + *tražiti* 'to search' > *istražiti* 'to investigate, perfective' + *-ivati* > *istraživati* 'to investigate, imperfective' + *-ač* > *istraživač* 'investigator' + *-ski* > *istraživački* 'investigative'. In the morphological structure of the adjective *istraživački* 'investigative', there is one verbal prefix (*iz-*), one verbal suffix (*-iv*), one nominal suffix (*-ač*) and one adjectival suffix (*-ski*[13]). How to account for such issues when structuring the database and how to simultaneously present derivational stages and morphological structure of words still raise many questions.
Secondly, the general adjectival structure presented above is probably tentative. Since it is based upon 1.000 analyzed adjectives, a more comprehensive picture will be formed once the rest of the collected material is processed and incorporated into the database.
In regard to the number of prefixes, suffixes and roots of Croatian adjectives, Marković [5] states that this POS can have up to three prefixes (although very rarely), e.g. ***ne-za-do-**voljan* 'displeased' and up to seven suffixes, e.g. *gost**-i-o-n-ič-ar-sk-i***[14]. Again, they are not all adjectival prefixes and suffixes, just as was the case with the adjective *istraživački* 'investigative' above. Moreover, Marković [5] lists adjectives with as many as six roots: ***tisuću*-devet*-sto*-sedam*-deset*-četvrt****-o godište* 'age 1974'. Although this kind of adjectives is not usually listed in dictionaries, it is frequent, especially in spoken language. Therefore, our intention is to collect them from available corpora and to incorporate them into our database in further stages of its development.

5 Concluding remarks and future work

In this paper we have presented the strategies applied in the building of CroDeriV, a derivational database for Croatian. Presently, CroDeriV contains only verbs, whereas words of other POS are being processed and will be introduced in future

[12] Four, if we also take inflectional ending into account. Marković [5] counts it in the maximal number of suffixes of Croatian adjectives, cf. later in the text.

[13] *-ki* is one of the possible allomorphs of the suffix *-ski*, cf. previous subsection.

[14] Adjectives like this one are common in Croatian.

stages of its development.

Each lemma in the database is morphologically analyzed. This analysis enables the recognition of derivational families via shared roots. The morphological analysis contains the full segmentation of words into morphemes as well as the recognition of stems used in particular derivational processes. The analysis is performed manually since the results of automatic approaches have so far been rather poor and unsatisfactory (cf. [13]). Manual analysis also enables the disambiguation of homographic roots and numerous affixal allomorphs.

Since the database contains only one POS, its expansion is necessary. We have presented procedures used in the processing of Croatian adjectives. Methodologically, we follow 'one POS at the time' approach for two reasons: 1) there are no available lists of derivational families for Croatian and therefore 2) it is easier to collect and process lemmas collected from various corpora and dictionaries. Although time-consuming and in many cases challenging, the processing of material in such an approach is more accurate and precise. The same approach will be used for the expansion of CroDeriV with nouns.

Acknowledgments

This paper is financially supported by the short-term research support of the University of Zagreb.

References

[1] Stjepan Babić. *Tvorba riječi u hrvatskome književnome jeziku*. HAZU : Nakladni zavod Globus, 2002.

[2] Nikola Ljubešić and Tomaž Erjavec. Corpus vs. lexicon supervision in morphosyntactic tagging: the case of slovene. In Nicoletta Calzolari et al., editor, *Proceedings of LREC 2016*, Paris, France, May 2016. ELRA.

[3] Ivan Marković. *Uvod u pridjev*. Disput, 2010.

[4] Ivan Marković. Tri radna ljuda: O supletivnosti i mogućim hrvatskim riječima. In *Ivan Slamnig, ehnti tschatschine Rogge! (Zbornik radova 10. kijevskih književnih susreta)*, pages 159–187. Općina Kijevo – Pučko otvoreno učilište Invictus – AGM, 2011.

[5] Ivan Marković. O najvećim (i) mogućim hrvatskim riječima. In Stjepan Blažetin, editor, *XI. međunarodni kroatistički znanstveni skup*, pages 43–58. Znanstveni zavod Hrvata u Mađarskoj,, 2013.

[6] Ivan Marković. *Uvod u jezičnu morfologiju*. Disput, 2013.

[7] Krešimir Mićanović. Posvojni pridjevi i izražavanje posvojnosti. *Suvremena lingvistika*, 49–50:111–123, 2000.

[8] Milan Moguš, Maja Bratanić, and Marko Tadić. *Hrvatski čestotni rječnik*. ZZL: Školska knjiga, 1999.

[9] Matea Srebačić, Krešimir Šojat, and Božo Bekavac. Croatian Derivational Patterns in NooJ. In J. Monti et al., editor, *Formalising natural languages with Nooj*, pages 55–62. Cambridge Scholars Publishing, 2015.

[10] Marko Tadić and Sanja Fulgosi. Building the Croatian Morphological Lexicon. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, pages 41–46. ACL, 2003.

[11] Branka Tafra. Razgraničavanje opisnih i odnosnih pridjeva (leksikološki i leksikografski problem). *Rasprave Zavoda za jezik*, 14:185–197, 1988.

[12] Marija Znika. Opisni i odnosni pridjevi. *Suvremena lingvistika*, 43–44:341–357, 1997.

[13] Krešimir Šojat, Matea Srebačić, and Tin Pavelić. Croderiv 2.0: Initial experiments. In Adam Przepiórkowski and Maciej Ogrodniczuk, editors, *Advances in Natural Language Processing*, pages 27–33. Springer, 2014.

[14] Krešimir Šojat, Matea Srebačić, and Vanja Štefanec. Croderiv i morfološka raščlamba hrvatskoga glagola. *Suvremena lingvistika*, 75:75–96, 2013.

On Compounding from Latin to Italian

M. Silvia Micheli

University of Pavia - University of Bergamo
E-mail: silvia.micheli@unibg.it

Abstract

This paper[1] deals with compounding from Latin to Italian. After a survey on how compounding is treated in the *Word Formation Latin* (WFL) lexicon, the fate of Latin compounds in Italian is analyzed from a morphological point of view, focusing on what of Latin Compounding (LC) has survived and what is lost. It is shown that most of Latin compounds have been totally lost or they were re-analyzed as derived or simple words. This causes a strong discontinuity between Latin and Italian compounding and a system reorganization, common to all Romance Languages, in compound word formation.

1 Introduction

All Romance Languages (RLs) show, from the early stages, a word-formation system essentially based on derivation (Dardano [4]; Rainer and Grossmann [10]). Even in Italian, the increase of the lexicon occurs through derivation, in particular through suffixation, which can be considered the most productive word-formation mechanism throughout its history. The link between Latin and Italian derivation is very clear: most of the Italian derived words are made up of affixes inherited from Latin (e.g. lat. *-ariu(m)* > it. *-aio*; lat. *mentu(m)* > it. *-mento*) and almost all still productive at the present stage. On the other hand, the relationship between Latin and Italian compounding seems to be much more opaque. At least until the XIX century, Italian Compounding (IC) is a rather rare, but not totally unproductive, word-formation mechanism which presents many differences compared to Latin Compounding (LC).[2] According to Dardano [4], the discontinuity between LC and IC depends on phonetic changes and gradual loss of transparency which involved Latin compounds in the transition from Vulgar Latin to Italian, during which much of the compound words that are fully transparent in Latin (e.g. *aurifice(m)* < *aurum* + *facere* or *iudicem* < *ius* + *dicere*) have become unanalyzable and opaque. This contribution focuses on the fate of Latin compounds in Italian through the analysis of

[1] I am grateful to three anonymous reviewers for useful comments on earlier draft of this paper. The usual disclaimers apply.

[2] For an overview on LC see Brucale [3] and Oniga [16].

data collected in WFL, i.e. a derivational lexicon for Latin. After a brief introduction on how compounding is represented in WFL and which are limits and potentials of this resource, an overview of what of LC has lost and what has survived in Italian will be provided.

2 Compounding in the *Word Formation Latin* lexicon

WFL is a derivational morphology resource for Latin (Litta et al. [14]) where lexical entries are analyzed in their morphological components and connected by word-formation rules (WFRs). Since the two main types of WFRs are derivation and compounding, WFL can be considered a powerful tool for the study of Latin compounds, especially for quantitative analysis.[3]

The design of WFL is consistent with the Item-and-Arrangement model (Hockett [11]), which considers morphemes, not words, the basic units for the study of utterance. Following this model, in WFL affixes are considered lexical elements as well as bases and recorded with the same status. In compounding, lexical bases are connected through WFRs which are automatically detected by considering all possible combinations of PoS (e.g. nouns, verbs, adjectives, pronouns and invariable lexemes) and further specified by the inflectional category of both input and output (e.g. N+V=N is the WFR that describes the creation of *pontifex*). For each compound, a derivational graph that shows the WFR is provided (as in figure 1).

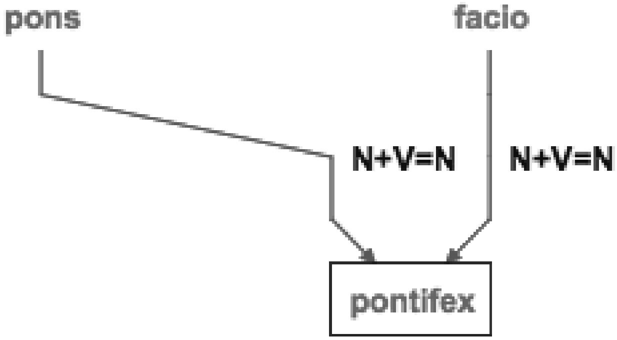


Figure 1: Derivation graph of *pontifex*

WFL provides information about compound constituents (i.e. input categories), the WFRs through which are connected and output categories. It allows to investigate the productivity of certain constituents or patterns and to compare them with

[3] LC can also be analyzed through CompoNet, i.e. a large database of compounds developed at the Department of Foreign Languages of the University of Bologna. However, unlike WFL, the CompoNet database is not a freely available resource.

other morphological strategies. However, since it does not provide either semantic information (i.e. constituents and whole compound meanings) or frequency information about compounds, it has to be combined with other resources for more in-depth investigation about Latin lexicon. As the portrait of LC shown in this contribution is exclusively based on WFL, it provides a strictly morphological analysis, which will be deepened and expanded in future studies.[4]

The sample collected in WFL consists of 1813 compounds created through 63 WRFs. As shown in Figure 2, the most productive WFRs are N+V=A (e.g. *calor* + *facio* = *calorificus*) and N+V=N (e.g. *aquila* + *fero* = *aquilifer*)[5]. The productivity of these two WFRs has not been preserved in Italian, in which the most productive WFR is V+N=N (e.g. *portapenne* 'penholder').[6] The change that occurred in the constituent order can be related to the syntactic change from OV to VO order between Latin and RLs. This correspondence between the constituent order in compounds and in syntax would support the hypothesis (Gaeta [8]) according to which, for this specific property, morphology is not autonomous from syntax. Furthermore, even the WFRs that form verbal and invariable compounds (particularly, A+V=V, N+V=V and I+I=I) show a good productivity that is not preserved in Italian, in which compounding is exploited to create almost exclusively nouns and adjective.

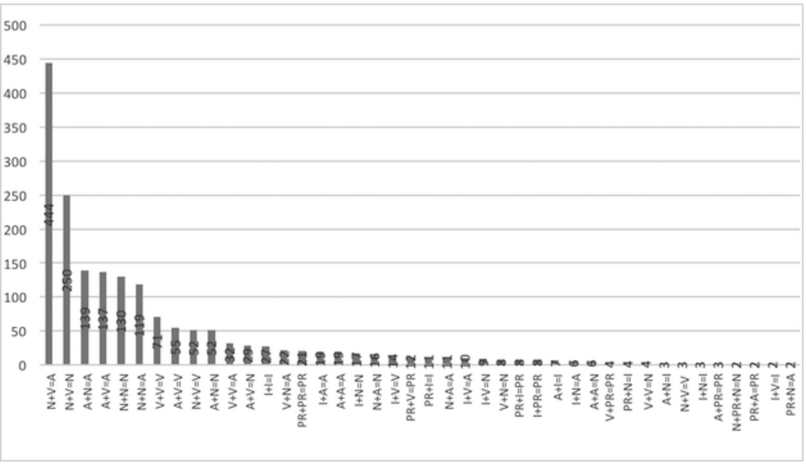


Figure 2: WFRs of compound words in WFL

[4] Context and frequency information are crucial in order to distinguish between compounds that are *hapax* (created by authors in literary works) and compounds which are more deeply entrenched in Latin lexicon.

[5] It should be noted that in WFL present and past participle are considered belonging to the category of verbs (e.g. *malevolens*).

[6] See Štichauer [18] for an overview of this type of Italian compounds from a diachronic point of view.

3 *Quid manet*?

Unlike other areas of morphology where one can perceive a significant continuity between Latin and Italian (e.g. derivation), very little of LC survives in Italian so much so that one can rather speak of two different word-formation systems.
In order to better understand this discontinuity, Latin compounds have been here classified into two types depending on their fate in Italian, i.e. compounds that have survived and compounds that were lost (Figure 3).

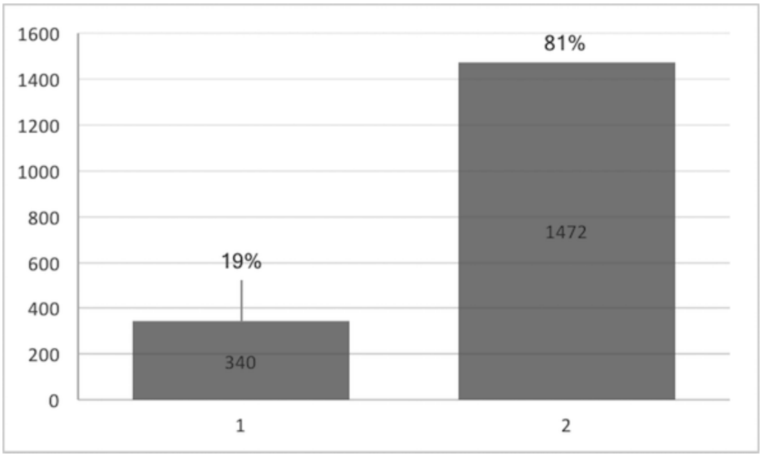


Figure 3: Survived (1) and lost (2) Latin compounds in Italian

Survived Latin compounds have been further sub-classified into three groups depending on their morphological structure (Figure 4):

1. compounds that survive in Italian as such (Group 1);
2. compounds that partially survive in Italian, as a constituent undergoes a grammaticalization process that leads it to become a productive suffix or affixoid of Neoclassical compounding (Group 2);
3. compounds that are kept in Italian but became partially or totally opaque and in some cases have been reanalyzed as simple words (Group 3).

The aliveness or death of each Latin compound in Italian has been determined checking the presence of each compound in Italian dictionaries (i.e. De Mauro [5], Treccani Online Dictionary, TLIO). In order to evaluate if, and to what extent, these forms are still in use in Contemporary Italian, the presence of each compound in a web corpus, i.e. itWaC (Baroni et al. [1]), has also been checked.
In the following sections, an overview of each type of Latin compounds fate will be outlined, on the basis of quantitative and qualitative data provided by WFL.

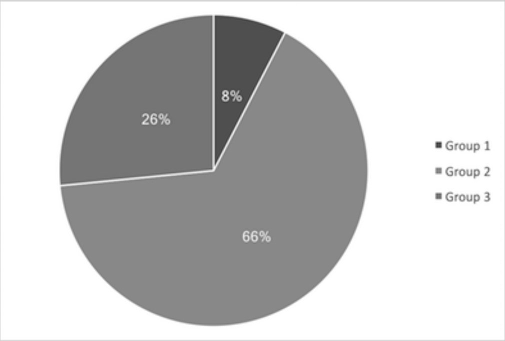


Figure 4: Latin compounds that are still used in Italian: distribution according to their morphological structure

3.1 Latin compounds that survived in Italian

The first group consists of Latin compounds that belong to the Italian lexicon and are still considered structurally compounded (Table 1).[7] They are described by Italian dictionaries as compound words and, in some cases, represent a compounding pattern that has been productive throughout the history of Italian language (e.g. compounds made up of *male* ’badly’ or *bene* ’well’ + present participle/adjective, such as *malfidus* ’dishonest’ or *benevolens* ’benevolent’).

They are structurally similar to Italian compounds, as they are made up of two autonomous words, even though some compounds (i.e. *altitonante*, *armipotente*, *capricorno*, *caprifico*, *crocifiggere*, *tragicommedia*, *verisimile*) maintain a clue of the original Latin compound structure, i.e. the linking element (-*i*-) between the two constituents.[8].

The second group is made up of Latin compounds which are made up of a constituent that undergoes a grammaticalization process that leads it to become a still productive affix (or affixoid). In Italian, these compounds are considered derived word or belonging to the category of Neoclassical compounds (Iacobini [13]).[9]

This group is made up of the following compound types:

[7] Obviously, the nature of this group of compounds depends on what is meant by "compound word" in Italian, which represents a still widely debated issue in literature (Grandi [9] and Masini and Scalise [15]). In this work, a definition of Italian compound as a word which is made up of two syntactically autonomous words has been assumed.

[8] It can be considered, following Ralli [17], as a compound marker that identifies the compounding process. According to Oniga [16], it is the result of the phonetic change of the thematic vowel of the first member. In WFL, about 71% of compounds shows the linking element -*i*-, which is kept in Italian as a residual element both in ancient compounds, such as *pettirosso* ‘robin’ or *capinera* ‘blackcap’ (created around the XV century), and in more recent formations, such as *altipiano* (attested from the XIX century).

[9] In order to identify this group of compounds, WFL data were compared to the set of affixoids which are collected and analyzed in Iacobini and Giuliani [12].

WFR	Latin comp.	Italian comp.	First occur.	Freq. in itWac
V+V=A	altitonans	altitonante	1332	4
N+N=N	arcuballista	arcobalista	XIII cent.	1
N+V=A	armipotens	armipotente	XIV cent.	0
I+V=V	benedico	benedire	XIII cent.	16.420
I+V=V	benefacio	benfare	1427	28
I+V=V	benefico	beneficare	1527	1.576
I+V=A	benevolens	benvolente	1294	9
N+N=N	capricornus	capricorno	1282	489
N+N=N	caprificus	caprifico	1340	22
N+N=N	carroballista	carrobalista	XVII cent.	5
N+V=V	crucifigo	crocifiggere	1321	3.189
I+V=V	maledico	maledire	1306	6.278
I+A=A	malefidus	malfido	1530	37
I+V=A	malevolens	malvolente	1400	8
N+N=N	malogranatum	melogranato	XIII cent.	1
N+V=V	manu-mitto	manomettere	1292	2.161
N+N=N	milifolium	millefoglio	XIV cent.	9
N+N=A	milleformis	milleforme	XIV cent.	1
A+N=A	primogenitus	primogenito	XIII cent.	4.203
A+A=A	sacrosanctus	sacrosanto	1313	7.639
N+N=N	tragicomoedia	tragicommedia	1543	287
A+A=A	verisimilis	verisimile	1311	6.532

Table 1: Latin compounds that are still compound words in Italian

- Latin compounds (55 in WFL) whose second constituent is a verbal root related to *facio* (Brucale and Mocciaro [2]), i.e. *-fic-*. In Italian, it can be considered a productive verbalizing suffix (e.g. lat. *clarificare* > it. *chiarificare*, lit. ‘to make clear’) or an adjectival suffix (e.g. lat. *salvificus* > it. *salvifico* ’peaceful’, lit. ’who/what gives safety’);

- Latin compounds whose constituents became affixoids (or semiwords) of the Italian Neoclassical compounding, e.g. *-fer-* (verbal roots related to *fero* ’to bring’) or *multi-* ’multi-’. Figure 5 shows the most productive patterns in which the second constituent represents a still productive suffixoid in Italian Neoclassical compounding.[10]

The last group includes Latin compounds whose structure is partially or totally opaque.[11] They can be ordered in a *continuum* from compounds in which the

[10] An example is provided for each pattern, e.g. lat. *fructifer* > ita. *fruttifero* ’frutiful’ represents an instance of the 55 compounds collected in WFL with *-fer* as second constituent.

[11] For a more appropriate analysis about these forms, their transparency should be tested through a

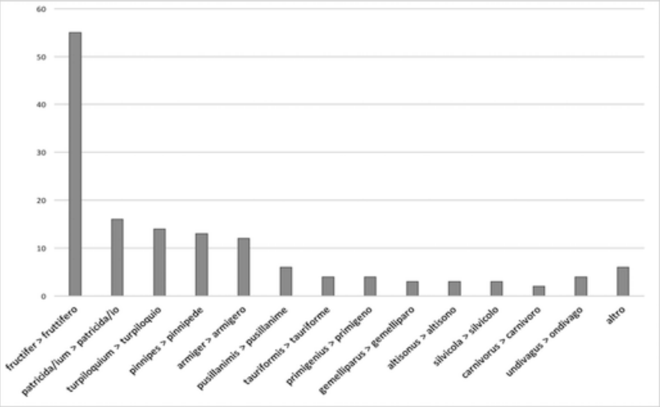


Figure 5: Latin compounds in which the second constituent represents a still productive suffixoid in Italian Neoclassical compounding

original Latin constituents are still recognizable, but they are no longer productive in Italian Neoclassical or native compounding (1), to compounds which undergo phonetic changes that make them totally opaque and are reanalyzed as simple words (2).
(1) *sanguisuga* (*sanguis+sugo*) 'bloodsucker' > it. *sanguisuga*
(2) *cordolium* (*cor+dolor+* SUFF) 'condolences' > it. *cordoglio*

More factors seem to play a role in determining the loss of transparency and the reanalysis of these forms:

1. the presence of bound forms as constituents that are often not easy to identify:
   (3) *manifestus* (*manus+fendo*) 'clear' > it. *manifesto*
2. the presence of the *–ium* suffix, that shows how the strong link and the fuzzy borders between compounding and derivation in Latin:
   (4) *plenilunium* (*plenus+lun*+suff) 'full moon' > it. *plenilunio*
3. the reduction of the number of syllables, especially in morpheme borders:
   (5) naufragium (*navis+frango*) 'shipwreck' > it. *naufragio*

All these factors seem to reinforce the cohesion of these words and contribute to hide their compounded nature. The disappearance of the boundaries between constituents has two consequences: Latin compounds are often reanalyzed as simple words and cease to be a model for the creation of new forms. RLs do not therefore inherit a productive system of compound words formation.

psycholinguistic experiment with Italian speakers.

3.2 Latin compounds that have been lost in Italian

As already pointed out above, most of the Latin compounds collected in WFL (81%; Figure 3) have not been preserved in Italian. Data collected in WFL allow to highlight which types of Latin compounds have suffered a greater decrease. Figure 6 shows how many forms have survived and how many have been lost for each type of compounds.

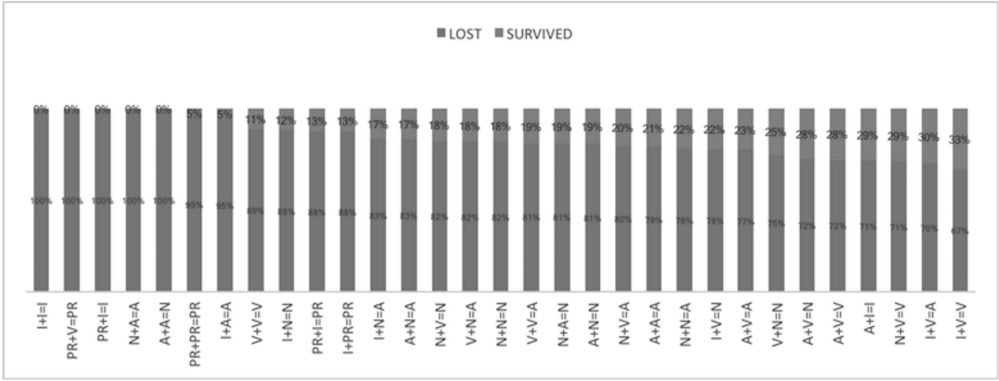


Figure 6: Quantitative distribution of compound words that have survived and have been lost for each compound type, identified by the WFR

Overall data collected in WFL show that each compound type has lost more than half of its elements. Invariable forms (i.e. conjunction or adverbs) and pronouns represent the compound types that have suffered a greater decrease. As far as adverbs are concerned, they were often replaced by other forms created by different morphological (e.g. derivation through *-mente* adverbial suffix) or syntactic mechanisms. This is the case of compound adverbs made up of an adjective and *–opere* (*opus* + SUFF) as second constituent (Table 2).

Latin compound	Italian compound
magnopere ‘strongly’	*intensamente*
maximopere ‘more overly’	*più intensamente*
nimiopere ‘overly’	*eccessivamente*
quantopere ‘as much as’	*tanto quanto*
tantopere ‘insomuch’	*talmente*

Table 2: Compound adverbs with *-opere* as second constituent

Latin pronouns created by compounding (e.g. *aliquis*, *aliquot*, *aliquot*, *quisquis*) have had a similar fate: only two of the 59 forms listed in WFL have survived in Italian, although their compound nature is now completely opaque (2).
(6) *aliquantus* ’rather’ > it. *alquanto*

qualiscumque 'whatever' > it. *qualunque*

More in-depth research is needed to try to better understand why these forms have been lost in the transition from Latin to Italian: WFL data allow to outline a partial picture only, as it does not provide semantic and quantitative information, which are crucial to analyze the life cycle of words. On the other hand, it has contributed to this research providing clues about where to focus our attention.

4 Conclusions

This survey based on WFL confirms the strong discontinuity between LC and IC due to the lost or the reanalysis as derived or simple words that Latin compounds undergo. It has been shown that most of the Latin compounds collected in WFL (especially invariable forms and pronouns) cease being used in Italian and that this decrease has affected all types of compounds. Very little of LC survives in Italian: only 19% of the sample is (or has been) attested in Italian. These forms can be classified in three groups depending on their morphological structure in Italian. The first group is made up of compounds which keep being productive pattern (e.g. compounds with *male* or *bene* as first constituents) in Italian. The second group includes compounds in which a constituent undergoes a grammaticalization process that leads it to become a productive suffix or affixoid of Neoclassical compounding. In the last group one can find Latin compounds whose structure is totally or partially opaque and which are therefore considered as simple word in Italian. This gap between Latin and Italian (as in other RLs) leads to a reorganization of compound word formation system and the rise of a new Romance Compounding.

References

[1] Baroni, M., Silvia Bernardini, Ferraresi, Alessandro and Zanchetta, Emanuele. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. In *Language resources and evaluation*, 43(3), pp. 209-26, 2009.

[2] Brucale, Luisa, & Mocciaro, Egle. Composizione verbale in latino: il caso dei verbi in facio, fico. In Paolo Poccetti (ed). *LATINITATIS RATIONES: Descriptive and Historical Accounts for the Latin Language*, pp. 279-97, Berlin/Boston: de Gruyter, 2016.

[3] Brucale, Luisa. Latin compounds. In *Probus*, 24, pp. 93–117, 2012.

[4] Dardano, Maurizio. *Costruire parole: La morfologia derivativa dell'italiano*. Bologna: Il Mulino, 2009.

[5] De Mauro, Tullio. *Grande dizionario dell'uso della lingua italiana*. Torino: Utet, 2009.

[6] Fruyt, Michèle. Constraints and productivity in Latin nominal compounding. In *Transactions of the Philological Society*, 100(3), pp. 259–87, 2002.

[7] Gaeta, Livio, Ricca, Davide. Composita solvantur: Compounds as lexical units or morphological objects. In *Rivista di Linguistica*, 21(1), pp. 35–70. 2009.

[8] Gaeta, Livio. Constituent order in compounds and syntax: typology and diachrony. In *Morphology*, 18(2), pp. 117–41, 2008.

[9] Grandi, Nicola. Considerazioni sulla definizione e la classificazione dei composti. In *Annali dell'Universitò di Ferrara. Sezione di Lettere*,1(1), pp. 31–52, 2006.

[10] Rainer, Franz, and Grossmann, Maria. *La formazione delle parole in italiano*, Tübingen: Niemeyer, 2004.

[11] Hockett, Charles F. Two models of grammatical description. In *Word*, 10(2-3), pp. 210–234, 1954.

[12] Iacobini, Claudio, Giuliani, Alessandro. A multidimensional approach to the classification of combining forms. In *Italian journal of linguistics*, 22(2), pp. 287–316, 2010.

[13] Iacobini, Claudio. Composizione con elementi neoclassici. *La formazione delle parole in italiano*, pp. 69–95, Tübingen: Max Niemeyer Verlag, 2004.

[14] Litta, Eeleonora, Passarotti, Marco, Culy, Chris. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC–it 2016)*. Napoli, pp. 185–89, aAccademia University Press, 2016.

[15] Masini, Francesca, Scalise, Sergio. Italian compounds. In *Probus*, 24, pp. 61–91, 2012.

[16] Oniga, Renato. Compounding in Latin. In *Rivista di linguistica*, 4(1), pp. 97–116, 1992.

[17] Ralli, Angela. Variation in word formation: the case of compound markers. In Pierluigi Cuzzolin, Maria Napoli (eds.). *Fonologia e tipologia lessicale nella storia della lingua greca*, pp. 238–64, Milano: Franco Angeli, 2006.

[18] Štichauer, Pavel. Verb-noun compounds in Italian from the 16th century onwards: An increasing exploitation of an available word-formation pattern. In *Morphology (Special issue: Modelling compound properties)*, 26(2), pp. 109–31, 2016.

Adding Morpho-phonological Features into a French Morpho-semantic Resource: the Demonette Derivational Database

Fiammetta Namer[(a)], Nabil Hathout[(b)] & Stéphanie Lignon[(a)]

[(a)]UMR 7118 ATILF, CNRS & Univ de Lorraine, France

[(b)]UMR 5263 CLLE-ERSS, CNRS & Univ de Toulouse-Jean Jaurès, France

E-mail: {fiammetta.namer, stephanie.lignon}@univ-lorraine.fr

nabil.hathout@univ-tlse2.fr

Abstract

Demonette (Hathout & Namer [13]) is a derivational database (DDB) of French with a relational structure: its entries describe a large number of properties of derivational relations connecting word pairs, such as LANCER 'launch'→ LANCEUR 'launcher' or LANCEUR → LANCEMENT 'launching'. The entries also specify the categorical, semantic and morpho-phonological properties of the connected words . We here present the morpho-phonological ones and show how Demonette's organization allows an original representation of these properties. Demonette's entries provide phonological transcriptions of the word pairs and syllabic decompositions. It also specifies their stems and the possible variations they display.

1 Introduction

Demonette (Hathout & Namer [13]) is a derivational database (DDB) of French which represents the morphological information in an original way: entries do not describe the properties of the derivatives; they describes the properties of the derivational relations connecting pairs of lexemes, such as LANCER 'launch'→ LANCEUR 'launcher$_{masc}$' or LANCEUR → LANCEMENT 'launching'. These relations specify the derivational properties of the lexemes they connect. One consequence of this conception is that the overall properties of a lexeme are the outcome of all the properties induced by each of the relations the lexeme occurs in. More generally, Demonette's structure is completely determined by this conception: The DDB is redundant, because relations are direct, indirect and bi-directional. Demonette describes relations between derivationally related pairs of lexemes [L1, L2], where L1 is morphosemantically motivated by L2. It includes relations between derived words and their bases (e.g. [LANCEUR, LANCER], where LANCEUR's meaning can be defined as "the one who performs the action of LANCER"), and relations between base words and their derivatives (eg. [LANCER, LANCEUR], where LANCER means "doing what a LANCEUR does"). The network also contains *indirect* relations between lexemes of the same derivational family, where none is the base of the other such as [LANCEMENT, LANCEUR]. In this relation, LANCEMENT can be defined as "activity performed by the LANCEUR". The relation is part of a network which contains [LANCER, LANCEUR], [LANCEUR, LANCER], [LANCEMENT, LANCER] (LANCEMENT is the "activity of LANCER"), [LANCER, LANCEMENT].

Derivational relations define derivational families, and are organized into paradigms. In previous publications, we focused on the morphosemantic characteristics of Demonette. We here address the morphophonological aspects of the DDB, and we show how these properties are described in Demonette and how morphophonological paradigms can be represented.

2 Derivational databases

One key feature of derivational morphology is its lexicality. Moreover, the analysis of complex lexemes relies on a large amount of memorized information.

In recent years, several efforts have improved the morphological analysis by using large corpora (Cotterrel [9], Lazaridou [15]), but progress on morphological information storage and harmonization has been weaker. A lot remains to be done: the accumulation of morphological knowledge is crucial for many researches in descriptive morphology, lexicology, teaching, etc.

The first DDBs where designed by psycholinguists in order to create experimental data. The best-known DBB is CELEX (Baayen *et al.* [2]) whose first version was released in the 90s. This resource covers English, German and Dutch and offers a broad range of phonological, morphosyntactic, inflectional and derivational information. It remains a reference with no real equivalent, despite its limited coverage, when compared to the size of the corpora available today.

Other large-scale resources have been created for English, such as CatVar (Habash & Dorr [11]), a lexicon of derivational family intended primarily to NLP applications. More recently, a similar resource has been developed for German: DerivBase (Zeller *et al.* [30]) was automatically built from corpora, with the help of distributed semantics methods. Another significant resource is DerivaTario (Talamo *et al.* [26]), a derivational dictionary of Italian; It provides analyses based on strong hypotheses regarding allomorphy and suppletion. For instance, BELLICOSO 'bellicose' is analyzed as a derivative of GUERRA 'war'. For French, the only comparable resource is Demonette. Its main characteristics are presented hereafter.

3 Demonette

One goal of Demonette (Hathout & Namer [13]) is to help satisfy the need for reliable and broad-coverage morphological resources of French. Demonette is a DDB characterized by an original structure based on the derivational relations. Moreover, it can host morphological descriptions from research works such as PhDs in morphology, or from manual-assessed NLP lexical resources, like VerbAction (Tanguy & Hathout [27]). In its current state (Hathout & Namer [14]), Demonette (version 1.3) includes information coming from four sources: DériF (Namer [18], [19]), Morphonette (Hathout [12]), VerbAction and Lexeur (Fabre *et al.* [10]). They have been added in three successive stages. Overall, Demonette contains 167,369 entries. Derived words in Demonette can be deverbal action nouns (ESSORAGE 'spin'), deverbal masculine or feminine agent nouns (RAMASSEUR 'collector', RAMASSEUSE 'collector') or deverbal adjectives (PRODUCTIF 'productive'). Demonette also includes simplex verb predicates (CONSTRUIRE 'build').

The fields used to describe the derivational relations in the Demonette do not form a closed list and can be extended if needed. Among the existing fields, the most original ones are probably those used for the semantic description, and include morphosemantic types (eg. @AGF for feminine agents), concrete definitions giving the meaning of L1 with respect to L2 (eg. MARCHEUSE in the relation [MARCHEUSE, MARCHER] is defined as "she who performs the action of MARCHER"), and abstract definitions generalizing the concrete ones where the meanings of L1 and L2 are replaced by their respective semantic type (eg. @AGF: "she who performs @"). Relations with the same abstract definition are inserted into the same morphosemantic paradigm. This is the case with the ones listed in Table 1.

L1, cat	**L2, cat**	**Type L1**	**Type L2**	**Concrete def**	**Abstract def**	**Affix L1**
MARCHEUSE, N_{Fem} 'walker$_{(fem)}$'	MARCHER, V 'walk'	@AGF	@	"she who performs the action of marcher"	"she who performs the action of @"	euse
ENSEIGNANTE, N_{Fem} 'teacher$_{(fem)}$'	ENSEIGNER ,V 'teach'	@AGF	@	"she who performs the action of enseigner"		ante
DIRECTRICE, N_{Fem} 'director$_{(fem)}$'	DIRIGER,V 'direct'	@AGF	@	"she who performs the action of diriger"		rice

Table 1: Concrete and abstract definitions of three feminine agent nouns

4 Morpho-phonological descriptions within Demonette

The 167,369 [L1, L2] entries of Demonette1.3 have been completed with morphophonological information: L1 and L2 phonological representations, the properties of their stems and exponents, and a description of the morphophonological variations that occur in the [L1, L2] relation. This information is mostly unpredictable in a language with rich morphology such as French and is therefore crucial for a comprehensive description of its derivational system. This additional knowledge is interconnected with the rest of the entries properties and in particular with the morphological and the morphosemantic ones.

In Demonette, morphophonological properties are described in a similar way to the morphosemantic ones: we distinguish concrete and abstract levels; some of the morphophonological descriptions of the L1 and L2 lexemes are induced by the derivational relation which connects them. Morphophonology is both easier and harder to describe than morphosemantics. On the one hand, it is simpler, because IPA transcriptions are part of the mainstream in linguistics: we don't have a similar standard for morphosemantic representation. On the other hand, it is more complex, because lexemes are abstract objects that do not have formal properties by themselves, unlike the inflected forms (or word forms) that realize them. We also consider that each word form can be decomposed into an inflectional stem and an inflectional exponent (Baerman *et al*. [3]).

Following Boyé [8], Bonami & Boyé [4] and Montermini & Bonami [17], we define word form exponents in French as the maximal rightmost strings that are common across the patterns, and interpret all the remaining variation as stem allomorphy (see Spencer [25] and Bonami & Boyé [7] for a discussion), where stems are pure forms (or morphomes, in Aronoff's terms [1]). As often discussed in the literature (see Bonami & Boyé [4, 6], Montermini & Boyé [16], Montermini & Bonami [17] among others), stems form a paradigmatic organization called *stem space* (deriving from Pirrelli & Battista's [21] 'Overall Distribution Schema'). Stem spaces are made of cells forming a graph where stems are in a dependency relation with each other. The value of a stem occupying one cell depends on the value of stem in one or several other cells. By default, this value is inherited from them without change. Allomorphic stems correspond to override of the default inheritance. The complexity of the stem space is language and part-of-speech dependent. For instance, the stem space of French verbs is a graph of at least 13 cells. Table 2 lists the 13 stems of the verb BOIRE 'drink'. Each stem is used by one or several inflection rules to produce one or several forms of the verb[1].

[1] C1 is used for the IND.PRS.SG; C2: IND.PRS.3PL; C3: IND.IPFV & IND.PRS.1PL & 2PL; C4: PTCP.PRS; C5: IMP.2SG; C6: IMP.1PL & 2PL; C7: SBJV.PRS.SG & 3PL; C8: SBJV.PRS.1PL & 2PL; C9: INF.PRS; C10: IND.FUT & COND.PRS; C11: IND.PST; C12: PTCP.PST.M; C13: PTCP.PST.F.

C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**
bwa	bwav	byv	byv	bwa	byv	bwav	byv	bwar
C10	**C11**	**C12**	**C13**					
bwa	by	by	by					

Table 2: Stem space of the verb BOIRE ‘drink’

Likewise, adjective and noun stems are organized in stem spaces: in French, a 3 cells space is required for adjectives (C1: M.SG; C2: F.SG; C3:M.PL; see Boye & Bonami [5]) and a 2 cells one for nouns (C1:SG; C2:PL; see Roché [23]). Table 3 shows the spaces of the adjective BEAU ‘beautiful’ and of the noun CHEVAL ‘horse’.

ADJ		
C1	**C2**	**C3**
bo	bɛl	bo

NOUN	
C1	**C2**
ʃəval	ʃəvo

Table 3: Stem spaces of the French adjective BEAU and noun CHEVAL.

Stem spaces also play a central role in word formation: word formation patterns use particular cells in the stem space of the input lexemes. For instance, *-able* suffixed deverbal adjectives are formed with the C3 verb stem. Therefore, the stem /byv/ is selected to derive BUVABLE ‘drinkable’ from BOIRE. Similarly, deadjectival prefixed verbs are generally built on the C2 adjective stem: EMBELLIR, for instance selects the /bɛl/ stem of the adjective BEAU.

In Demonette we only provide the morphophonological properties of lexemes (or more precisely, of the wordforms that realize them) relevant for word formation (Plénat [22], Roché [23]). Therefore, stems and exponents are listed in the [L1, L2] description only if they are involved in derivational constructions. For French, this means that, out of the stem spaces illustrated in Tables 2 and 3, only the following are needed:

- For nouns, C1, e.g. CHEVAL ‘horse’ → CHEVALIER /ʃəvalje/ ‘horseman’
- For adjectives, C1 and C2 are relevant: the M.SG stem /bo/ of BEAU is used to form the property noun BEAUTÉ /bote/ ‘beauty’, and the F.SG stem /bɛl/ to form the pejorative noun BELLÂTRE /bɛlɑtʀ/ ‘fop’.
- Six stems are required for verbs: C1, C4, C12 and C13 are used by V-to-N conversion patterns (C1: SOUTENIR ‘support$_V$’ → SOUTIEN /sutjɛ̃/ ‘support$_N$’, C4: COURIR ‘run’ → COURANT /kurɑ̃/ ‘flow’, C12: DEVOIR ‘owe’ → DÛ /dy/ ‘due’, C13: MÉPRENDRE ‘be mistaken’ → MÉPRISE /mepriz/ ‘mistake’, cf. Tribout [28]), C2, used in *-ment* suffixed deverbal event nouns (SOULEVER ‘lift$_V$’ → SOULÈVEMENT

/sulεvmã/ ‘lift$_{N}$’), and C3 for *-able* suffixed adjectives (BOIRE ‘drink’ → BUVABLE /byvabl/ ‘drinkable’).

The main source of automatic acquisition for the IPA transcription of the selected stems is the freely available database GLÀFF (Sajous *et al.* [24]), which contains more than 1.4 million entries of inflected forms annotated with phonetic representation encoded in SAMPA (Wells [29]). When needed, it is completed with data coming from Lexique3 (New [20]), which uses phonetic transcriptions very similar to SAMPA, which makes the mapping task relatively trivial.

All but one of the stems of the lexemes present in GLÀFF or Lexique3 can be directly retrieved from the word forms for which they have been used. The exception is C3 for verbs, because this stem is always concatenated to an exponent in the word forms: the C3 stem is thus computed from the IND.PRS.1PL form by stripping off the final /ɔ̃/ exponent (eg. *buvons* ‘(we) drink’, /byvɔ̃/ = /byv/ ⊕ /ɔ̃/). The entries also contain various other pieces of information that describe the morphophonological specificity of L1, L2 and the [L1, L2] relation (see Tables 4 and 5).

In Table 4, the features *Rad1* and *Rad2* can be compared to determine the formal distance between L1 and L2. When [L1, L2] are in a base/derivative relation, as in [BOIRE, BUVEUR], *Rad2* is obtained by removing the suffix *Suf2* (e.g. /œr/) from the word form of the derivative (e.g. BUVEUR). *Rad1* is selected from the stem space of the base (e.g. BOIRE) in such a way that it is the most similar to one of the possible values of *Rad2*. In the example [BOIRE, BUVEUR], it is C3 (see Table 2).

When L1 and L2 are in an indirect relation, as in [ADMIRATEUR$_{N}$, ADMIRATION$_{N}$], both words being derived from ADMIRER$_{V}$ ‘admire$_{V}$’, the value of Rad_i is obtained by depriving L_i from the suffix Suf_i. For [ADMIRATEUR$_{N}$, ADMIRATION$_{N}$], we get *Rad1* = /admirat/ and *Rad2* = /admiras/. For each L_i, the Rad_i description also includes the number of syllables $Size_i$, and the properties of its last syllable (onset, vowel, final consonant), as value of, respectively, $LastOnset_i$, $LastV_i$ and $LastC_i$.

L1						**L2**					
Rad1	**Size1**	**LastOns1**	**LastV1**	**LastC1**	**Suf1**	**Rad2**	**Size2**	**LastOns2**	**LastV2**	**LastC2**	**Suf2**
BOIRE ‘drink$_{V}$’						BUVEUR ‘drinker (masc)’					
byv	1	b	y	v	--	byv	1	b	y	v	œr
ADMIRATEUR ‘admirer (masc)’						ADMIRATION ‘admiration’					
admirat	3	r	a	t	œr	admiras	3	r	a	s	jõ

Table 4: Rad1 and Rad2

Table 5 gives more examples of [L1, L2] formal properties, and shows how allomorphy is described.

		L$_i$	**Rad**$_i$	**Suf**$_i$	**Z**	**Alternation**	**Concrete Phon Rel**	**Abstract Phon Rel**	**Phon Rule**
1	L1 L2	boire buveur	byv byv	-- œr	byv	=	byv / byvœr	Z / Zœr	=
2	L1 L2	admirateur admiratif	admirat admirat	œr if	admirat	=	admiratœr / admiratif	Zœr / Zif	=
3	L1 L2	admirer admirateur	admir admirat	-- œr	admir	| at	admir / admiratœr	Z / Zatœr	NONE
4	L1 L2	admirateur admiration	admirat admiras	œr jõ	admira	t | s	admiratœr / admirasjõ	Ztœr / Zsjõ	→ [+sib]
5	L1 L2	extincteur extinction	ekstɛ̃kt ekstɛ̃ks	œr jõ	ekstɛ̃k	t | s	ekstɛ̃ktœr / ekstɛ̃ksjõ	Ztœr / Zsjõ	→ [+sib]
6	L1 L2	éteindre extincteur	etɛ̃ ekstɛ̃kt	-- œr	e	--	etɛ̃ / ekstɛ̃ktœr	etɛ̃ / ekstɛ̃ktœr	NONE
7	L1 L2	aliment ‘food’ alimentaire ‘alimentary’	alimã alimãt	-- ɛr	alimã	| t	alimã / alimãtɛr	Z / Ztɛr	+C
8	L1 L2	cheval ‘horse’ hippique ‘equine'	ʃəval ip	-- ik	--	--	ʃəval / ipik	ʃəval / ipik	NONE

Table5: Identity and variation in a derivational relation

Columns ***L***$_i$, ***Rad***$_i$, ***Suf***$_i$ contain the orthographic representation, the radical and suffix of each of the lexemes L1 and L2 (cf. above Table 4).
The other columns describe the properties of the [L1, L2] relation.

- The field ***Concrete Phon(ological) Rel(ation)*** reproduces the sequences formed by the concatenation of *Rad1* and *Rad2* (cf. Table 4).
- The longest common subsequence of *Rad1* and *Rad2* is given in ***Z***. *Z* can be identical to *Rad1* and *Rad2* as with /byv/ and /admirat/, in raw 1 and 2; it can be identical to *Rad1* and included in *Rad2*: both *Rad1* /admir/ in raw 3 and /alimã/ in raw 7 are included in their corresponding *Rad2*; it can be a subpart of *Rad1* and *Rad2*, such as /admira/ (raw 4), /ekstɛ̃k/ (raw 5) and /e/ (raw 6), or it be empty (raw 8).

- When *Z* has a non-null value, the difference between *Rad1* and *Rad2* is given in the field ***Alternation***. When *Rad1* and *Rad2* are identical, the value of *Alternation* is '=' (raws 1 and 2). The value 't|s' (in raws 4 and 5) says that the variation between *Rad1* and *Rad2* is a change in their last consonant; the value '|at' in raw 3 (resp. '|t' in raw 7) says that *Rad2* is the concatenation of *Rad1* with /at/ (resp. with /t/). We consider the *Alternation* value to be not relevant (value '--') when the two stems are completely different (raw 8), or when their difference (*i*) is not reproduced elsewhere in the lexicon, and (*ii*) is greater than their likeliness (raw 6).
- When relevant, *Alternation* is characterized phonologically. The explanation (assibilation, insertion, sonorization, etc.) is encoded as a rule in the ***Phonological Rule*** field (last column). The rule is identity, symbolized by '=' in raws 1 and 2. It contains the value 'NONE', e.g. in raw 3 because the difference between *Rad1* /admir/ and *Rad2* /admirat/ does not have a phonological origin but an historical one (/admirat/ is the Latinate bound stem of the verb ADMIRER). Likewise, the rules in raws 6 and 8 have a 'NONE' value because the stem variations between L1 and L2 are not phonologically motivated. Conversely, the 't|s' alternation in the relations of raws 4 and 5 can be qualified as a case of palatalization (or sibilantization), represented by the '→[+sib]' rule. The insertion of /t/ at the stem/suffix boundary of /alimã**t**ɛr/ in raw 7 is phonologically motivated (as opposed to the /at/ insertion in raw 3): it is the sonorization (symbolized with '+C') of the latent final consonant on the orthographical form *aliment*.
- *Z* is used to generalize the *Concrete Phon Rel* into an ***Abstract Phon(ology) Rel(ation)*,** where the Z symbol substitutes for the value of the *Z* attribute. This abstract relation emphasizes the morphophonological organization of the lexicon, in particular in terms of stem and exponent variation. This abstract representation also identifies the set of morphophonological relations that connect each lexeme to the rest of its derivational family.

The descriptions exemplified in Tables 4 and 5 allows us to separate the derivational relations into four categories according to morphophonological criteria, based on their identity, the variation between their stems, and the nature of their formal relation. This categorization uses the values of *Alternation* and *Abstract Phon Rel* fields. The four categories are:

(i) no stem variation (raw 1, 2);
(ii) phonologically motivated variation (raws 4, 5, 7);
(iii) stem variation surfacing as an alternation not phonologically motivated (raw 3);

(iv) suppletion, i.e. no sequence in common (raws 6, 8).

The alternations define morphophonological classes of derivational (sub-)families: for instance, the same set of A ↔ B stem variations are shared by [L1, L2] pairs in several derivational families, as shown in Table 6. Stem variations are evidenced by the *Abstract Phon Rel* value in each of the relevant [L1, L2] entries.

Table 6 shows that (COMPOSER$_V$ ‘compose’, COMPOSITEUR$_N$ ‘composer (m)’, COMPOSITRICE$_N$ ‘composer (f)’, COMPOSITION$_N$ ‘composition’) and (INHIBER$_V$ ‘inhibit’, INHIBITEUR$_N$ ‘inhibitor (m)’, INHIBITRICE$_N$ ‘inhibitor (f)’, INHIBITION$_N$ ‘inhibition’) share the same set of stem variations, and have the same suffix exponents *-eur, -rice* and *-ion*. Moreover, the indirect relations in Demonette highlights the formal organization of the lexicon. These relations make it possible to identify sub-regularities, for instance between EXTINCTEUR$_N$ ‘extinguisher’ and EXTINCTION$_N$ ‘extinction’ (raw 5, Table 5) or between PRÉDATEUR$_N$ ‘predator’ and PRÉDATION$_N$ ‘predation’: whereas the standard derivational connections between the first noun pair can be retrieved from their individual relations with their verb base ÉTEINDRE ‘extinguish’, as shown in Table 5, raw 6, there is no such direct base/derived relation in the French contemporary lexicon, between PRÉDATEUR or PREDATION and a common verb base.

		PRED(V)	M. AGENT(N)	F. AGENT(N)	EVENT(N)
Deriv. families		COMPOSER	COMPOSITEUR	COMPOSITRICE	COMPOSITION
		…			
		INHIBER	INHIBITEUR	INHIBITRICE	INHIBITION
A ↔ B	Z ↔ Zit	A	B		
		A		B	
	Z ↔ Z		A	B	
	Z ↔ Zis	A			B
	Zt ↔ Zs		A		B
				A	B

Table 6: Morphophonological organization of derivational families

5 Paradigmatic view of the derivational lexicon

With the organization we outlined above, Demonette has a triple network of morphological, morphosemantic and morphophonological relations able to capture paradigmatic regularities and sub-regularities at different levels. Just like morphosemantics, morphophonological information is described at two levels, a concrete one and an abstract one, which multiplies the perspectives of observation.

For instance, at the concrete level, noun pairs EXTINCTEUR ↔ EXTINCTION, ADMIRATEUR ↔ ADMIRATION and PRÉDATEUR ↔ PRÉDATION behave in the same way, whereas at the abstract level, (ADMIRER, ADMIRATEUR, ADMIRATION) and (ÉTEINDRE, EXTINCTEUR, EXTINCTION) belong to two distinct series.

Examined at different levels, the same data leads to different findings. For example, crossing morphology and morphophonology leads to the insertion of [PRÉDATEUR, PRÉDATION] in the sub-paradigm of the paradigm (ADMIRATEUR, ADMIRATION, ADMIRER).

If we consider the morphosemantic / morphophonology opposition, triplets (ADMIRER, ADMIRATEUR, ADMIRATION) and (CONSPIRER 'conspire$_V$', CONSPIRATEUR 'conspirator(m)$_N$', CONSPIRATION 'conspiracy$_N$') belong to two different morphosemantic paradigms (ADMIRER and ADMIRATION are stative predicates, whereas CONSPIRER and CONSPIRATION are eventive ones), but to the same morphophonological paradigm; conversely (ENSEIGNER 'teach', ENSEIGNANT 'teacher(m)', ENSEIGNEMENT 'teaching') is in the same morphosemantic paradigm as (CONSPIRER, CONSPIRATEUR, CONSPIRATION), but the two sub-families belong to distinct morphophonological paradigms.

Finally, the two families presented in Table 6 illustrate a case of uniform paradigm: members of the same morphophonological category share the same semantic category and the same part-of-speech (INHIBER and COMPOSER are verbal predicates, COMPOSITEUR and INHIBITEUR, masculine agent nouns, INHIBITRICE and COMPOSITRICE, feminine agent nouns, and COMPOSITION and INHIBITION event nouns). They result from the same derivational processes (the verbs are simplex, and the nouns are suffixed in *-eur*, *-rice* and *-ion* respectively) and are two by two in the same phonological relations, as shown in Table 6.

References

[1] Aronoff, Marc. *Morphology by Itself*, MIT Press: Cambridge, 1994.

[2] Baayen, R. Harald, Piepenbrock, Richard, and Gulikers, Leon. *The CELEX lexical database* (release 2). Linguistic Data Consortium, Philadelphia, 1995.

[3] Baerman, Matthew, Corbett, Greville, Brown, Dunstan and Hippisley, Andrew. *Deponency and Morphological Mismatches*, British Academy and Oxford University: Oxford, 2007.

[4] Bonami, Olivier and Boyé, Gilles. Supplétion et classes flexionnelles dans la conjugaison du français, pp. 102–126. *Langages* 152, 2003.

[5] Bonami, Olivier and Boyé, Gilles. Construire le paradigme d'un adjective, pp. 77–98. *Recherches Linguistiques de Vincennes* 34, 2005.

[6] Bonami Olivier and Boyé Gilles. Remarques sur les bases de la conjugaison. In Delais-Roussarie, Elisabeth and Labrune, Laurence (Eds.), *Des sons et des sens. Données et modèles en phonologie et en morphologie*, pp. 77–90. Hermès-Lavoisier: Paris, 2007.

[7] Bonami, Olivier and Boyé Gilles. De formes en thèmes. in Villoing, Florence, Leroy, Sarah and David, Sophie (eds.), *Foisonnements*

morphologiques : études en hommage à Françoise Kerleroux, pp. 19–44. Presses Universitaires de Paris-Ouest: Nanterre, 2014.

[8] Boyé, Gilles. *Problèmes de morphophonologie verbale en français, espagnol et italien*. PhD dissertation, University Paris 7, 2000.

[9] Cotterell, Ryan and Schütze Hinrich. Joint Semantic Synthesis and Morphological Analysis of the Derived Word. *Transactions of the Association for Computational Linguistics,* 2017.

[10] Fabre, Cécile, Floricic, Franck and Hathout, Nabil. Collecte outillée pour l'analyse des emplois discordants des déverbaux en -eur. Paper presented at the *Journées d'étude sur la place des méthodes quantitatives dans le travail du linguiste*, ERSS, Toulouse, 2004.

[11] Habash, Nizar and Dorr, Bonnie (2003) A categorial variation database for English. In *Proceedings of NAACL 2003*, pp. 96-102.

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

[13] Hathout, Nabil and Namer, Fiammetta. Démonette, a French derivational morpho-semantic network, pp. 125-168. *LiLT* 11, 2014.

[14] Hathout, Nabil and Namer, Fiammetta. Giving Lexical Resources a Second Life: Démonette, a Multi-sourced Morpho-semantic Network for French. In *Proceedings of LREC 2016*, Portorož, pp. 1084–1091, 2016. European Language Resources Association (ELRA): Slovenia.

[15] Lazaridou, Angeliki, Marelli, Marco, Zamparelli, Roberto and Baroni, Marco. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of ACL 2013,* pp. 1517–1526, Association for Computational Linguistics: Sofia, 2013.

[16] Montermini, Fabio and Boyé, Gilles. Stem relations and inflection class assignment in Italian. *Word Structure 5* (1), pp. 69–87, 2012.

[17] Montermini, Fabio and Bonami, Olivier. Stem spaces and predictability in verbal inflection, *Lingue e linguaggio 2013 (2)*, pp. 171–90, 2013.

[18] Namer, Fiammetta. *Morphologie, Lexique et TAL: l'analyseur DériF* Hermes Sciences Publishing: London, 2009.

[19] Namer, Fiammetta. A Rule-Based Morphosemantic Analyzer for French for a Fine-Grained Semantic Annotation of Texts. in Mahlow Cerstin and Piotrowski, Michael (eds.), *SFCM 2013*, pp. 93–115, Springer: Heidelberg, 2013.

[20] New, Boris, Lexique 3: Une nouvelle base de données lexicales. In *Proceedings of TALN*, Louvain, pp. 892–900, 2006.

[21] Pirrelli, Vito and Battista, Marco. The paradigmatic dimension of stem allomorphy in Italian. *Rivista di linguistica* 12 (2), pp. 307–380, 2000.

[22] Plénat, Marc. Le conditionnement de l'allomorphie radicale en français. *Mémoires de la Société de Linguistique de Paris*, Nouvelle série, n° 17, pp. 119–140, 2009.

[23] Roché, Michel. Base, thème, radical, *Recherches linguistiques de Vincennes*, 39 (1), pp. 95–134, 2010.

[24] Sajous, Franck, Hathout, Nabil and Calderone, Basilio. Glàff, un gros lexique à tout faire du français. In *Actes de la 20e conférence TALN*, pp. 285–298, 2013.

[25] Spencer, Andrew. Identifying stems *Word Structure* 5 (1), pp. 88–108, 2012.

[26] Talamo, Luigi, Celata, Chiara, & Bertinetto, Piermarco. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure 9* (1), pp: 72–102, 2016.

[27] Tanguy, Ludovic and Hathout, Nabil. Webaffix: un outil d'acquisition morphologique dérivationnelle à partir du Web. In *Proceedings of the 9e Conférence TALN*, pp. 245–254. ATALA: Nancy, 2002.

[28] Tribout, Delphine. Verbal stem space and verb to noun conversion in French. *Word Structure*, 5, pp. 109–28, 2012.

[29] Wells, J. SAMPA computer readable phonetic alphabet. *Handbook of Standards and Resources for Spoken Language Systems*. D. Gibbon, R. Moore and R. Winski. Berlin/New York, Mouton de Gruyter. Part IV, section B, 1997.

[30] Zeller, Britta, Snajder, Jan and Padó, Sebastian. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceeding of the 51th ACL 2013,* pp. 1201-1211. ACL: Sofia, 2013.

Internationalisms with the Suffix -*ácia* and their Adaptation in Slovak

Renáta Panocová[1]

Pavol Jozef Šafárik University in Košice, Slovakia
E-mail: renata.panocova@upjs.sk

Abstract

In modelling the system of derivation of a language, we generally assume that affixed words are derived from their non-affixed counterparts.This paper investigates the direction of motivation in pairs of mostly Latin origin such as *diverzifikovať* 'diversify' → *diverzifikácia* 'diversification'. In the Slovak linguistic tradition, such pairs were analogically modelled as derivation from verbs to nouns. In this paper I discuss two types of evidence which suggest that the direction of motivation is rather the opposite. One type is based on frequency, the other on the meaning of the two members of the pair.

1 Introduction

In the Slavic tradition international words or internationalisms are generally defined as words of Greek or Latin origin which occur in at least three genetically unrelated languages (Jiráček [7]) for instance, *communication* in English, *comunicazione* in Italian, and *komunikácia* in Slovak. At present, especially in the word formation of Slavic languages, internationalisation of languages is better understood as a tendency rather than as a single process (Buzássyová [2]). Internationalization as a tendency can only be observed when comparing different languages. The individual processes in each language reflect this tendency and are subsumed by it (Gutschmidt [4], [5]). In Slavic languages, many internationalisms were borrowed via French and German, but at present they most commonly arrive via English (Buzássyová [2]).

An example of an international suffix in contemporary Slovak is -*ácia* (with the variants -*izácia*, -*fikácia*). It attaches to international bases and is productive (Mistrík [10], Horecký et al., [6]). In terms of Horecký et al.'s [6] theoretical framework these words name actions, usually the process as a whole without differentiation into process, action proper and result. Such names of actions with international elements enter Slovak as nouns in -*ácia*, e.g. *komunikácia* 'communication', *integrácia* 'integration', *špecifikácia* 'specification', *kvantifikácia* 'quantification'. They are borrowings.

[1] This work was supported by the Slovak Research and Development Agency under the Contract No APVV-16-0035.

In general, lexical borrowings are considered unmotivated units in the recipient language. On the other hand, borrowings tend to gradually adapt to the word formation system of the recipient language. For instance, in Slovak there are more than 60 words derived from *rádio* ‘radio’, including derivatives, such as *rádiový* ‘radio$_{\mathrm{ADJ}}$’, *radista* ‘radio operator’, and compounds, such as *rádiotechnika* ‘radio engineering’, *rádiouzol* ‘radio node’, *rádioopravovňa* ‘radio repairs/service’ (Furdík [3]).

The *Retrograde Slovak Dictionary* by Mistrík [10] lists more than 2000 action nouns with the suffix *-ácia*. The majority of them are of Latin origin. As mentioned above, such borrowings are unmotivated lexical units in Slovak, for instance, *konštelácia* ‘constellation’ or *relácia* ‘relation’ (Furdík [3]). Interestingly, due to the fact that such nouns name actions, verbs derived from them are usually formed or more precisely backformed very soon (Furdík [3], Horecký et al. [6]). Furdík [3] illustrates this by the examples in (1).

(1) devastácia ‘devastation’ – devastovať ‘devastate’
distribúcia ‘distribution’ – distribuovať ‘distribute’
integrácia ‘integration’ – integrovať ‘integrate’

Despite the fact that diachronically, the verbs in (1) were formed later than the corresponding nouns, the status of the nouns in the synchronic perspective of the word formation system of Slovak is reevaluated (Horecký et al. [6]). Furdík [3] takes verbs with an international (originally Latin) element as primary motivation (motivating words) to form nouns with *-ácia*. This phenomenon is referred to as *remotivation* (Furdík [3], Mistrík [11], Horecký et al., [6]). The different stages are illustrated in (2).

(2) a. špecifikácia ‘specification’
b. špecifikácia ‘specification’ – špecifikovať ‘specify’
c. špecifikovať ‘specify’ → špecifikácia ‘specification’

In (2a) we see an example of the first stage in the process. A noun is borrowed and orthographically, phonologically and morphologically adapted to Slovak. In the second stage in (2b), a verb motivated by the borrowed noun is formed. In the last stage in (2c), the nature of remotivation is given. The direction of the motivation is reversed, the noun is derived from the verb. Furdík [3] gives two reasons for this remotivation. The first is based on an analogy with the same direction of motivation in native pairs, as illustrated in (3).

(3) čítať → čítanie : read$_{\mathrm{V}}$ → reading$_{\mathrm{N}}$,
strieľať → streľba : shoot$_{\mathrm{V}}$ → shooting$_{\mathrm{N}}$

The examples in (3) demonstrate that the motivating items in native pairs of a verb and action noun are clearly the verbs. On the same basis, Furdík [3] suggests analogical application of the direction of motivation to pairs with international elements with the suffix *-ácia*. As a second reason he mentions that, by accepting this direction of motivatedness, the degree of motivation of Slovak vocabulary is not reduced by the borrowing of nouns in *-ácia*.

It is interesting to observe how these theoretical assumptions are reflected, for instance, in lexicographic practice. The *Concise Etymological Slovak Dictionary* (CESD) by Králik [8] lists mostly verbs such as *asimilovať* 'assimilate', *deportovať* 'deport', *demilitarizovať* 'demilitarise', *devastovať* 'devastate', but not the corresponding nouns *asimilácia* 'assimilation', *deportácia* 'deportation', *demilitarizácia* 'demilitarisation', *devastácia* 'devastation'. This can be explained by the fact that the verbs are considered to be the motivating words for noun formation in such pairs.

In this paper I argue that in Slovak, international nouns with *-ácia* serve as the basis for verb formation. Different types of evidence can be brought to support this argumentation line. The most straightforward evidence would come from the date of first attestation. However, in Slovak dictionaries, unlike, for instance, in the Oxford English Dictionary (OED) this information is in general not available. This means it cannot be used in the case of Slovak data. I will discuss two other types of evidence in the sections below. The first type is based on frequency, the second on meaning.

2 Frequency in the Slovak National Corpus as evidence

Sambor [15] and Furdík [3] showed that, in a pair of a motivating and a motivated word, the motivated word tends to be the word with the lower frequency. Their research was based on a corpus of one million words. My research is based on the Slovak National Corpus (SNC), which has a size of 972 million words. SNC provides several types of frequency lists including a frequency list of lemmas and a frequency list of word forms based on wordclass. I used a full frequency list of lemmas for nouns and verbs.

First I looked at the frequencies of native *verb → noun* pairs with the suffix *-anie*. The suffix *-anie* is a competing native suffix of the Latin-based suffix -*ácia*. Frequencies for the ten most frequent native Slovak verbs in SNC for which a noun in *-anie* or *-enie* is attested and for their corresponding nouns are given in Table 1.

verb	absolute frequency	noun	absolute frequency	absol. freq. of verb /absol. freq. of noun
hovoriť 'speak'	901080	hovorenie	1097	821.4
vidieť 'see'	626194	videnie	16069	38.97
myslieť 'think'	568041	myslenie	47490	11.96
hrať 'play'	502558	hranie	10820	46.45

verb	absolute frequency	noun	absolute frequency	absol. freq. of verb /absol. freq. of noun
čakať 'wait'	383019	čakanie	16074	23.83
vrátiť 'return'	370408	vrátenie	11654	31.78
viesť 'lead'	342718	vedenie	311525	1.10
dodať 'supply'	322201	dodanie	5354	60.18
tvrdiť 'claim'	320130	tvrdenie	48160	6.64
pokračovať 'continue'	308067	pokračovanie	48797	6.31

Table 1: Native Slovak verbs and corresponding deverbal nouns

Table 1 clearly demonstrates that these verbs are significantly more frequent than the nouns derived from them. As shown in the last column, for *hovoriť* 'speak' the verb is even more than 800 times as frequent as the noun. This is fully in line with Sambor's [15] and Furdík's [3] finding that motivated words tend to be of a lower frequency. The frequency values in Table 1 support the hypothesis that motivating words, in this case verbs, are more frequent than the motivated nouns.

It is interesting to compare these results with some observations about English examples of backformation such as *edit* derived from *editor*. Bauer et al. [1] note that the longer forms have higher frequencies than the back-formed ones. This means that also in this case, motivating units tend to be more frequent than motivated ones. Therefore, if the nouns with the suffix *-ácia* are motivated by the corresponding verbs, the verbs are predicted to be more frequent.

SNC includes nearly 8000 nouns formed by *-ácia/-izácia/-fikácia*. The frequencies of the top 20 nouns with their corresponding verbs are in Table 2.

verb	absolute frequency	noun	absolute frequency	absol. freq. of verb /absol. freq. of noun
informovať 'inform'	310355	informácia	439292	0.71
organizovať 'organise'	68424	organizácia	253282	0.27
komunikovať 'communicate'	31899	komunikácia	120876	0.26
reprezentovať 'represent'	43729	reprezentácia	94999	0.46
operovať 'operate'	11872	operácia	82363	0.14

verb	absolute frequency	noun	absolute frequency	absol. freq. of verb /absol. freq. of noun
realizovať ‘realize’	70606	realizácia	67953	1.04
dotovať ‘dotate’	4803	dotácia	67435	0.07
kvalifikovať ‘qualify’	7319	kvalifikácia	60520	0.12
asociovať ‘associate’	769	asociácia	53313	0.01
privatizovať ‘privatise’	4044	privatizácia	52966	0.08
kombinovať ‘combine’	11013	kombinácia	51074	0.22
prezentovať ‘present’	69600	prezentácia	50958	1.37
aplikovať ‘apply’	14469	aplikácia	49748	0.29
publikovať ‘publish’	16646	publikácia	42081	0.40
nominovať ‘nominate’	11428	nominácia	40810	0.28
motivovať ‘motivate’	18027	motivácia	34157	0.53
orientovať ‘orientate’	21488	orientácia	33419	0.64
integrovať ‘integrate’	5804	integrácia	30464	0.19
likvidovať ‘liquidate’	9378	likvidácia	28363	0.33
interpretovať ‘interpret’	11341	interpretácia	27761	0.41

Table 2: International verbs of Latin origin and corresponding deverbal nouns

In Table 2 the frequency values contrast sharply with the frequencies in Table 1. If the noun behaves in the same as the noun in *-enie*, we expect a range of relative frequencies as in Table 1. However, in Table 2, the nouns in *-ácia* have in most cases higher frequency scores than the corresponding verbs, resulting in a score below 1 in the final column. In the few cases where the verb is more frequent, the score is just over 1. Higher frequencies of nouns indicate the reverse direction of motivation is more likely. Nouns are the motivating lexical items for the motivated verbs. The cases in Table 2 are not a random sample, but they show a strong tendency. Some more examples with lower absolute frequencies in SNC are given in Table 3.

verb	absolute frequency	noun	absolute frequency	absol. freq. of verb /absol. freq. of noun
sebalikvidovať ‘selfliquidate’	0	sebalikvidácia	35	0
redislokovať ‘redislocate’	4	redislokácia	35	0.11
katolizovať ‘catholise’	6	katolizácia	35	0.17
intimizovať ‘intimise’	9	intimizácia	35	0.26
elaborovať ‘elaborate’	4	elaborácia	35	0.11
efektivizovať ‘effectivise’	3	efektivizácia	35	0.09
cyklizovať ‘cyclise’	3	cyklizácia	35	0.09
prioritizovať ‘prioritise’	13	prioritizácia	34	0.38
palatalizovať ‘palatalise’	0	palatalizácia	34	0
nukleovať ‘nucleate’	2	nukleácia	34	0.06

Table 3: International verbs of Latin origin and corresponding deverbal nouns with low absolute frequencies

In Table 3, we see that also for nouns in *-ácia* with lower frequencies, the absolute values are higher than the absolute scores of the corresponding verbs. For the nouns *sebalikvidácia* ‘self-liquidation’ and *palatalizácia* ‘palatalisation’ no corresponding verbs occur in the corpus.

Using the Excel feature RANDOM, I extracted a randomized sample of nouns in *-ácia* with a frequency of more than 30 from the full frequency list. Then the verbs and their frequencies were added. The frequency comparison in the randomised sample is summarized in Figure 1.

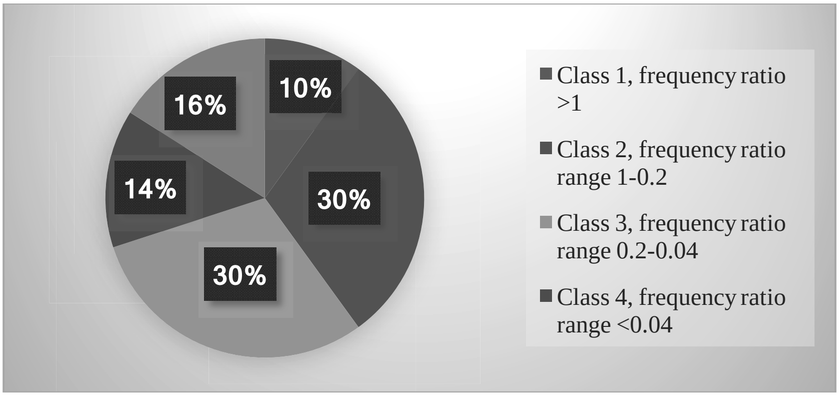


Figure 1: Frequencies of nouns in *-ácia* and corresponding verbs

Figure 1 shows five classes based on the ratio of absolute frequency of verbs and absolute frequency of nouns. In Class 1 the ratio is higher than 1, which means that the frequency of the verbs is higher than the frequency of the corresponding nouns. The higher the ratio value, the greater the difference in the frequency values between verbs and nouns.

In the remaining classes, the frequency of nouns is higher. Class 2 and class 3 are proportionally equal categories. In Class 2 in many cases the absolute frequency of the verb is less than half of the absolute frequency of the noun, e.g. *inštitucionalizácia* 'institutionalisation' occurs 586 times in SNC whereas the verb *inštitucionalizovať* 'institutionalise' only 262 times with the ratio 0.44. In Class 3 the absolute frequency of the verb tends to be much lower than the absolute frequency of the noun. This can be illustrated by *demokratizácia* 'democratisation' with the absolute frequency 3084 and the absolute frequency 276 of the verb *demokratizovať* 'democratise' resulting in a ratio of 0.08. Class 4 is approximately half the size of Class 2 and Class 3. Similarly, there is a tendency for much higher absolute frequency scores of the nouns in *-ácia* than their corresponding verbs. In Class 4, the frequency of the verbs is almost negligeable compared to that of the nouns.

In addition, this evidence is supported by 16% of the nouns without a corresponding verb in the corpus in Class 5. These include, for instance *trunkácia* 'truncation' but not the verb *trunkovať* 'truncate' or similarly *peroxidácia* 'peroxidation' but not *peroxidovať* 'peroxidate'. Although these verbs do not occur in SNC, they are used in scientific contexts and sometimes also in other contexts. A Google search gives 6 hits for *trunkovať* 'truncate', all in research paper in linguistics, and 25 hits for *peroxidovať* 'peroxidate', both in scientific and less formal contexts.[2]

The data in Fig. 1 provide strong evidence for the tendency observed in Tables 1-3. Whereas for native Slovak *-enie*, the relative frequency of nouns and verbs corresponds to what is expected when the noun is derived from the verb, for *-ácia* the opposite frequency distribution is found.

Another interesting example of a noun in *-ácia* without a derived verb is *biodegradácia* 'biodegradation'. SNC does not list the verb *biodegradovať* 'biodegrade'. Similarly to the examples above, the Google search gives several hits of this verb mostly in academic and environmental context. The most recent *Dictionary of foreign words (academic)* [17] in Slovak does not include an entry for the noun *biodegradácia* 'biodegradation' or the verb *biodegradovať* 'biodegrade'. It is well known that at present many internationalisms enter Slovak (and other languages) from English. Therefore it is useful to compare the situation in Slovak with the information given in the *Oxford English dictionary* (OED [13]), summarized in (4).

(4) a. biodegradation, n. 1941.

[2] Retrieved 11 August, 2017

Origin: Formed within English, by compounding.
Etymology: < bio- comb.form + degradation n.1 Compare later biodegrade v.

b. biodegrade, v. 1961
Origin: Formed within English, by compounding.
Etymology: < bio- comb.form + degrade v. After biodegradation n.

The information about the entries in (4) includes the date of attestation, origin and etymology. A comparison of (4a) and (4b) shows that the noun in (4a) was formed in English and probably earlier than the verb in (4b). The difference between the dates of attestation in (4a) and (4b) is twenty years. Given the fact that the 20th century is well documented by OED, the time difference can be seen as evidence that verb was backformed from the noun. OED (2017) also gives information about the frequency of current use. The noun in (4a) is in Frequency band 4.[3] This means this word may not necessarily be used on daily basis but its meaning will not present a problem for most speakers of English. For the verb in (4b) the Frequency band is 3.[4]

The frequency of occurrence of the Slovak internationalism *biodegradácia* in SNC is 34, which is 0.03 per million. The verb *biodegradovať* does not occur in the corpus but it can be found on Google. The data suggest that *biodegradácia* is likely to be borrowed from English. The verb *biodegradovať* was formed later and is obviously used in scientific contexts, but it is still not used frequently enough to be included in SNC. However, it can be expected that after some time, when new, especially academic texts are added to SNC, the verb *biodegradovať* will also appear there. This may be viewed as another piece of evidence for the claim that in Slovak international nouns with -*ácia* serve as the basis for verb formations.

3 Meaning as evidence

Let us now turn to the semantic relation between the nouns and the verbs. As mentioned above, the basic meaning of the word formation type [international base + *-ácia/-cia, -izácia, -fikácia*] is action, or homogeneous process (Horecký et al., [6]). This is illustrated in (5).

(5) a. nacionalizácia ‘nationalisation’

[3] Frequency Band 4 contains words which occur between 0.1 and 1.0 times per million words in typical modern English usage. Such words are marked by much greater specificity and a wider range of register, regionality, and subject domain than those found in bands 8-5. However, most words remain recognizable to English-speakers, and are likely be used unproblematically in fiction or journalism. (OED [13]).

[4] Frequency Band 3 contains words which occur between 0.01 and 0.1 times per million words in typical modern English usage. These words are not commonly found in general text types like novels and newspapers, but at the same they are not overly opaque or obscure. (OED [13]).

prevod, prevzatie súkromných podnikov do štátnej správy al. do štátneho vlastníctva
‘transfer, taking over of private enterprises to national administration or to the ownership of the state’

b. modernizácia ‘modernisation’
prispôsobovanie, prispôsobenie novej dobe, móde, novým požiadavkám
‘adaptation, adjustment to new era, fashion, new requirements’

c. špecifikácia ‘specification’
bližšie určenie, vymedzenie niečoho s uvedením podrobností, presných, rozlišujúcich údajov
‘closer determination, delimitation of something with mentioning details, precise, distinctive data’

The examples in (5) and their definitions are taken from the *Dictionary of foreign words (academic)* [17] in Slovak. The definitions demonstrate that the nouns with *-ácia* usually denote a process. It seems interesting to compare the meaning of the nouns in (5) with the meaning of the corresponding verbs in (6).

(6) a. nacionalizovať ‘nationalise’
uskutočňovať, uskutočniť nacionalizáciu
‘carry out nationalisation’

b. modernizovať ‘modernise’
uskutočňovať, uskutočniť modernizáciu
‘carry out, perform, undergo modernisation’

c. špecifikovať ‘specify’
(u)robiť špecifikáciu
‘carry out, perform, undergo specification’

In the examples in (6) we can see that the meaning of the nouns in (5) is typically included in the meaning of the verb. For backformations in English, Nagano [12] views inclusion of the meaning of the noun in the meaning of the verb as a relevant proof. The same can be applied to the Slovak cases in (5) and (6).

Semantic evidence plays an important role in determining the direction in the case of conversion, especially in English. In this context, Plag [14] applies a parallel reasoning of a general assumption that derived words are semantically more complex. This means that the derived or converted word “should be semantically more complex than the base word from which it is derived” (Plag, [14]). This is illustrated in (7).

(7) a. house$_N$ – a building for habitation, and related senses (OED, [13])
b. house$_V$ – to take, receive, or put into a house (OED, [13])

The example in (7) is a case of a noun to verb conversion. The verb in (7b) is semantically richer than the noun in (7a). In addition, the meaning interpretation of (7b) is dependent on (7a). Conceptually, the verb in (7b) requires the existence of the noun in (7a). In such cases, in line with Plag [14] "we have good evidence that the dependent member is derived from the other form". A similar semantic parallel can be observed between the nouns in *-ácia* in (5) and the corresponding verbs in (6). Obviously the verbs in (6) are semantically more complex than the nouns in (5) and the interpretation of the verbal meaning includes and depends on the meaning of the noun. This means that this type of semantic evidence should not be overlooked in determining the direction of motivation of the internationalisms in *-ácia*.

4 Conclusion

In this paper I investigated whether the direction of motivation in pairs of mostly Latin origin such as *integrovať* 'integrate' → *integrácia* 'integration' can be supported by data taken from the Slovak National Corpus. The complete sample includes nearly 8000 nouns. The results indicate that the direction is rather opposite to what can be found for native pairs of deverbal nouns, e.g. *čítať* 'read'→ *čítanie* 'reading$_N$', which were traditionally considered as a model for analogy. In native pairs, motivating verbs usually have higher frequency scores than motivated nouns. In contrast, in pairs with nouns of Latin origin and the suffix *-ácia*, the verbs display lower frequency scores. Another type of evidence is semantic. In many cases, the meaning of the noun is included in the meaning of the verb. Therefore, there is strong evidence that in Slovak, nouns in *-ácia* are morphologically prior to the corresponding verbs.

References

[1] Bauer, Laurie, Lieber, Rochelle, and Ingo Plag. *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press, 2015.

[2] Buzássyová, Klára. Vzťah internacionálnych a domácich slov v premenách času. [Relation between International and Original Words in a Metamorphosis of Time.] *Jazykovedný časopis*, Vol. 61, No. 2, pp. 113–130, 2010.

[3] Furdík, Juraj. Slovotvorná motivovanosť slovnej zásoby v slovenčine. [Word formation motivatedness in Slovak lexis.] In: Mistrík Jozef (ed.) *Studia Academica Slovaca. 7. Prednášky XIV. letného seminára slovenského jazyka a kultúry*. Bratislava: Alfa, pp. 103–115, 1978.

[4] Gutschmidt, Karl. Der Begriff der Tendenz in der slawischen Sprachen. [The concept of tendency in Slavic languages.] In: Gladrow, W. (Ed.). *Das Russische in seiner Geschichte, Gegenwart und Literatur.*. Műnchen: Sagner, pp. 52 – 69, 1995.

[5] Gutschmidt, Karl. Tipologični tendencii. [Typological tendencies.] In: Ohnheiser Ingeborg (Ed.) *Komparacja systemów i funkcjonowania współczesnych języków słowiańskich. 1. Słowotwórstwo/Nominacja*. Innsbruck/Opole: Universität Innsbruck – Institut fűr Slawistik, Uniwersytet Opolski – Institut Filologii Polskiej, pp. 341 – 355, 2003.

[6] Horecký, Ján, Buzássyová, Klára, Bosák, Ján. a kol. *Dynamika slovnej zásoby súčasnej slovenčiny*. [Dynamics of contemporary Slovak lexis.] Bratislava: Vydavateľstvo Slovenskej akadémie vied, 1989.

[7] Jiráček, Jiří. *Adjektíva s internacionálními sufixálními morfy v současné ruštině (v porovnání s češtinou)*. [Adjectives with international suffixal morphs in contemporary Russian (in comparison with Czech).] Brno: Univerzita J. E. Purkyně v Brně, 1984.

[8] Králik, Ľubor. *Stručný etymologický slovník slovenčiny*. [Concise etymological dictionary of Slovak.]. Bratislava: Veda, 2015.

[9] Mistrík, Jozef, *Frekvencia slov v slovenčine*. [Frequency of words in Slovak] Bratislava: Vydavateľstvo SAV, 1969.

[10] Mistrík, Jozef. *Retrográdny slovník slovenčiny*. [Retrograde dictionary of Slovak.] Bratislava: Univerzita Komenského, 1976.

[11] Mistrík, Jozef. *Frekvencia tvarov a konštrukcií v slovenčine*. [Frequency of word-forms and constructions in Slovak] Bratislava: Veda, 1985.

[12] Nagano, Akiko. *Conversion and back-formation in English: Toward a Theory of morpheme-based morphology*. Tokyo: Kaitakusha, 2008.

[13] OED *Oxford English Dictionary*, Third edition, edited by John Simpson, www.oed.com, 2017.

[14] Plag, Ingo. *Word-formation in English*. Cambridge: Cambridge University Press, 2003.

[15] Sambor, Jadwiga. *O słownictwie statystycznie rzadkim*. [About statistically rare vocabulary.] Wyd. 1. Warszawa, 1975.

[16] *Slovenský národný korpus* – prim-7.0-public-all. [Slovak national corpus – prim-7.0-public-all.] Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2015. Available at WWW: http://korpus.juls.savba.sk.

[17] *Slovník cudzích slov (akademický)*. [Dictionary of foreign words (academic)]. 2., doplnené a prepracované vyd. Spracoval kolektív autorov pod

vedením V. Petráčkovej a J. Krausa. Preklad Ľ. Balážová, J. Bosák, J. Genzor, I. Ripka, J. Skladaná. Ed. Ľ. Balážová – J. Bosák. Bratislava: Slovenské pedagogické nakladateľstvo – Mladé letá, 2005.

Evaluating and Improving a Derivational Lexicon with Graph-theoretical Methods

Sean Papay, Gabriella Lapesa, and Sebastian Padó

Institute for Natural Language Processing, Stuttgart University
E-mail: name.surname@ims.uni-stuttgart.de

Abstract

We employ a graph-theoretical approach to evaluate and improve a German derivational lexicon, DERIVBASE. We represent derivational families (that is, groups of derivationally related words) as labelled directed graphs in which words (*friend*, *friendly*) are nodes and derivational relationships (*friend* → *friendly*) between words are directed edges, labeled with the derivation rule (*-ly*).

This graph-theoretical approach allows us to carry out a large-scale comparison of the structure of different derivational families and identify, in a completely automatic fashion, possible errors in the resource. We conduct a manual evaluation of the predictions of our method and find that it successfully identifies instances which are missing from DERIVBASE; the predictions of our approach can be interpreted as the result of interplay among productivity constraints.

1 Introduction

Derivational lexicons encode knowledge about derivational relations between words. Minimally, they group lemmas into derivational families, but optionally provide additional information, such as semantic transparency, morphological structure, or instantiation of specific derivational rules. Examples include CELEX for English, German and Dutch (Baayen et al. [1]), CatVar for English (Habash and Dorr, [3]), DERIVBASE for German (Zeller et al. [13]), DERIVBASE.HR for Croatian (Šnajder [11]), Démonette for French (Hathout and Namer [4]), and DeriNet (Žabokrtský et al. [12]) for Czech. Derivational lexicons are employed in NLP applications (Shnarch et al. [9], Padó et al. [7]) and can serve for the selection of the experimental items in psycholinguistic experiments and corpus-based modeling (Smolka et al. [10], Padó et al. [8]). In particular when extracted automatically or semi-automatically, they enable large-scale investigations of the structure of the underlying morphological systems (Lazaridou et al. [5], Padó et al. [6]). At the same time, (semi-)automatically constructed derivational lexicons cannot guarantee completeness: any resource is likely to both miss some instances of derivational relations and to contain spurious instances. It is therefore crucial to properly evaluate them and, ideally, improve them by both removing incorrect derivations and filling in missing derivations.

In this paper, we introduce a graph-theoretical approach for the targeted evaluation and improvement of derivational lexicons. We apply our method to DERIVBASE (Zeller et al. [13]), a high-coverage German derivational lexicon. Our approach is however applicable to any derivational lexicon that can be interpreted as a graph with lemmas as nodes and derivational relations as labeled edges.

Our method is centered around the concept of a *fingerprint* of a derivational family, a structure which represents morphological connections between words, while abstracting away individual words. Our central assumption in this paper is that if the fingerprints of two families are shared *almost, but not completely*, this is a strong indication that (at least) one of the two families is incorrect. We further hypothesize that the decision of which of the families is correct can again be made automatically on the basis of *frequency* information: If one family misses a node that is present in a large number of families, this is an indicator of a false negative (missing family member). Conversely, a rare surplus node that a family adds to a frequent fingerprint indicates a false positive (spurious family member). We discuss below to what extent these assumptions are warranted.

2 Data

DERIVBASE is a derivational lexicon for German (Zeller et al. [13]). It is based on a set of 158 finite state rules describing German derivation patterns (including prefixation, suffixation, stem changes, and combinations thereof). The rules were hand-crafted to maximize coverage and minimize errors on a development set.

DERIVBASE forms a large directed graph. Its nodes are the 280k lemmas that occur in SdeWaC (Faaß and Eckart [2]) with a frequency of four or more. They are annotated automatically with part-of-speech and gender information. Edges connect derivationally related words, and each edge is labeled with one of the rules. The edges group the 280k nodes into 20k non-singleton derivational families, and 220k singleton families.[1] DERIVBASE edges are created whenever a word pair in SdeWaC matched a rule; edges therefore express morphological (but not necessarily semantic) relatedness. Even at the morphological level, though, errors arise from the fully automatic construction of the resource. DERIVBASE was evaluated against a small manually annotated sample in (Zeller et al. [13]) and was found to have a precision of 83% and recall of 71%. The imperfect precision results from false positives, that is, spurious edges that arise from chance matches (e.g., *Celle* (German town) → *Cellist* (cello player)). The imperfect recall indicates missing edges, which

[1] The high number of singleton families is due to the prevalence of compounding in German. As DERIVBASE does not group compounds together with their bases, and compounds typically exhibit less derivation than the bases, these compounds tend to form singleton families.

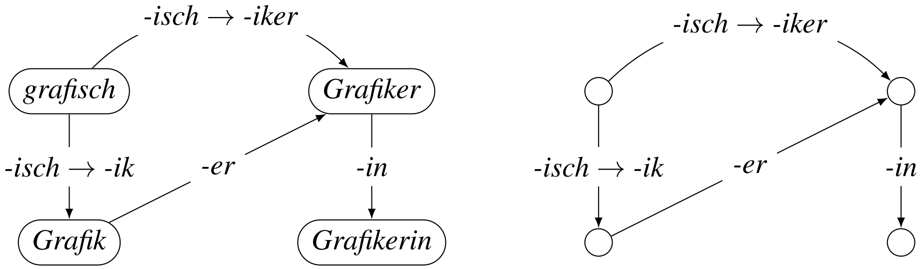


Figure 1: Illustration of a German derivational family (left) and its fingerprint (right)

are due to a range of factors, including lemmatization problems, words being too infrequent, or simply orthographic variation that was overlooked in the formulation of the rules.

3 Method

We begin by finding the *fingerprints* of the families in DERIVBASE. A family's fingerprint is a representation of the derivational relationships within a family, which abstracts away information about individual words. This can best be understood in the context of graphs – if a family is taken as a directed graph as described in Section 2, its fingerprint is simply the same graph with all node labels removed. Figure 1 illustrates the derivational family of the word *Grafik*, and that family's fingerprint. Two families which undergo the same patterns of derivation will have the same fingerprint. For example, the family above shares its fingerprint with the families {*Musik*, *musisch*, *Musiker*, *Musikerin*} and {*Tragik*, *tragisch*, *Tragiker*, *Tragikerin*}, among many others. Mathematically, two families will share their fingerprint if and only if their graphs are *isomorphic*.

The 20k non-singleton families of DERIVBASE were grouped into equivalence classes, with families grouped together if and only if they shared a fingerprint. As the database contained 4539 distinct fingerprints, 4539 such classes were constructed, with an average of 4.5 families per class. Families' fingerprints were compared by checking for graph isomorphism.[2]

As motivated in Section 1, our hypothesis is that the (semi-)regularity of morphology leads to *consistency* across derivational families: the structures of any two families should either be identical or show *major* differences; conversely, *minor* differences are indicators of mistakes. While there are a number of potential ways to operationalize what counts as a minor difference, in this paper we focus on one type of difference, namely the presence or absence of exactly one node, respectively. Formally, this corresponds to the concept of *induced subgraphs*.

[2] We used the Python3 package `networkx` for all graph-theoretical operations. While no polynomial-time algorithm is known for the problem of graph isomorphism, the general small size of derivational families made asymptotic complexity largely irrelevant.

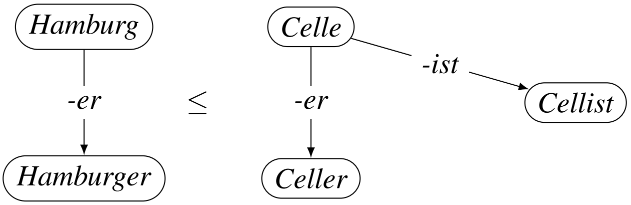


Figure 2: The left family is an induced subgraph of the right one: it is isomorphic to the right family sans *Cellist*.

An induced subgraph G' of a graph G is obtained by removing one or more nodes from G and removing all edges adjacent to the removed nodes. Our procedure is therefore as follows. We consider all pairs of fingerprints (F_1, F_2) where F_2 is an induced subgraph of F_1 such that $||V(F_2)|| = ||V(F_1)|| - 1$, that is, they differ in one node. We call these pairs of fingerprints our *error candidates*. Our linguistic interpretation of the pairs in this set is determined by the ratio of the number of derivational families in the F_1 and F_2 equivalence classes, respectively. Our concrete hypotheses are as follows:

1. If the larger fingerprint was found for many more families than the smaller one, the smaller one is very likely to be incomplete: this is a false negative.
2. If, conversely, the smaller fingerprint was found more often than the larger one, the larger one is likely to contain an incorrect node: this is a false positive.
3. When both fingerprints occur roughly equally often, we cannot make a judgment, and they may be equally valid.

Figure 2 illustrates this on a concrete example of a family (right) and an induced subfamily with one node less (left). If the fingerprint of the right-hand family were much more frequent, we would (incorrectly) infer that the left-hand family were missing the node **Hamburgist*. However, since the fingerprint of the left-hand family is in fact much more frequent, we can (correctly) infer that *Cellist* is a spurious member of this family.

This method has a number of convenient properties. In contrast to other error detection methods, it does not compare individual families, but equivalence classes of families. As a result, it can take consistency across families in account. In addition, due to the isomorphism underlying the induced subgraph relation, the method can pinpoint exactly where in the family there is a potential gap (or spurious node, respectively) and which derivation rule is responsible. Note that we do not consider the prediction of a concrete surface form for a missing node. In the case of DERIVBASE, this would be possible by applying the morphological transformation that the resource associates with each derivation rule. However, since these transformations typically overgenerate, this would require a disambiguation setup that goes beyond the focus of this paper. At any rate, during our manual evaluation (described in Section 4), we found that native annotators have no trouble whatsoever judging the appropriateness of proposed derivations even without a

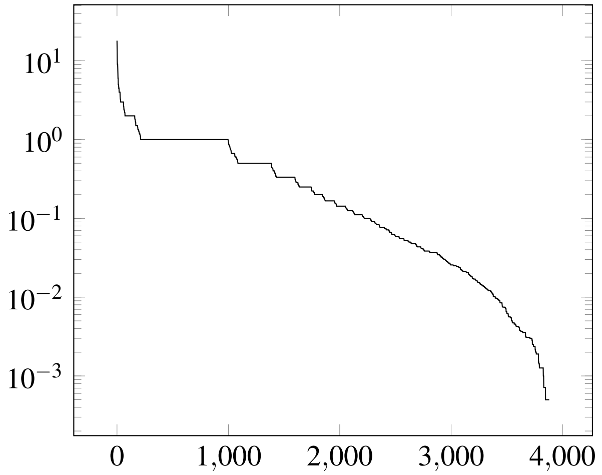


Figure 3: The ratio of the number of families for each error candidate, plotted by list index. Ratios are plotted on a logarithmic scale, so as to better illustrate differences in ratios which lie very close to zero.

concrete surface form proposal.

In closing, we note that there is reason to believe that there is an assymetry between cases (1) and (2) that is due to the *semi*-regularity of derivational morphology. While some derivational rules are applicable almost universally within their domain (e.g., almost all verbs can be nominalized), other rules apply only to very specific semantic classes (e.g., nationalities: *Schweden* → *Schwede*, *Polen* → *Pole* etc.). Thus, the *absence* of a frequent node from a family (as in (1)) is presumably a more reliable indicator than the *presence* of a rare node in a family (as in (2)). Fortunately, the evaluation numbers for DERIVBASE reported above indicate that false negatives, which are found by (1), are also a larger problem in practice than false positives.

4 Annotation

When we applied the fingerprint computation and comparison method to DERIVBASE, we obtained 2471 fingerprints and 3882 error candidates. We ranked the error candidates by the ratio of the number of participating families. The ratio is 18 : 1 for the top-ranked error candidate, and 1 : 2005 for the bottom-ranked error candidate. Figure 3 shows how these ratios vary with list position.

Since a full annotation of all error candidates was impractical, we extracted the top and bottom 250 error candidates, since these should be most interesting according to our hypotheses. For each class present in these error candidates, we selected one family at random to represent that class. In order to avoid annotator biases about predominant case types at the top and bottom of the list, we shuffled these 500 error candidates. In addition, candidates from both samples had to be presented in exactly the same form. We chose an “analogy-style” presentation as follows:

`[LHS-1] [rule]→ [LHS-2] :: [RHS-1] [rule]→ ???`

In these analogies, `LHS-2` is the word in the larger family which has no corresponding node in the smaller family. `LHS-1` and `rule` are populated with values from some edge adjacent to `LHS-2`.[3] `RHS-1` is the word in the smaller family which corresponds to `LHS-1`. We will use the name `RHS-2` to describe the hypothetical word, which might exist in the place of (`???`) in the analogy.

A native speaker with graduate-level knowledge in linguistics was presented with the 500 analogies and asked to categorize each analogy according to the following schema:

FN is the false-negative case, where `RHS-2` is correct but missing from the resource. According to our hypothesis (1), these cases should predominate at the top of the sorted candidate list.

FP is the corresponding false-positive case, where `LHS-2` is not a derivation of `LHS-1` even though it is present in the resource. According to our hypothesis (2), these cases should predominate at the bottom of the sorted candidate list.

OK is the case where the left-hand derivation is correct but the right-hand derivation is not. This corresponds to cases in which DERIVBASE was correct as-is, and no error was present to be identified. We expect theses cases to be rare, since they run counter to our assumption that “small differences” between fingerprints are generally errors.

LER, RER are cases where linguistic preprocessing (lemmatization or gender determination) failed either on the left-hand side or the right-hand side, respectively.

Table 1 shows examples for each of these categories.

5 Results

The main results are shown in Table 2. We first discuss the percentage of the annotation labels in the top-250 and bottom-250 lists shown in the first two rows.

The Top-250 candidates. In this list, false negatives (FN, gaps in the resource) account for 79% of the error candidate pairs. This is a very strong confirmation of our hypothesis (1) from above: almost 80% of the instances that our method

[3] We attempt to choose an edge at random which points towards `LHS-2`. If no such edge exists, we select a random edge which points away from `LHS-2`. In these cases, `rule` was marked with an asterisk, to notate the reversed direction of derivation.

Tag	Definition	LHS-1	LHS-2	RHS-1	(RHS-2)
FN	RHS-2 valid derivation for RHS-1	Ehrenbürger honorary citizen (m.)	Ehrenbürgerin honorary citizen (f.)	Einzeltäter lone offender (m.)	Einzeltäterin lone offender (f.)
FP	words on LHS unrelated	pazifisch pacific	Pazifismus pacifism	ökosozial eco-social	Ökosozialismus eco-socialism
LER	preprocessing error on LHS	niedersächsisch low saxonian	*Niedersachs	westfälisch westphalian	N/A
OK	RHS-2 not a derivation of RHS-1	Unterwanderung subversion	unterwandert subverted	Bergwanderung mountain tour	*bergwandert
RER	preprocessing error on RHS	Dusel fluke	duselig flukey	*Hark	N/A

Table 1: Annotation categories and examples (RHS-2 as determined by annotator)

	FN	FP	LER	OK	RER
percentage in top 250	**78.8**	1.2	3.2	14.4	2.4
percentage in bottom 250	8.0	4.4	8.8	**78.8**	0.0
Pearson's *r* with list rank	-0.6432	0.0920	0.1384	0.5720	-0.0900
p-values	<0.0001	0.04	0.002	<0.0001	0.04

Table 2: Results: Tag frequency and correlation with list rank

identifies as gaps in DERIVBASE are indeed gaps. Of the rest, only 1% is due to erroneous entries in DERIVBASE, some 5% are due to preprocessing errors (lemmatization and gender detection), and 14% are cases where the small difference is actually correct. To illustrate this category, consider

(1) *Geschäftspartner* dNN02→ *Geschäftspartnerin* :: *Ort* dNN02→ ???
business partner (m.) dNN02→ business partner (f.) :: place dNN02→ ???

where dNN02 is the rule deriving a female from a male profession or role noun, which is appropriate for LHS-1 (*business partner*) but not for RHS-1 (*place*), which belongs to another semantic category. The next example,

(2) *abschieben* dVN07→ *Abschiebung* :: *anfliegen* dVN07→ ???
to deport dVN07→ deportation :: to approach dVN07→ ???

arises from the fact that German has several nominalization patterns, including the *−ung* suffix (dVN07 in DERIVBASE), which is however not applicable to all verbs. Thus, for *anfliegen* the derivation **Anfliegung* is not attested; instead, the stem nominalization *Anflug* (dNV09) is used. These examples illustrate two limits of our current schema: (a) the derivation rules do not take semantic classes into account that affect their applicability; (b) the fingerprint comparison does not take relations among derivation rules into account.

The Bottom-250 candidates. The bottom-250 candidate list shows a very different picture. According to our hypothesis (2), we would expect the majority of analogies to fall into category FP/false positives: cases where the existing (LHS) derivation relation is incorrect. This however turns out to be true for only some 4% of all cases, a lower percentage than even the false negatives (FN, 8%) and preprocessing errors (LER+RER, 8.8%) account for. The majority of bottom candidates actually consists of cases where the (rare) LHS is a valid and the (frequent) RHS an invalid derivation.[4] In other words, the bottom end of the error candidate list consists of edges that are rather rare, but still valid, and which can *not* be generalized to other families.

A qualitative analysis of the OK cases found that about 80% of them could be grouped into three main classes. The largest class, accounting for about 40%, consisted of borderline derivation/composition instances like

(3) *Wehrdienstleistende* dNN46.1→ *Grundwehrdienstleistende* ::
conscript dNN46.1→ conscript in basic training ::
Nächstenliebe dNN46.1→ ???
altruism dNN46.1→ ???

where the prefix *Grund- 'basic'* is only applicable to a very specific set of base nouns, and **Grundnächstenliebe* does not exist.

The second class (20%) was composed of cases of morphological alternatives (e.g. multiple nominalization rules) similar to those we found for the top-250 candidates. The third class (20%) concerned a specific problem in German morphology, namely prefix verbs. These behave in many respects like base verbs, but not with regard to further prefixation:

(4) *stöpseln* dVV22.2→ *einstöpseln* :: *errechnen* dVV22.2→ ???
to plug dVV22.2→ to plug in :: to compute dVV22.2→ ???

Here, the prefix verb *errechnen* cannot serve as a base to derive **einerrechnen*, while this is possible for its base verb *rechnen > einrechnen / to calculate > to include*.

These observations support and strengthen our caveat from above regarding the *semi*-regularity of derivational morphology, even though to a considerably more extreme degree that we initially assumed.

Correlation Analysis. A correlation analysis, shown in the lower half of Table 2, bolsters this picture. We compute the Pearson correlation *r* between the occurrence of the different categories and the rank in the list.[5] We find that there is an extremely strong negative correlation for FN, that is, false negatives occur overwhelmingly towards the top of the list. There is an almost equally strong positive correlation for OK, that is, idiosyncratic yet valid edges tend strongly to occur towards the end

[4] The fact that the percentages of Y for top-250 and NN for bottom-250 are identical is purely coincidental.

[5] We use the ranks of entries in the original list of 3882 error candidates, not the ranks in our list of 500 annotated entries

of the list. As the p-values show, the values for the remaining categories (FP, LER, RER) are also significant, but considerably less so. We conclude that preprocessing errors and false positives tend to occur towards the end of the list, but much less strongly so.

6 Discussion and Conclusion

We have presented a graph-theoretical method to evaluate derivational lexicons; through a manual classification of the predictions of our model on a German lexicon, DERIVBASE, we have shown that we can predict with high confidence those cases where possible derived words are missing from the resource. Our predictions concerning spurious words in the resource turned out to be less strikingly correct, and current work targets a better understanding of our treatment of false positives. A further potential improvement of our method is the identification better score to rank the candidates, beyond the simple ratio of the cardinality of the equivalence classes. Future work also targets the automatic integration of the gaps identified through our method.

Acknowledgments

We gratefully acknowledge funding of our research by the DFG, SFB 732 (project B9: Lapesa and Padó).

References

[1] R. Baayen, Richard Piepenbrock, and Léon Gulikers. CELEX2 LDC96L14. *Web Download. Philadelphia: Linguistic Data Consortium*, 1995.

[2] Gertrud Faaß and Kerstin Eckart. Sdewac – a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg, 2013.

[3] Nizar Habash and Bonnie Dorr. A categorial variation database for English. In *Proceedings of NAACL-HLT*, pages 17–23, Edmonton, AL, 2003.

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

[5] Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of ACL*, pages 1517–1526, Sofia, Bulgaria, 2013.

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

[7] Sebastian Padó, Jan Šnajder, and Britta D. Zeller. Derivational smoothing for syntactic distributional semantics. In *Proceedings of ACL*, pages 731–735, Sofia, Bulgaria, 2013.

[8] Sebastian Padó, Britta Zeller, and Jan Šnajder. Morphological priming in German: The word is not enough (or is it?). In *Proceedings of NetWords*, pages 42–45, Pisa, Italy, 2015.

[9] Eyal Shnarch, Jacob Goldberger, and Ido Dagan. A probabilistic modeling framework for lexical entailment. In *Proceedings of ACL/HLT*, pages 558–563, Portland, Oregon, 2011.

[10] Eva Smolka, Katrin H. Preller, and Carsten Eulitz. 'verstehen' ('understand') primes 'stehen' ('stand'): Morphological structure overrides semantic compositionality in the lexical representation of German complex verbs. *Journal of Memory and Language*, 72:16–36, 2014.

[11] Jan Šnajder. Derivbase.hr: A high-coverage derivational morphology resource for Croatian. In *Proceedings of LREC*, Reykjavík, Iceland, 2014.

[12] Zdeněk Žabokrtský, Magda Sevcikova, Milan Straka, Jonáš Vidra, and Adéla Limburská. Merging data resources for inflectional and derivational morphology in Czech. In *Proceedings of LREC*, pages 23–28, Portoroz, Slovenia, 2016.

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

The Question of Proto-Indo-European Inheritance in Latin Word Formation[1]

Lucie Pultrová

Institute of Greek and Latin Studies, Faculty of Arts
Charles University (Prague)
E-mail: Lucie.Pultrova@ff.cuni.cz

Abstract

The paper concerns diachronic word-formation. It argues that the diachronic distinction of being inherited or non-inherited has important implications for the synchronic semantic and formal analysis of word formative types. Moreover, it shows that the (not always trivial) distinction between inherited and non-inherited (= analogical) formations also plays a pivotal role in the description of the phonological system of the language.

1 Introduction

This paper concerns diachronic word-formation, yet, as will be argued, is relevant to research taking a predominantly synchronic approach, such as that adopted for the WFL Project thus far.

I will start with the rather trivial statement that some Latin word-formative types were inherited from Proto-Indo-European (PIE), whereas others originated much later in Latin itself. The claim that a word-formative type was inherited from PIE assumes that the type existed in PIE before the period of its final disintegration. The evidence for this assumption lies in the records of formally and semantically analogous formations in at least a few branches of Indo-European. For example, the Latin "perfect passive participles" ending in *-tus* have their direct or partial equivalents in many Indo-European languages,[2] and thus are evidently inherited, whereas the "future active participles" ending in *-tūrus* have no such equivalents outside Latin, hence we may infer that they must have evolved as late as the period by which Latin was an independent language, and in any case no earlier than the period of Proto-Italic.[3]

[1] This study was supported by the Charles University project Progres 4, *Language in the shiftings of time, space, and culture*.

[2] Cf. OInd. *ta-tás* stretched', Gr. τατός 'stretchable', Lith. *giñtas* 'driven', etc. (Examples taken from Brugmann [1, p. 395].)

[3] The suffix *-(t)ūrus* obviously did not evolve out of nothing; there must have been some base in PIE. Nevertheless, in this form and with this semantics, such a suffix exists only in Latin. The suffix is complex: it is apparently a combination of other suffixes (which PIE suffixes they

The distinction between inherited and non-inherited word-formative types is manifested in both their semantics and form. The more recent (= Latin) formations are, from the synchronic point of view, easy to analyse: a suffix — both formally and semantically transparent — attaches to a clearly defined stem (cf. _mīrā-bilis_, _dēlē-bilis_ [verbal present stem + _-bilis_], _laudāt-ūrus_, _monit-ūrus_ ["supine" stem + _-ūrus_]). Inherited formations, on the other hand, resist synchronic analysis: the form of Latin adjectives ending in _-tus_ is unpredictable; the base to which the suffix (_-tus_ or _-sus_) is attached cannot be easily defined (_lēc-tus_ [× _leg-ō, lēg-ī_], _cēn-sus_ [× _cēns-eō, cēns-uī_], _pāc-tus_ [× _pang-ō, pepig-ī_], etc.); nor is the semantics entirely consistent (_datus_ = pas. 'given' × _cēnātus_ = act./med. 'having dined').

From the diachronic point of view, however, the situation is reversed: inherited formations conform to the phonological (accent-ablaut) system of PIE and to the sound laws of the given language or branch,[4] whereas for Latin neologisms such a phonological analysis at the level of PIE is inapplicable.[5]

All this is clear and obvious. What is less routinely taken into account, however, is that within an inherited word-formative type there often occur formations which in themselves are not inherited, but are analogically formed at a later stage. I shall argue that the distinction between inherited and non-inherited (analogical) formations within a word-formative type plays a major role in correctly interpreting the given word-formative type and, moreover, in specifying the sound laws of the given language.

2 Inherited vs. non-inherited formations in Latin

We shall consider several Latin word-formative types in order to illustrate the distinction between inherited and non-inherited formations both formally and semantically. As we have observed, from the synchronic point of view analogical formations are easy to analyse, whereas inherited formations may be problematic or unpredictable; from a diachronic viewpoint the opposite is true.

2.1 Adjectives ending in _-tus_

are, however, is not clear; cf. e.g. Sihler [13, p. 621] or Leumann [6, p. 618]) that started to live a life of its own in Latin.

[4] Or, better, the aim is that they should conform; if not, then this signals a need for revising the PIE reconstruction and/or sound laws.

[5] For the example _mīrābilis_ mentioned above, the fact alone that the two initial syllables take a form corresponding to the PIE full ablaut grade (_mī-_ < _*(s)méi̯-_, cf. IEW [4, p. 967]; _-r-ā-_ < _*-r-é_h_2_(i̯)-_) makes nonsense of such an analysis. Only primary derivatives can be analysed in terms of the PIE phonological ("accent-ablaut") system. In this case such a primary derivative is the adj. _mīrus_ (< _*(s)méi̯-ro-_); all other derivatives are secondary (_mī-rus_ > _mīr-ā-rī_ > _mīrā-bilis_), created as late as within the Latin phonological (= no longer "accent-ablaut") system.

The PIE structure of the word-formative type we are considering is R(z)-*tó-* (where R(z) is the label for an unaccented root in the ablaut zero-grade); from the roots, e.g., *deh$_3$-[6] 'to give' (Lat. *dare*) or *peh$_2$g- 'to firm' (Lat. *pangere*), the adjectives in question appear to be *dh$_3$-tó-s > Lat. *datus* and *ph$_2$g-tó-s, which yields Lat. *pāctus*.[7] If, however, we take as an example the secondary *laudātus*, referring to the PIE *R(z)-*tó-* makes no sense: the *-tus* is not affixed to the root, but to the secondary stem.[8]

However, there are a number of less trivial examples. Many perfect passive participles from primary verbs are also not inherited, but rather analogical, e.g.:

- *iūnctus* (the root *ieug- 'to harness' × inherited *iug-tó-s would yield, according to Latin sound laws, *iuctus* or *iūctus*),[9]
- *mīnctus* (*h$_3$meig'^{h}- 'to urinate' × *h$_3$mig'^{h}-tó-s would yield *mictus* or *mīctus*),[10]
- *sparsus* (*sp^{h}erh$_2$g- 'to scatter' × *sp^{h}r̥h$_2$g-tó-s would probably yield *sprāctus*),[11]
- *mānsus* (*men- 'to remain' × *mn̥-tó-s would yield *mentus*),[12]
- *crētus* (*k'erh$_3$- 'to fill up' × *k'r̥h$_3$-tó-s would yield *crātus*),[13]
- *doctus* (*dek'- 'to acquire' × *d$_e$k'-tó-s would yield *dektus*),[14]

[6] Unless indicated otherwise, the PIE roots here and below are quoted according to LIV [7].

[7] The interconsonantal laryngeal (= the sequence CHC) develops generally into *-a-* in Latin (here *dh$_3$t- > *dat-*, *ph$_2$g- > *pag-). The long *-ā-* and the devoiced velar occlusive in *pāctus* are in accordance with Lachmann's law, by which the root vowel lengthens in participles ending in *-tus* when the root ends in a voiced occlusive, while the occlusive itself undergoes devoicing (see e.g. Meiser [8, p. 79]). However, Lachmann's law is one of the most disputed Latin "sound laws" (see e.g. Drinka [3]); the substantiation and the extent of the phenomenon it describes is a matter of unending debate. See also section 3 below.

[8] See note 5 above. Adj. *laudā-tus* 'praised' < *laud-ā-re* 'to praise' < *laud-* 'praise' < the root *leu- 'to sing' (see de Vaan [2, p. 330]).

[9] The quantity of the vowel would depend on the actual scope of Lachmann's law (the same for the next *mictus* × *mīctus*) — see note 7 above and section 3 below.

[10] The initial preconsonantal laryngeal (HC-) is supposed to have dropped in Latin; see e.g. Weiss [15, p. 50].

[11] The sequence CRHC should yield CR*ā*C in Latin, cf. e.g. *tl̥h$_2$-tó- > lat. *(t)lātus*, *g'n̥h$_3$-ró- > lat. *gnārus*, etc. The change *-gt-* > *-ct-* is a common devoicing assimilation.

[12] The sonant *n̥* vocalizes into *en* in Latin; see e.g. Meiser [8, p. 65].

[13] See note 11.

[14] The double-stop root in ablaut zero-grade (i.e. with the "schwa secundum" in old terminology) vocalizes into *-e-* or *-a-* in Latin initial syllables; cf. e.g. Meiser [8, p. 31], Weiss [15, p. 368].

and many others.

Adjectives ending in *-tus* in Latin (= perfect passive participles) function as a means of forming the passive past tense, hence they constitute a virtual functional unit with active perfect forms. The active perfect forms in Latin are multifarious, which is due to the fact that in the category of the Latin perfect there merged forms of the original PIE perfect and aorist, and moreover, other, secondary perfect forms established themselves besides the inherited forms of aorist and perfect.[15]

The Latin perfect passive participles are also formally multifarious. In an effort to relate them to the reconstructed PIE form *R(z)-*tó-*, various phonological sub-rules and exceptions had to be introduced, until a simple principle was identified that interconnects the form of the perfect passive participles with the active perfect forms,[16] namely:

1) If the active perfect is itself a Latin neologism (= "simple" perfects, *u-/v*-perfects, some reduplicated perfects, some *s*-perfects),[17] then the perfect passive participle is a neologism too (formed analogically to the active perfect), cf.

- *iūnxī – iūnctus*,
- *mīnxī – mīnctus*,
- *sparsī – sparsus*,
- *mānsī – mānsus*,
- *crēvī – crētus*,
- *docuī – doctus*,

etc.

2) However, besides the active perfect forms that Latin had inherited directly from PIE (= original root aorists, reduplicated perfects and *s*-aorists) we can find the perfect passive participles that are direct successors of the PIE *R(z)-*tó-*, e.g.:

- *rūpī – ruptus* (< **reup- – *rup-tó-s*),
- *vīcī – victus* (< **u̯eik- – *u̯ik-tó-s*),

[15] This very non-trivial topic has been treated thoroughly in a monograph by Meiser [9].

[16] See Pultrová [10] and [11, p. 21].

[17] See Meiser [9].

- _pepigī – pāctus_ (< _*pe-ph_$_2$_g- – *ph_$_2$_g-tó-s_),

- _tetinī – tentus_ (< _*te-tn̥- – *tn̥-tó-s_),

- _clepsī – cleptus_ (< _*klep-s- – *kl̥p-tó-s_),

- _dūxī – ductus_ (< _*deu̯k’-s- – *duk’-tó-s_),

and many others.[18]

2.2 Adjectives ending in _-uus/-vus_

Adjectives ending in _-uus/-vus_ are generally thought to have issued from PIE adjectives with the structure R(z)-_u̯ó-_ (that is, accented on the suffix and with the root in zero-grade).[19] Some Latin adjectives ending in _-uus/-vus_ are in accord with this reconstruction:

- _vīvus_ ‘alive’ < _*g_^{u}_ih_$_3$_-u̯ó-s_ (the root _*g_^{u}_i̯eh_$_3$_-_ ‘to live’),

- _mortuus_ ‘dead’ < _*mr̥-t-u̯ó-s_ (_*mer-_ ‘to die’),

- _curvus_ ‘bent, crooked’ < _*(s)k_$^{(’)}$_r̥-u̯ó-s_ (_*(s)ker-_ ‘to bend (oneself)’),

- _(g)nāvus_ ‘industrious < efficient < experienced’ < _*gn̥h_$_3$_-u̯ó-s_ (_*g’neh_$_3$_-_ ‘to learn’),

- _prāvus_ ‘crooked, perverse’ < _*pr̥H-u̯ó-s_ (_*preH-_ or _*perH-_ ‘to bend’),[20]

etc.

However, a larger number of adjectives ending in _-uus/-vus_ are evidently secondary:[21]

- _cōnspicuus_ ‘clearly seen, visible’ (from _cōnspicere_ ‘to see’),

- _dīviduus_ ‘divided, divisible’ (from _dīvidere_ ‘to divide’),

[18] A comprehensive list of the relevant passive perfect participles may be found in Pultrová [10] and Pultrová [11, p. 21–30].

[19] Cf. e.g. de Vaan [2], _s. v. arduus_ and others.

[20] The roots of the adj. _curvus_ and _prāvus_ are quoted according to IEW [4, p. 935 and 842, resp.], as LIV [7] does not list these adjectives.

[21] Not derived directly from a root, but from a “complete” Latin word.

- *assiduus* ‘constantly present’ (from *assidēre* ‘to sit by, watch over’),

- *continuus* ‘unremitting, continuous’ (from *continēre* ‘to hold together’),[22]

- *pāscuus* ‘used or suitable for pasture’ (from *pāscere* ‘to pasture’),[23]

- *nocuus* ‘harmful’ (from *nocēre* ‘to harm’),[24]

- *arvus* ‘ploughed, arable’ (from *arāre* ‘to plough’),[25]

and many others.[26]

Let us now turn to how the inherited versus non-inherited distinction is manifested in the semantics of adjectives ending in *-uus*/*-vus*. Members of the inherited group have the meaning of the resultative perfect:

- **mer-* ‘to die’ > *mortuus* = ‘that (has died, and hence) is dead’,

- **g*^{u}*i̯eh*$_3$*-* ‘to live’ > *vīvus* = ‘that is alive’,

- **(s)ker-* ‘to bend (oneself)’ > *curvus* = ‘that (has bent and hence) is curved’,

- **g’neh*$_3$*-* ‘to learn’ > *(g)nāvus* = ‘that (has learned something and hence) is experienced’,

- **preH-* or **perH-* ‘to bend’ > *prāvus* = ‘that (has bent and hence) is crooked’,[27]

etc.

By contrast, members of the non-inherited (analogical) group simply copy the meaning of their founding verb; as a general principle, we can say that in

[22] In all four adjectives mentioned above (*cōnspicuus*, *dīviduus*, *assiduus*, *continuus*) the suffixal derivation came after the prefixation of the base verbs.

[23] The root is **peh*$_2$*(i̯)-* ‘to graze’; *-sc-* is a formant of inchoative verbs; the adjectival derivation is secondary, only coming after the verbal suffix *-sc-* had been applied.

[24] The root is **nek’-* ‘to disappear, get lost’; the ablaut *o*-grade is a feature of causative verbs (**nok’-ei̯e-* ‘to cause death’); the adjective *nocuus* has been derived from the causative verb, not directly from the root.

[25] Secondariness is debatable here: as concerns the form alone, the adjective *arvus* could have been derived secondarily from the verb *arāre*, but the primary derivation from the root **h*$_2$*r̥h*$_3$*-* is not excluded either. The semantics of the adjective indicates the former.

[26] A comprehensive list may be found in Pultrová [11, p. 33–37].

[27] The sound laws relevant to the form of the last five adjectives are the following: *r̥* > *or* (**mr̥-* > *mor-*) or *ur* before *-u̯-* (*(s)kr̥-* > *cur-*); *g*^{u}*-* > *v-* (*vīvus*; cf. e.g. *venīre* < **g*^{u}*em-*), *-i*H- > *-ī-* (*vīvus*); CRHC > CR*ā*C (*gnāvus, prāvus*; cf. note 11 above).

this group using the suffix *-uus*/*-vus* with transitive verbs yields adjectives with a passive meaning (often with an added element of modality), while applying the suffix to intransitive verbs yields active, or more precisely, medial adjectives.

In short, if we had not divided the rather confusing mass of Latin deverbative adjectives with the suffix *-uus*/*-vus* into two groups, inherited and non-inherited (analogically formed), then we would not have had a generally valid method of defining the stem to which the suffix is attached, nor would we have been able to say anything about the semantics of the formations beyond the fact that they are deverbatives. However, after having made this division we saw that in the older layer the suffix is attached to the root in ablaut zero-grade and the semantics of the suffix corresponds to the resultative perfect, while in the younger formations it is the verbal present stem (without the thematic vowel) to which the suffix is attached and the meaning is active/passive (a distinction that did not exist in PIE).

2.3 Other adjective types

A similar principle is also apparent in other types of deverbative adjectives (and some action nouns too): the old, inherited formations are marked in terms of imperfective action / perfective action / state / process,[28] whereas the semantics of the younger, Latin formations is expressed on the active-passive axis.

For example, among Latin adjectives with the suffix *-ilis* there is only one representative that is clearly a primary (inherited) formation, namely the adjective *fragilis*.[29] The meaning of this adjective is ‘liable to break, brittle’, and its semantics thus corresponds to that of the grammatical category of medium (= process, see note 28), with the added meaning of “easiness”. The same characteristics can also be observed in other adjectives ending in *-ilis*

[28] This fourfold distinction corresponds with the theory of Kurzová [5, p. 120], according to which the cornerstone of the PIE verb system was the opposition of “active” × “inactive” diathesis. Active verbs express intentional actions ascribed to an external agent oriented to an external goal, namely imperfective (= present) or perfective (= aorist), whereas inactive verbs express processes (= medium) and states (= perfect), which have no such ascription to external actants.

For the sake of clarity, the system is displayed in the following table:

active		inactive	
imperfective	perfective	state	process
(= present)	(= aorist)	(= perfect)	(= medium)

[29] Adj. *fragilis* does not contain the present nasal infix as the correspondent verb does (*fra-n-gere*), which implies that it is not a secondary derivation from this verb, but an inherited, primary formation with the reconstruction *R(z)-*lí-s*.

that, from the formal point of view, can be inherited formations, e.g. *agilis* ‘that moves easily, nimble’, *fūtilis* = ‘that leaks easily’.[30] On the other hand, the clearly secondary *docilis* ‘apt to learn, teachable’ (from *docēre* ‘to teach’ × the primary formation would be *decilis* < *d_ek’-lí-s*) and *ūtilis* ‘useful, serviceable’ (from *ūtī* ‘to use’ × the PIE *$h_3$it-lís* would yield *ītilis* in Latin) have the meaning of passive “aptitude”.

The same applies for the not very numerous group of Latin deverbative adjectives ending in *-ius*. The two members of this group that are likely to be primary formations, namely *fluvius* (acting as a noun in Latin, i.e. as a substantivized adjective with the meaning ‘a stream’) and *pluvius*, both have a medial meaning (‘that flows’, resp. ‘that rains’). By contrast, the clearly secondary *eximius* (from the verb *eximere* ‘to take out’) has the passive meaning ‘excepted, outstanding’ = ‘that is to be taken out’.

3 Conclusions

The diachronic distinction of being inherited or non-inherited thus has important implications for the synchronic semantic and formal analysis of word-formative types. Moreover, the (not always trivial) distinction between inherited and non-inherited (= analogical) formations also plays a pivotal role in the description of the phonological system of the language. If we treated analogical formations as if they were inherited then we would find ourselves in a situation where the presumed form of the word (according to the reconstruction of PIE and sound laws as currently formulated) simply does not correspond to reality. A consequence of not observing the diachronic distinction between inherited and non-inherited is the need to endlessly define minor and not generally valid rules, supplemented in each case by a list of exceptions. This can be seen in the fact that historical grammars and etymological dictionaries abound in sound laws (actually not sound laws at all) which very often result precisely from treating an analogical formation as inherited.

Lachmann’s law, to which we alluded in note 7 above, is an instance of such a sound law. Part of the problem with this “law” consists in the fact that there are several examples not in accord with its definition; that is, there exist Latin perfect passive participles from roots ending in a voiced occlusive where the root vowel does not lengthen, e.g. *fossus* (*b^hed^h(h_2)-*) or *tractus* (*d^hre$g^{(’)h}$*). This complicates defining the actual scope of the phonological phenomenon described by the law: Is it only some voiced occlusives that are involved in the phonological process in question (devoicing and lengthening the preceding vowel)? And if so, which and why? As a matter of fact, the two adjectives ending in *-tus* already mentioned, *fossus* and *tractus*, are examples of non-inherited, analogical formations (in the case of *fossus* this is seen immediately due to its *o*-vocalism), and as such they fall out of the scope of

[30] Cf. Pultrová [11, p. 62].

sound laws. (It should however be pointed out that not all the exceptions to the basic definition of Lachmann’s law can be resolved this way — the debate on this law is far from being ended.)

Continuing with the same adjective type, what has thus far also complicated the definition of Lachmann’s law is the fact that there are — reverse to what has been said above about “exceptions to Lachmann’s law” — many perfect passive participles with long vowels with roots ending in consonants other than a voiced occlusive, thus falling beyond the scope of Lachmann’s law, and at the same time not being in accord with the PIE reconstruction *R(z)-_tó_-. There are also many other instances of non-compliance with this reconstructed PIE form. As indicated in section 2.1, all these instances of non-compliance result from having been formed analogically to the corresponding active perfect forms. However, Latin historical grammars have thus far not reflected this rule, and sometimes introduce a completely unsystematic sub-rule according to which in some adjectives ending in _-tus_ a “secondary full-grade root” had evolved, or they resort to the equally unsystematic _o_-grade for their reconstruction (see e.g. Vine [14], Meiser [8, p. 112]: _nōtus_ < *_g’noh$_3$-tó-_, etc.).

The same failure to distinguish between inherited and analogical formations within one word-formative type makes it impossible to set clear rules of laryngeal development in Latin. For example, the initial preconsonantal laryngeals (HC-) are generally dropped in Latin.[31] Nevertheless, Schrijver in his systematic treatise on the development of laryngeals in Latin [12] introduces another rule of HC- > _a_C- in Latin, giving as instances, however, _āctus_ or _arvus_, both in all likelihood being analogical formations,[32] and thus not directly reflecting the PIE phonological system.

One could continue with further examples, but let me sum up by saying that what I wished to demonstrate above all is, first, that in word formation, perhaps even more than in other linguistic fields of study, a combination of synchronic and diachronic approaches may bear fruit; and, second, that we must, at all times, apply the prism of word-formation to solving phonological problems, i.e. not treat phenomena within individual formations, but regard them always in the context of the whole word-formative type.

References

[1] Brugmann, Karl. _Grundriss der vergleichenden Grammatik_, II,1 (_Lehre von den Wortformen und ihrem Gebrauch_). Strassburg: K. J. Trübner, 1906.

[31] E.g. *_h$_1$rud^h-ro-s_ > Lat. _ruber_ (× Greek ἐρυθρός); *_h$_3$d-n̥t-_ > Lat. _dēns_ (Greek ὀδών), etc.

[32] For _arvus_ see section 2.2 above. For _āctus_ see Meiser [9, p. 207]: the author explains the active perfect _ēgī_ not as an original root aorist, but as an analogical formation to _fēcī_, _iēcī_, etc. (see also LIV [7, p. 256]). Thus the participle _āctus_ is also not inherited but analogical — see section 2.1.

[2] de Vaan, Michiel. *Etymological Dictionary of Latin and the other Italic Languages*. Leiden – Boston: Brill, 2008.

[3] Drinka, Bridget. Lachmann's Law: A Phonological Solution. *Indogermanische Forschungen* 96, pp. 52–74, 1991.

[4] IEW = Pokorny, Julius. *Indogermanisches etymologisches Wörterbuch*. Bern – München: Francke, 1959.

[5] Kurzová Helena. *From Indo-European to Latin. The Evolution of a Morphosyntactic Type*. Amsterdam – Philadelphia: John Benjamins, 1993.

[6] Leumann, Manu. *Lateinische Laut- und Formenlehre, Lateinische Grammatik I*. München: C. H. Beck, 1977.

[7] LIV = Rix, Helmut et al. *Lexikon der indogermanischen Verben*. Wiesbaden: Dr.-Ludwig-Reichert-Verlag, 2001^{2}.

[8] Meiser, Gerhard. *Historische Laut- und Formenlehre der lateinischen Sprache*. Darmstadt: Wissenschaftliche Buchgesellschaft, 1998.

[9] Meiser, Gerhard. *Veni Vidi Vici. Die Vorgeschichte des lateinischen Perfektsystems*. München: C. H. Beck, 2003.

[10] Pultrová, Lucie. The Formation of the Latin Perfect Passive Participles. *AUC Philologica – Graecolatina Pragensia* XXI, pp. 101–139, 2006.

[11] Pultrová, Lucie. *The Latin Deverbative Nouns and Adjectives*. Prague: Karolinum, 2011.

[12] Schrijver, Peter. *The Reflexes of the Proto-Indo-European Laryngeals in Latin*. Amsterdam – Atlanta (GA): Rodopi, 1991.

[13] Sihler, Andrew L. *New Comparative Grammar of Greek and Latin*. Oxford University Press, 1995.

[14] Vine, Brent. On PIE Full Grades in Some Zero-Grade Contexts: **-tí-*, **-tó-*. In Clackson, James and Olsen, Birgit Anette (eds.) *Indo-European Word Formation (Proceedings of the Conference held at the University of Copenhagen October 20th–22nd 2000)*, pp. 357–379. Copenhagen: Museum Tusculanum Press, 2004.

[15] Weiss, Michael. *Outline of the Historical and Comparative Grammar of Latin*. Ann Arbor: Beech Stave Press, 2009.

A Computational Implementation of Pāṇini's Derivational Morphology of Sanskrit

Peter M. Scharf

President, The Sanskrit Library
Visiting Professor, Department of Humanities and Social Sciences
Indian Institute of Technology Bombay
scharf@sanskritlibrary.org

5 September 2017

Abstract

The most extensive analysis of derivational morphology ever undertaken is the analysis of Sanskrit by Pāṇini by the fourth century BC. Pāṇini achieved an extraordinary degree of abstraction in systematically describing the language of his time and of inherited literature. Pāṇini's linguistic system consists of a set of about four thousand rules formulated in compact aphorisms (*sūtra*s) in the *Aṣṭādhyāyī*, a set of basic phonological units (*akṣarasamāmnāya*) ordered for the purpose of making abbreviatory terms (*pratyāhāra*), a list of about two thousand verbal roots (*dhātupāṭha*), and two hundred and eighty-two minor lexical lists (*gaṇapāṭha*). Rules classify semantic objects, add affixes to basic roots and nominal bases under semantic and cooccurrence conditions, and make morphophonemic and phonetic modifications to reconstruct utterances of the language. One of the most complicated sections is the section that generates secondary nominal derivates, the taddhita section where Pāṇini segregates formal conditions from semantic conditions.

The author has produced an XML formalization of Pāṇini's linguistic system amenable to the production of a computational implementation. Each rule organizes a set of regular expressions and attributes into a tree consisting of XML elements. XML elements may contain a phone attribute that refers to a subsegment of the phonetic string in the data structure, and attributes of that subsegment. The computational implementation tracks rules, associates semantic conditions with morphological units, and preserves dependency relations and information about expected complements. The taddhita section is formalized as a constrained many-to-many mapping of formal affixation rules to semantic conditions.

1 Introduction

By the fourth century BC, Pāṇini composed a fairly complete generative grammar of the language consisting of about four thousand rules (*Aṣṭādhyāyī*), a classified list of about two thousand verbal roots (*dhātupāṭha*), nearly 250 lists of nominal bases (*gaṇapāṭha*), and a structured list of basic phonetic segments (*akṣarasamāmnāya*). The grammar reconstructs Sanskrit utterances by classifying semantic objects and basic phonetic segments and introducing affixes after basic lexical elements under various semantic and cooccurrence conditions. The grammar achieves an extraordinary degree of efficiency by employing various techniques of rule formulation, organization and interaction some of which are explicitly described in metarules. Aside from precisely describing inter-word phonetic changes, tonal details, inflectional morphology and some phrasal syntax, the grammar extensively describes derivational morphology. Pāṇini derives secondary verbal roots, denominative verbs, deverbative and denominative nominal derivates, and compounds. Scharf [16] briefly summarizes the extensive tradition of commentary on Pāṇini's grammar and references the several bibliographic sources to editions, translations, and the abundant research on it in the modern era. Recently several projects have implemented sections of Pāṇini's *Aṣṭādhyāyī* computationally and described plans for comprehensive modeling. Scharf et al. [20, pp. 165–170], who described the computational implementation of rules concerning voice, preverb, and transitivity restrictions in verbs, surveyed these projects and the most pertinent discussions. Scharf [14] demonstrates and Cardona [3] explicates in detail how Pāṇini generates speech forms from semantic and cooccurrence conditions. Scharf [15] examines cases of rule conflict (*vipratiṣedha*) throughout the grammar and casts doubt on their consistent solvability by simple rule-selection principles.

The XML formalization of Pāṇini's *Aṣṭādhyāyī* I have constructed over the past four years, called *Paitāmbarī*, implements a bottom-up formalization that accurately captures the provision of each rule taking into consideration information that recurs by inheritance from other rules and incorporating metarules difficult to segregate. Scharf [19] previously described how the structure of *Paitāmbarī* captures the general sweep of Pāṇinian linguistic description. There I explain how the formalization represents strings analyzed into components introduced under semantic and cooccurrence conditions which are subject to combination, replacement, deletion, and augmentation under additional semantic, cooccurrence, and phonetic conditions. Ajotikar, Ajotikar, and Scharf [1] explain some Pāṇinian techniques and illustrate how *Paitāmbarī* captures them in a number of cases. Two principles relevant to the taddhita section are discussed in detail here.

The approach of carefully modeling Pāṇinian rules and procedures differs from the comparison of contemporary computational architectures with generalized abstractions of rule structures. Comparison of Pāṇini's procedures with contemporary computational models is useful to illuminate the extent to which Pāṇini may have employed such models as well as to suggest possible solutions to contemporary issues based upon Pāṇinian methods. Yet care must be taken not to impose contemporary models anachronistically on ancient work either with a view to claim that the ancient work anticipated contemporary work or with a view to claim to explain the procedure of the ancient work. For example, as Scharf [13] and Scharf [12] demonstrated, Houben [5] was right to critique the proposition that Pāṇini operated with distinct levels as articulated by Kiparsky and Staal [9]. Kiparsky [8] himself hedges his earlier attribution of levels to Pāṇini calling them, "what we (from a somewhat anachronistic modern perspective) could see as different levels of representation."

Recently, Kiparsky's student, Deo [4] compared the architecture of the section of rules that derive secondary nominal stems ending in taddhita affixes from other nominals to modern inheritance-based lexica. She argues that the interleaving of formal and semantic rules in a single-inheritance hierarchy with defaults in the taddhita section provides a constrained separation of the form and meaning of affixes that elegantly represents the homonymy and synonymy of these affixes in the complex derivational morphology of Sanskrit. Her comparison is interesting and inspired Krishna and Goyal [10] to produce a Java implementation of the taddhita section that models the structure Deo [4] described. I concur that the taddhita section provides a constrained separation of the form and meaning of affixes, and *Paitāmbarī* formalizes such a separation and mapping. Yet I disagree that the taddhita section conforms to a single-inheritance hierarchy. Rather, both the taddhita section and other sections of the *Aṣṭādhyāyī* appear to be characterized by a multiple-inheritance hierarchy.

Overlapping domains are the essence of a multiple-inheritance structure; a single-inheritance structure is one in which domains occur in a hierarchical tree structure and do not overlap. Deo [4, p. 13] clearly defines the criteria for overlapping domains and writes, "The rules in the *Aṣṭādhyāyī* must be ordered in this relation for multiple-inheritance to obtain. However, the fact of the matter is that they are not." She continues, "it is rare in the *Aṣṭādhyāyī* for two or more adhikāras to partially overlap in their domains. This shows that the *Aṣṭādhyāyī* relies on single inheritance for representing shared information."

2 Anuvr̥tti, adhikāra, and classification

In order to clarify that the *Aṣṭādhyāyī* as a whole has a multiple-inheritance hierarchy rather than a single-inheritance hierarchy it is essential to note that the device of the heading (*adhikāra*) does not differ categorically from that of recurrence (*anuvr̥tti*). This is essential because it is rather common for domains of anuvr̥tti, which is ubiquitous, to overlap; in contrast, the scope of headings rarely do simply because they are far fewer. Deo [4, p. 12] rightly indicates that "An adhikāra may be considered to be a special type of anuvr̥tti." Yet she emphasizes the distinction between the two in order to isolate the former from the latter. She writes, "The adhikāra device in the Aṣṭādhyāyī, unlike anuvr̥tti, does not facilitate the information of procedural information in rules, but the inheritance of information about the classification of rules." On the contrary, both headings (*adhikāra*s) and recurrence (*anuvr̥tti*) facilitate procedural information in rules. Classification belongs to a different sūtra type from the adhikāra, namely, rules that introduce technical terms (*sañjñāsūtra*). Headings convey information about classification of rules only insofar as they also either themselves consist of technical terms or utilize technical terms introduced elsewhere.

In an extended use of natural language ellipsis, the *Aṣṭādhyāyī* has terms from preceding sūtras recur in subsequent sūtras. In this way, Pāṇini factors out conditions common to a number of rules and groups rules with common conditions together in sections. When the term that states the common condition occurs within a rule that provides an operation, the recurrence is simply termed *anuvr̥tti* 'recurrence'. Terms may recur in just the following rule or in numerous rules. When the term that states a common condition is placed as a heading above subsequent rules, but is not a complete operational rule in itself, it is termed *adhikāra* 'heading'. Headings recur in numerous subsequent rules. While both simple recurrence and headings serve to create sections with common characteristics, headings create larger thematic divisions. Yet a heading is not necessarily a classifier. Rather Pāṇini employs technical terms (*sañjñā*) to classify items.

While some headings introduce technical terms and thus are explicit classifiers, many headings serve to state that operations apply to classes of objects where the classification has been achieved elsewhere by a rule that introduces a technical term (*sañjñāsūtra*). Headings that explicitly classify items include *A.* 2.1.3 *prāk kaḍārāt samāsaḥ*, *A.* 3.1.1 *pratyayaḥ*, and *A.* 4.1.76 *taddhitāḥ*. These rules term items introduced by subsequent rules *samāsa* 'compound', *pratyaya* 'affix', and *taddhita* respectively. Only ten of the seventy-one adhikāras in the *Aṣṭādhyāyī* explicitly undertake classification by introducing a technical term in this way. Other sañjñāsūtras are valid within the scope of a heading under which they occur. For example, *A.* 3.1.93 *kr̥d atiṅ*, which uccurs under the heading *A.* 3.1.91 *dhātoḥ* 'after a verbal root', terms affixes introduced under that heading *kr̥t*. Others still refer to an external list. Thus *A.* 1.3.1 *bhūvādayo dhātavaḥ* terms items listed in the *dhātupāṭha dhātu* 'root'. This and many other *sañjñāsūtra*s introduce technical terms that are used in headings elsewhere. Thus *A.* 1.4.13 *yasmāt pratyayavidhis tadādi pratyaye 'ṅgam* terms that which begins with that after which an affix is introduced *aṅga* 'stem' with respect to that affix. *A.* 1.4.14 *suptiṅantam padam* terms speech forms that end in a nominal or verbal termination *pada* 'word'. The following three rules term additional speech forms before certain affixes *pada* as well thereby extending the classification. *A.* 1.4.18 *yaci bham* and a few additional rules term certain stems *bha*. *A.* 1.2.45 *arthavad adhātur apratyayaḥ prātipadikam* terms meaningful speech forms that are not a verbal root or affix, and do not end in an affix, *prātipadika* 'nominal base', and *A.* 1.2.46 *kr̥ttaddhitasamāsāś ca* so terms speech forms that end in a kr̥t or taddhita affix and speech forms termed *samāsa* 'compound'.

The technical terms introduced by the sañjñāsūtras that are not headings, such as *dhātu*, *prātipadika*, *aṅga*, *bha*, and *pada*, serve to classify various speech forms as root, nominal base, stem, a special subtype of stem, and word. These terms are then used elsewhere in headings of sections of rules that provide operations on each class of item. Twenty-five of the seventy-one headings employ technical terms introduced elsewhere in the *Aṣṭādhyāyī* to indicate procedural function in subsequent rules while a few additional headings use terms familiar from prior usage (e.g. *uttarapada*). Thus 3.1.91 *dhātoḥ* heads a section of rules that introduce affixes (*pratyaya A.* 3.1.1) after a verbal root. *A.* 4.1.1 *ṅyāpprātipadikāt* heads a section of rules that introduce affixes after a nominal base (*prātipadika*) or an item ending in a feminine affix *ī* or *ā*. *A.* 6.4.1 *aṅgasya* introduces a section that provides operations relevant to stems, and *A.* 6.4.129 *bhasya* to a special subclass of stems. *A.* 8.1.16 *padasya* heads a section of rules that provide operations relevant to a word, while *A.* 8.3.55 *apadāntasya mūrdhanyaḥ* heads a section of rules that provide a retroflex (*mūrdhanya*) replacement for a sound that is not word-final (*a-pada-anta*). While just over half of the headings employ classificatory terms, most of these serve a procedural function in subsequent rules, and only ten are themselves sañjñāsūtras. Hence, classification, which properly belongs to the sañjñāsūtra, cannot be said to be the function of headings nor be accepted as a criterion to distinguish headings fundamentally from recurrence.

3 Overlapping domains

It is quite common for recurring terms to occur in overlapping domains. For example, *A.* 3.2.1 includes the term *karmaṇi* which recurs through *A.* 3.2.58. *A.* 3.2.3 includes the term *anupasarge* which recurs through *A.* 3.2.60. The domains of the term *karmaṇi* and *anupasarge* overlap in *A.* 3.2.3–58 but are independent in *A.* 3.2.2 and *A.* 3.2.59. The domains of recurring terms overlap with headings too. For example, the term *dīrghaḥ* in *A.* 6.3.111 *ḍhralope pūrvasya dīrgho 'ṇaḥ*, which is not a heading, recurs through *A.* 6.4.18, past the heading *A.* 6.4.1 *aṅgasya* which is valid through *A.* 7.4.97, the end of the seventh adhyāya. Even the domains of headings overlap with each other. For example, the headings *A.* 4.1.3 *striyām* and *A.* 4.1.14 *anupasarjanāt* both recur through *A.* 4.1.81, several sūtras after the heading *A.* 4.1.76 *taddhitāḥ* which continues through the end of fifth adhyāya. Similarly, the heading *A.* 8.1.16 *padasya* continues through *A.* 8.3.54, into the middle of the section headed by *A.* 8.2.1 *pūrvatrāsiddham* which continues through *A.* 8.4.68, the end of the *Aṣṭādhyāyī.*

Finally, the taddhita section itself contains headings whose domains overlap. *A.* 4.1.87 *strīpuṁsābhyāṁ nañsnañau bhavanāt* mentions its range as terminating prior to the occurrence of the term *bhavana* which occurs in *A.* 5.2.1 *dhānyānāṁ bhavane kṣetre khañ*. *A.* 4.1.87 provides the affixes ***nañ*** and ***snañ*** after the nominal bases *strī* and *pums* respectively under semantic conditions stated in *A.* 4.1.88–5.1.136, that is, up to the end of the first pāda of the fifth adhyāya. The domain includes the domain of several headings in the taddhita section that provide default affixes. Yet in exception to these defaults, the rule provides these two affixes after the two specific bases mentioned. While this appears to be the only heading in the taddhita section that violates the single-inheritance structure, it nonetheless demonstrates the inappropriateness of Deo’s [4, p. 13] claim, “The taddhita hierarchy, ..., is based only on single inheritance.” This claim is hardly necessary, however, for Deo [4, p. 8] to establish her main point, namely, that Pāṇini’s treatment of derivational morphology in the taddhita section, “[b]y allowing for flexible, many-to-many correspondences between affixal form and semantics, and at the same time, constraining the range of these correspondences,” provides a “model of constrained separationism embodied in the formalism of an inheritance hierarchy.”

4 The taddhita section

Joshi [7] already studied Pāṇini’s taddhita affixation rules in his doctoral dissertation, and Cardona [2, pp. 229–255] described the structure of the section. To a large extent, the section abstracts formal derivational factors from semantic and syntactic conditions. Thus five major headings provide certain affixes by default under semantic conditions stated in subsequent rules:

- *A.* 4.1.83 *prāg dīvyato ’ṇ* provides the affix *a**ṇ***
- *A.* 4.4.1 *prāg vahateṣ ṭhak* provides the affix *ṭha**k***
- *A.* 4.4.75 *prāg ghitād yat* provides the affix *ya**t***
- *A.* 5.1.1 *prāk krītāc chaḥ* provides the affix *cha*
- *A.* 5.1.18 *prāg vateṣ ṭhañ* provides the affix *ṭha**ñ***

Each of these rules is followed by rules that state the semantic conditions under which the affix is provided, as well as by rules that provide different affixes in exception to the default affixes. Thus among the numerous semantic conditions stated in the domain of *A.* 4.1.83 are included for example the following:

- *A.* 4.1.92 *tasyāpatyam* ‘his offspring’
- *A.* 4.2.1 *tena raktaṁ rāgāt* ‘dyed by that after a speech form denoting a dying agent’
- *A.* 4.2.24 *sāsya devatā* ‘after a speech form denoting a divinity, something related to it’
- *A.* 4.2.37 *tasya samūhaḥ* ‘its group’

Two issues are of paramount importance in this section: (1) whether taddhita affixes are provided after inflected words or after nominal bases, and (2) the relation between formal and semantic factors, namely, between the rules that provide the affixes and the sūtras that state semantic conditions. The first issue arises because of two competing headings. As mentioned above, the heading *A.* 4.1.1 *ṅyāpprātipadikāt*, valid through *A.* 5.4.130, the end of the entire taddhita section, indicates that subsequent rules provide affixes after nominal bases or items ending in a feminine affix. On the other hand, the heading *A.* 4.1.82 *samarthānāṁ prathamād vā*, valid through *A.* 5.2.140, indicates that subsequent rules provide affixes after the first of syntactically related inflected words. Cardona [2, 254–255 ¶366] makes clear that rules such as *A.* 4.1.77, which precede the latter heading, provide affixes after nominal bases, while rules such as *A.* 5.3.7, which occur after the termination of the domain of the latter heading, provide affixes after inflected words. Pāṇini provides a mechanism for deleting the nominal termination of the inflected word after which a taddhita affix is taught in *A.* 2.4.71 *supo dhātuprātipadikayoḥ*, the same rule by which the nominal terminations of the inflected words that are the constituents of a compound are deleted.

As Cardona [2, 246 ¶351] notes, in the domain of the heading *A.* 4.1.82, however, in which the heading *A.* 4.1.1 is also valid, two major commentaries on the *Aṣṭādhyāyī* paraphrase taddhita affixation rules differently. The *Kāśikā* commentary often states that the taddhita affix occurs after a nominal base, while the *Siddhāntakaumudī* states that affixes should occur after the first inflected word stated in semantic conditions provided the base conforms to constraints stated in formal affixation rules. Here Cardona [2, 255 ¶366] notes that the fact that *A.* 6.3.17 provides for the non-deletion of nominal terminations in exception to *A.* 2.4.71 demonstrates that certain taddhita affixes are indeed introduced after inflected words, not after nominal bases. In concurrence, Scharf [17] concluded that certain rules require that the taddhita affixes be added after inflected words. These rules include in particular *A.* 4.3.23 *sāyañcirampráhṇeprage'vyayebhyaṣ ṭyuṭyulau tuṭ ca* and *A.* 4.3.24 *vibhāṣā pūrvāhṇāparāhṇābhyām* which occur in the domain of the heading *A.* 4.1.82. This conclusion is reached despite the fact that Scharf [18] considers a similar problem in the derivation of primary deverbative nominal derivates that occur only as the final elements in upapada-tatatpuruṣa compounds and concludes that even where rules explicitly refer to the nominal termination of the subordinate element, the nominal termination cannot be present. In that context, the kāraka that conditions the nominal termination must be inferred in the rule instead. Regarding these rules in the taddhita section, however, the fact that a rule specifically negates the deletion of the inflectional termination necessitates that the base be an inflected word. Scharf [17] considers previously unnoticed difficulties that ensue if the bases are inflected words rather than nominal bases, particularly in the applicability of rules that apply to stems (termed *aṅga*) rather than inflected words (termed *pada*). One is required to accept that the term *aṅga* provided by *A.* 1.4.13 retains applicability even where the term *pada* provided by *A.* 1.4.14–17 is applicable in exception to the heading governing those rules that require that just the latter term be applicable.

Concerning the second issue regarding the relation between formal and semantic factors in the taddhita section, commentators and modern scholars generally consider that the default affixes provided by the general rules *A.* 4.1.83 etc. that head the several major subsections of the taddhita section recur in the sūtras that state semantic conditions that are governed by them. The assumption is that the rules are organized in a tree structure with subordinate notes inheriting information from higher nodes. For example, Deo [4, p. 16] writes, "The general affix is inherited by each of the arthādhikāras. The arthādhikāra node, therefore is more specific than the pratyayādhikāra node, since it contains information about both the formal affix prescribed (via inheritance) as well as the semantic conditions under which it is attached to the derived forms." *A*. 4.1.92 *tasyāpatayam*, which explicitly states just a semantic condition 'his offspring', is therefore interpreted as an operational rule (*vidhi*) that provides the affix *a**ṇ*** (inherited from *A*. 4.1.83). In accordance with the heading *A*. 4.1.82, the affix is provided after an inflected word that is a value of the genitive singular pronoun *tasya* in the meaning 'his offspring'. Particular rules that provide different affixes under certain formal conditions stated beneath the semantic heading *A*. 4.1.92 are taken to be exceptions to this rule. For example, *A*. 4.1.95 *ata iñ* states that the affix *i**ñ*** after a nominal base that ends in *a*. Taken together with headings, the rule provides that the affix *i**ñ*** occurs after an inflected word in the genitive whose nominal base ends in *a* in the meaning 'his offspring'. Here, according to the normal assumption of inheritance to a subordinate node in a tree structure, Deo [4, p. 17] writes that the affix *i**ñ*** "overrides *á**ṇ*** (4.1.92)".

There are problems with the assumption that rules in this section are structured in a tree, that subordinate nodes inherit information from higher nodes, and that the most subordinate node whose conditions are met provides the affix. First of all, there are other rules besides those that provide default affixes that provide affixes under the semantic conditions stated under them that are not specific exceptions to the semantic conditions. Some of the general headings that provide default affixes are followed by rules that provide different affixes after specific bases in all the semantic conditions mentioned under these headings. For example, the following rules provide other affixes instead of the affix *a**ṇ*** in the domain headed by *A*. 4.1.83 in all the semantic conditions stated in that domain:

- *A*. 4.1.85 *dityadityādityapatyuttarapadāṇ ṇyaḥ* (*ṇyaḥ* 84)
- *A*. 4.1.86 *utsādibhyo 'ñ*

A. 4.1.85 provides the affix ***ṇ**ya* after the bases *diti*, *aditi*, and *āditya* and compounds whose final constituent is *pati*. *A*. 4.1.86 provides the affix *a**ñ*** after the bases in the list beginning with *utsa*. Other rules provide other affixes or delete the default affix in all the semantic conditions stated in the domain of *A*. 4.1.83. Similarly, three rules provide the affix *ya**t*** instead of the affix *cha* in the domain headed by *A*. 5.1.1 in all the semantic conditions stated in that domain. The first of these, *A*. 5.1.2 *ugavādibhyo yat*, for instance, provides the affix *ya**t*** after bases that end in an *u*-vowel and after bases in the list beginning with *go*. Similarly, *A*. 5.3.71 *avyayasarvanāmnām akac prāk ṭeḥ* provides the affix *ak**ac*** instead of the affix *ka* in the domain headed by *A*. 5.3.70 *prāgivāt kaḥ* after an indeclineable or pronominal in all the semantic conditions stated under that heading.

If the numerous sūtras that state semantic conditions under the major headings that provide default affixes inherit the default affix and themselves provide that affix, they must also inherit the information of the exceptions to the default affix, including the specifics regarding the stems after which those exceptional affixes occur, and must state provisions of those exceptional affixes under those conditions as well. If this is so, then rules such as *A*. 4.1.92 become quite complex. *A*. 4.1.92 would have to mean that the affix *aṇ* (inherited from *A*. 4.1.83) occurs after an inflected word that is a value of the genitive singular pronoun *tasya* in the meaning 'his offspring' unless the nominal base of that word is *diti*, *aditi*, or *āditya* or a compound whose final constituent is *pati*, in which case the affix *ṇya* occurs (information inherited from *A*. 4.1.85), or unless the base is included in the list beginning with *utsa*, in which case the affix ***añ*** occurs (information inherited from *A*. 4.1.86), etc. Moreover, every semantic condition stated under all five of the major headings that provide default affixes listed at the beginning of this section would have to state in addition, "unless the nominal base is *strī* or *pums*, in which case the affix *na**ñ*** or *sna**ñ*** occurs respectively (information inherited from *A*. 4.1.87.

The proliferation of redundancy by the inheritance of multiple formal headings on lower semantic nodes of a tree in this manner by assuming that inheritance operates without distinguishing formal from semantic criteria is extremely prolix. However, it is not necessarily the case that Pāṇini or his commentators operated in this manner. One sign that they did would be if commentators considered more specific rules stated under intermediate semantic headings to be exceptions to the semantic sūtras. However, the *Kāśikā* uniformly states that such rules are exceptions to the default affixes provided by the higher level formal headings, not to semantic sūtras. This is so because affixes provided in the nominative in more specific rules are exceptions to the affixes provided in the nominative in the general headings. Yet it indicates awareness of a structure in which the specific rules that provide affixes are directly related to the general rules that provide affixes without the intermediary of the semantic conditions even though those semantic conditions define the semantic domain in which the specific rules apply.

Another indication that Pāṇini segregated the structure of semantic conditions from formal conditions rather than configuring them in a single tree is the manner in which he refers to the terminal points of the domain of the headings that provide the default affixes. These rules refer to the domains in which they apply in terms of the relevant semantic conditions, not to specific points in the sequence of sūtras. The text of the *Aṣṭādhyāyī* is a single line of sūtras owing to the fact that speech occurs in single sequence of sounds in the dimension of time, and a manuscript of a text in a single sequential string of characters. If a single line of inheritance were at the fore in the composers mind, he would have stated the terminus in terms of the single line of sūtras, but he doesn't.

Although the reference to the termination of the scope of the default affix *aṇ* provided by the formal heading *A*. 4.1.83 is to the semantic condition *dīvyati* in *A*. 4.4.2, the affix *aṇ* recurs only through *A*. 4.3.168, the end of *A*. 4.3. The next rule *A*. 4.4.1 provides *ṭha**k*** as the general affix for the next section; hence it is not in the scope of *A*. 4.1.83. Likewise, although the reference to the termination of the scope of the default affix *ṭha**k*** provided by the formal heading *A*. 4.4.1 is to the semantic condition *vahati* in *A*. 4.4.76, the affix *ṭha**k*** recurs only through *A*. 4.4.74. *A*. 4.4.75 provides the affix *yat* generally for the next section. The *Kāśikā* on *A*. 4.4.74 states *ṭhakaḥ pūrṇo 'vadhiḥ. ataḥ param anyaḥ pratyayo vidhīyate.* Again, although the reference to the termination of the scope of the default affix in *A*. 4.4.75 is to the semantic condition *hita* in *A*. 5.1.5, the affix *yat* recurs through *A*. 4.4.144, the end of the 4th adhyāya. The next affix adhikāra is stated in *A*. 5.1.1. The *Kāśikā* on *A*. 4.4.144 states *yataḥ pūrṇo 'vadhiḥ. ataḥ param anyaḥ pratyayo 'dhikriyate.* Once again, although the reference to the termination of the scope of the default affix *cha* in *A*. 5.1.1 is to the semantic condition *krīta* in *A*. 5.1.37, the affix *cha* recurs only through *A*. 5.1.17. The affix *yat* provided after bases end in an *u*-vowel or in the list beginning with *go* by *A*. 5.1.2 has the same scope. Under *A*. 5.1.17 the *Kāśikā* states *chayatoḥ purṇo 'vadhiḥ. itaḥ param anyaḥ pratyayo vidhīyate.* Thus in four of the five major formal headings that provide default affixes in the taddhita section, reference to the termination of the scope is in terms of the semantic condition rather than to the actual sūtra at which the scope terminates. This is not accidental. It is clear that Pāṇini segregates semantics from formal conditions and is specifying the semantic conditions in which the affixes occur, not simply the terminus in a single list of sūtras. As Cardona [2, 246 ¶[352] notes, "in sūtras like *A*. 4.1.83, Pāṇini is concerned with sections of meaning which condition the introduction of affixes. ...Pāṇini is providing for sections of meanings." He then cites the *Kāśikā* on *A*. 5.1.1, *artho 'vadhitvena gr̥hītaḥ na pratyayaḥ. tena prāk ṭhañaś cha iti noktam*, and translates "Meaning is taken as the boundary, not an affix. Therefore, (Pāṇini) has not said *prāk ṭhañaś chaḥ*." The headings that provide general default affixes specify a set of semantic conditions in which those affixes apply, and the exceptions to those defaults that have the same range apply to the same set of semantic conditions. This constitutes a constrained mapping of multiple affixes to multiple semantic conditions.

5 Formalization of the taddhita section

Paitāmbarī formalizes the rules in the taddhita section by clearly delineating semantic conditions from affix provision. Semantic conditions are collected in a set and given a the name by which they are referenced in the formal heading that provides the default affix and specifies the terminus. For example, the semantic conditions specified under the heading *A*. 4.1.83 are given the name *prāgdīvyatīya*. The rules *A*. 4.1.83–86 apply to any form that satisfies any of those semantic conditions. Specific rules such as *A*. 4.1.95 are considered exceptions to *A*. 4.1.83, not to *A*. 4.1.92 which states the semantic condition.

References

[1] Tanuja Ajotikar, Anuja Ajotikar, and Peter M. Scharf. "Some issues in the computational implementation of the Aṣṭādhyāyī". In: *Sanskrit and computational linguistics. select papers presented at the 16th World Sanskrit Conference in the 'Sanskrit and the IT world' section 28 June – 2 July 2015, San-*

skrit Studies Center, Silpakorn University, Bangkok. Ed. by Amba Kulkarni. New Delhi: D. K. Publishers, 2016, pp. 101–123.

[2] George Cardona. *Pāṇini. his work and its traditions*. Vol. 1: *Background and Introduction*. Second edition, revised and enlarged. Delhi: Motilal Banarsidass, 1997.

[3] George Cardona. “On the structure of Pāṇini’s system”. In: *Sanskrit computational linguistics. first and second international symposia, Rocquencourt, France, October 2007; Providence, RI, USA, May 2008; Revised selected and invited papers*. Ed. by Gérard Huet, Amba Kulkarni, and Peter M. Scharf. Lecture Notes in Artificial Intelligence 5402. Berlin; Heidelberg: Springer-Verlag, 2009, pp. 1–32.

[4] Ashwini Deo. “Derivational morphology in inheritance-based lexica. insights from Pāṇini”. In: *Lingua* 117.1 (2007), pp. 175–201.

[5] Jan E. M. Houben. “‘Meaning statements’ in Pāṇini’s grammar: on the purpose and context of the Aṣṭādhyāyī”. In: *Studien zur Indologie und Iranistik* 22 (1999), pp. 23–54.

[6] Gérard Huet, Amba Kulkarni, and Peter M. Scharf, eds. *Sanskrit computational linguistics. first and second international symposia, Rocquencourt, France, October 2007; Providence, RI, USA, May 2008; Revised selected and invited papers*. Lecture Notes in Artificial Intelligence 5402. Berlin; Heidelberg: Springer-Verlag, 2009.

[7] Dayashankar Mohanlal Joshi. “Pārini’s taddhita affixation rules”. Ph.D. dissertation. Philadelphia: University of Pennsylvania, 1969.

[8] Paul Kiparsky. “On the architecture of Pāṇini’s grammar”. In: *Sanskrit computational linguistics. first and second international symposia, Rocquencourt, France, October 2007; Providence, RI, USA, May 2008; Revised selected and invited papers*. Ed. by Gérard Huet, Amba Kulkarni, and Peter M. Scharf. Lecture Notes in Artificial Intelligence 5402. Berlin; Heidelberg: Springer-Verlag, 2009, pp. 33–94.

[9] Paul Kiparsky and J. F. Staal. “Syntactic and semantic relations in Panini”. In: *Foundations of Language* 5 (1969), pp. 83–117.

[10] Amrith Krishna and Pawan Goyal. “Towards automating the generation of derivative nouns in Sanskrit by simulating Pāṇini”. In: *Sanskrit and computational linguistics. select papers presented at the 16th World Sanskrit Conference in the ‘Sanskrit and the IT world’ section 28 June – 2 July 2015, Sanskrit Studies Center, Silpakorn University, Bangkok*. Ed. by Amba Kulkarni. New Delhi: D. K. Publishers, 2016, pp. 157–193.

[11] Amba Kulkarni, ed. *Sanskrit and computational linguistics. select papers presented at the 16th World Sanskrit Conference in the ‘Sanskrit and the IT world’ section 28 June – 2 July 2015, Sanskrit Studies Center, Silpakorn University, Bangkok*. New Delhi: D. K. Publishers, 2016.

[12] Peter M. Scharf. “Levels in Pāṇini’s *Aṣṭādhyāyī*”. In: *Sanskrit computational linguistics. third international symposium, Hyderabad, India, January 2009, proceedings*. Ed. by Amba Kulkarni and Gérard Huet. Lecture Notes in Artificial Intelligence 5406. Berlin; Heidelberg: Springer-Verlag, 2009.

[13] Peter M. Scharf. “Modeling Pāṇinian grammar”. In: *Sanskrit computational linguistics. first and second international symposia, Rocquencourt, France, October 2007; Providence, RI, USA, May 2008; Revised selected and invited papers*. Ed. by Gérard Huet, Amba Kulkarni, and Peter M. Scharf. Lecture Notes in Artificial Intelligence 5402. Berlin; Heidelberg: Springer-Verlag, 2009.

[14] Peter M. Scharf. “On the semantic foundation of Pāṇinian derivational procedure: the derivation of *kumbhakāra*”. In: *Journal of the American Oriental Society* 131 (2011), pp. 39–72.

[15] Peter M. Scharf. “Rule selection in the *Aṣṭādhyāyī* or Is Pāṇini’s grammar mechanistic?” In: *Studies in Sanskrit grammars*. proceedings of the Vyākaraṇa section of the 14th World Sanskrit Conference, 1–5 September 2009, Kyoto University, Kyoto. Ed. by George Cardona, Ashok Aklujkar, and Hideyo Ogawa. New Delhi: D. K. Printworld, 2011, pp. 319–350.

[16] Peter M. Scharf. “Chapter 11. Linguistics in India”. In: *Oxford handbook of the history of linguistics*. Ed. by Keith Allan. Oxford: Oxford University Press, 2012, pp. 230–264.

[17] Peter M. Scharf. “Are taddhita affixes provided after prātipadikas or padas?” In: Pāṇini and the Pāṇinīyas of the 16th–17th century C.E., trosième atelier du projet ANR PP16-17. (Institut Français de Pondichéry, Pondicherry, Oct. 14–16, 2014). 2014. Forthcoming.

[18] Peter M. Scharf. “On the status of nominal terminations in upapada compounds”. In: *Proceedings of the Vyākaraṇa section of 16th World Sanskrit Conference, 28 June–2 September 2015, Sanskrit Studies Center, Silpakorn University, Bangkok*. Ed. by George Cardona and Hideyo Ogawa. New Delhi: D. K. Printworld, 2015, pp. 287–316.

[19] Peter M. Scharf. “An XML formalization of the *Aṣṭādhyāyī*”. In: *Sanskrit and computational linguistics. select papers presented at the 16th World Sanskrit Conference in the ‘Sanskrit and the IT world’ section 28 June – 2 July 2015, Sanskrit Studies Center, Silpakorn University, Bangkok*. Ed. by Amba Kulkarni. New Delhi: D. K. Publishers, 2016, pp. 77–102.

[20] Peter M. Scharf, Pawan Goyal, Anuja Ajotikar, and Tanuja Ajotikar. “Voice, preverb, and transitivity restrictions in Sanskrit verb use”. In: *Sanskrit syntax. selected papers presented at the seminar on Sanskrit syntax and discourse structures, 13–15 June 2013, Université Paris Diderot, with a bibliography of recent research by Hans Henrich Hock*. Ed. by Peter M. Scharf. Providence: The Sanskrit Library, 2015, pp. 157–201.

Identification of Aspectual Pairs of Verbs Derived by Suffixation in the Lexical Database DeriNet

Magda Ševčíková, Adéla Kalužová, and Zdeněk Žabokrtský

Institute of Formal and Applied Linguistics
Charles University, Czech Republic
E-mail: sevcikova|kaluzova|zabokrtsky@ufal.mff.cuni.cz

Abstract

The present paper describes a semi-automatic method of adding derivational links to the lexical database DeriNet by identifying verbs which are derived by suffixation and constitute aspectual pairs. It briefly introduces the notion of aspect in Czech and discusses the account of aspect in the Czech linguistic literature and in existing data resources. As its main focus, it presents an approach to identifying aspectual pairs based on extraction of such pairs from the VALLEX valency dictionary, identification of suffix substitution rules and subsequent manual annotation, which resulted in the addition of almost 6,000 derivational links into the DeriNet database.

1 Introduction

Aspect has been discussed mainly as a grammatical (inflectional) category of verbs in the theoretical description of Czech and other Slavic languages [8, 2, 12]. However, since it is expressed by derivational affixes in Czech verbs, we have recognized and used the category as an important feature in modelling the verb-to-verb derivation in the large lexical database of derivational relations DeriNet [18].[1]

The present paper describes the process of how aspectual pairs of verbs based on suffixation were identified in DeriNet and how the corresponding derivational links between these verbs were established in the database data. The described results are part of the changes between DeriNet version 1.3 and the newest release, DeriNet 1.4. Both of these versions of the database contain an identical set of lemmas, they only differ in the way individual lemma nodes are connected.

In Section 2, we resume basic linguistic facts about the category of aspect in Czech and provide an overview of language data resources available for Czech that contain some aspect-relevant information. After a brief introduction into the DeriNet database with a focus on verbs (Sect. 3), the procedure of identifying and

[1] http://ufal.mff.cuni.cz/derinet

establishing derivational links between verbs related by aspect is presented as two subsequent subtasks. First, a core set of aspectual pairs was extracted from an existing valency dictionary of Czech verbs (Sect. 4). As the second subtask (Sect. 5), the data was used to compile a list of final strings in which the aspectual pairs differ. The final string patterns were used to search the DeriNet database for further aspectual pairs which were, subsequently, organized according to the adopted criteria.

2 The category of aspect in the linguistic literature and in existing language data resources for Czech

2.1 Aspect as the linguistic category

In Czech, perfective and imperfective verbs are distinguished with respect to the category of aspect. If two verbs share the lexical meaning and differ just in the aspectual meaning (i.e. in the complex vs. continuous representation of the given action; [14]), they are considered to constitute a pure aspectual pair. Only pairs of verbs that differ in suffixes are considered pure aspectual pairs in a narrow sense (e.g. [6]; ex. (1)), while according to broader approaches (e.g. [8] or [3]) counterparts derived by prefixes are accepted as pure aspectual counterparts, too (see ex. (2)).

(1) $dát_{Vpf}$ – $dávat_{Vimpf}$ ‘to give’

(2) $vařit_{Vimpf}$ – $uvařit_{Vpf}$ ‘to cook’

Within the complex grammatical system of Czech, aspect is one of six inflectional categories that are conveyed by verbal forms in Czech (besides person, number, tense, mood, and verbal voice; cf. [7]).[2] Out of these categories, aspect is the only one that is not expressed in Czech by inflectional affixes cumulatively with other inflectional categories, but rather by agglutinative (derivational) affixes, namely by suffixes and (as we adhere to the broader approach to pure aspectual pairs) by prefixes.

However, if we change the function-to-form approach above (i.e., the aspectual meanings are expressed by affixes) to the form-to-function approach, we see that the changes in aspect are just one of the functions of suffixation and prefixation of verbs. Apart from it, the prefix modifies the meaning of the base verb (ex. (3)); the aspect may, or may not change at the same time ((3a) vs. (3b)). Suffixes are further used to derive iterative verbs from base imperfectives (ex. (4)), or secondary imperfectives from prefixed perfectives (ex. (5)).[3]

[2] In addition to the above mentioned categories, the category of gender is marked in the past and passive forms of verbs, too.

[3] These types of suffixation and prefixation are omitted in the sequel of the paper.

(3) (a) *vařit*$_{Vimpf}$ ‘to cook’ → *převařit*$_{Vpf}$ ‘to boil’
(b) *dávat*$_{Vimpf}$ ‘to give’ → *přidávat*$_{Vimpf}$ ‘to add’

(4) *dávat*$_{Vimpf}$ ‘to give’ → *dávávat*$_{Vimpf.iter}$ ‘to give’

(5) *převařit*$_{Vpf}$ ‘to boil’ → *převařovat*$_{Vimpf}$ ‘to boil’

As the functions of the affixes are closely interconnected and there is no theoretical consensus on how to differentiate them, in our project of establishing the database of derivational relations DeriNet (Sect. 3), we decided to deal with all prefixed and suffixed verbs equally without distinguishing those in which the affix fulfills the aspectual (inflectional) function. Pure aspectual pairs of verbs thus have been treated as derivationally related in the database. The paper is limited to pairs formed by suffixation.

2.2 Aspect in language data resources

The category of aspect was assigned to individual verbs in several data resources existing for Czech. A set of resources, which, we know, is not exhaustive but which we find sufficiently representative, is listed in this section. Table 1 presents their features relevant for adding aspectual links into DeriNet:

- By machine tractability we mean that an electronic version exists that contains an explicit markup of the resource’s logical structure (i.e., not only formatting).
- By a permissive license we mean that the resource is available to us under a license that allows us using it for the development of DeriNet and releasing DeriNet under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.
- Perfective and imperfective verbs are distinguished in all presented resources, but are not organized into tuples in some resources (i.e., corresponding aspectual counterparts are not interlinked).
- Similarly, only some of the resources explicitly connect iterative verbs to their base imperfective non-iterative counterparts.
- Last but not least, there are order-of-magnitude differences in the number of verbs covered by the individual resources.

First, there are high-quality traditional dictionaries such as *Slovník spisovného jazyka českého* (SSJČ, Dictionary of Standard Czech Language; [1]). However, such dictionaries were created for human users and are not fully machine tractable, at least not on the level of dictionary microstructure. In the case of SSJČ we are aware of an XML-ized version, but it contains more or less only formatting markup (and is not publicly available anyway).

Another example of a Czech dictionary intended for human users, this time focused specifically on verbs and their valency, is *Slovesa pro praxi* (SPP, Verbs for practice; [17]).

Then we are aware of several primarily electronic dictionaries, whose main focus is on verb valency in Czech again, but they contain some information on aspect too, such as the dictionaries BRIEF [13], VerbaLex [5], and VALLEX [10].

In addition, information about aspect of particular verbs is contained in two high-coverage morphological analysers: MorfFlex CZ [4] and Ajka [16]. However, none of them interconnects aspectual counterparts.

Table 1 shows that the valency lexicon VALLEX is the only resource that provides information on verbal aspect as well as on grouping of lexically related verbs differing in aspect, and is available to us under a permissive license at the same time. That is why this lexicon was used as the primary resource on aspectual linking information in Section 4.

Obviously, VALLEX has substantially smaller vocabulary coverage compared to the above mentioned morphological analysers. Thus we complemented it with a broader coverage approach which, however, required some annotation effort (Sect. 5).

Resource property	SSJČ	SPP	BRIEF	VerbaLex	VALLEX	MorfFlex	Ajka
Fully machine tractable	NO	NO	YES	YES	YES	YES	YES
Permissive license	NO	NO	NO	NO	YES	YES	NO
Pf./impf. connected	YES	YES	NO	YES	YES	NO	NO
Iter. connected to base impf.	YES	NO	NO	NO	YES	YES	NO
Number of verbs	25k	0.7k	15k	11k	4.5k	44k	42k

Table 1: Language data resources relevant for aspect in Czech.

3 Verbs in the lexical database DeriNet

DeriNet is a lexical resource containing more than 1 million nouns, adjectives, verbs, and adverbs of Czech; verbs are the smallest group containing 54,617 lexemes. Pairs of base and derived words have been searched for by semi-automatic methods and connected by edges that represent derivational relations. The edges are oriented from the base lexeme towards the derived lexemes. In the current design of the database, each derived lexeme can be assigned to at most one base lexeme. Thus, the derivational nests can be viewed as rooted trees. In accord with the terminology commonly used to describe trees, we use the term parent to refer to the base lexeme, and the term child to refer to the lexeme derived from it.

As for the deverbal derivatives in DeriNet, deverbal nouns, adjectives, and adverbs were connected with the particular base verbs in the previous version of De-

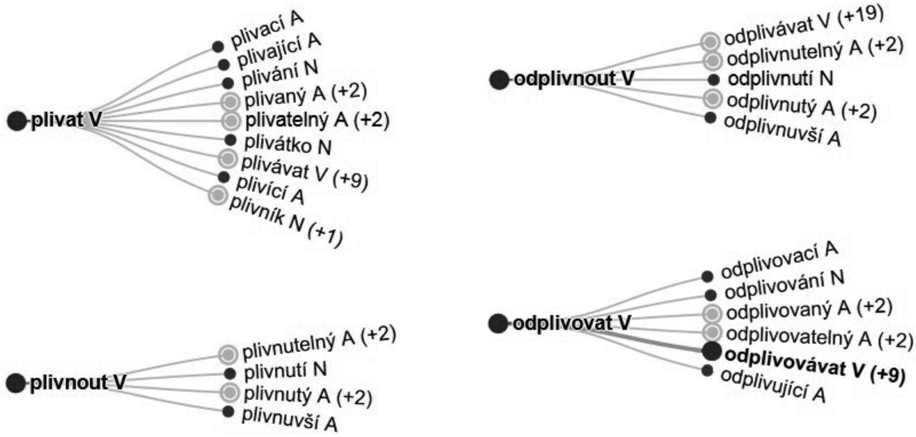


Figure 1: DeriNet 1.3: Four separate trees with the roots corresponding to the verbs *plivat*.impf 'to spit', *plivnout*.pf 'to spit', *odplivnout*.pf 'to spit out', and *odplivovat*.impf 'to spit out'

riNet (DeriNet 1.3). Concerning the derivation of verbs from verbs, links between iterative verbs and their base imperfective verbs have been satisfactorily resolved already in DeriNet 1.3, mainly because of the fact that this type of derivation is fairly regular. Thus, about 24,000 iterative verbs were assigned to their derivational ancestors. All other types of verb-to-verb derivation had been left unprocessed in DeriNet 1.3.

Fig. 1 shows an example of verbs which are derivationally related but were not connected to each other in DeriNet 1.3. The only exception is the derivational link between *odplivovat*$_{Vimpf}$ 'to spit out' and the iterative *odplivovávat*$_{Vimpf.iter}$ 'to spit out' marked red in the bottom right tree.

In this paper, we use a method of tree representation in which the trees are ordered from left to right, i.e. the leftmost node belongs to the base word, the words immediately at its right side are derived from it etc. The method also allows us to collapse edges which are not interesting for the example at hand. If edges leading from a certain node to its children have been collapsed, their number is given in brackets after the corresponding lemma. For example, the node description *odplivovávat V (+9)* means that there are 9 more words derived from the lemma *odplivovávat* in the database but they are not shown in the tree.

4 Extraction of aspectual pairs from the valency dictionary

4.1 Identifying verbs related by aspect

The very first step in identifying aspectual pairs of verbs in DeriNet was to extract groups of lexically related verbs differing in their aspect labels from the valency lexicon VALLEX. Verbs that do not constitute aspectual pairs had to be excluded from the extracted groups. Thus, the iterative verb *dávávat* was excluded from the group in (6) (getting the aspectual pair *dát* – *dávat* mentioned in (1)), whereas in ex. (7) and (8) two verbs that share both the lexical and aspectual meaning are included (cf. two imperfectives *střetat* and *střetávat* that are at hand for the perfective *střetnout* in (7); in (8), two perfective counterparts *hnout* and *hýbnout* are available for the imperfective *hýbat*). In such cases, the decision on which of the competing verbs should be preferred, i.e. marked as the direct aspectual counterpart of the single verb with a different aspect, was based on length of its affix or on corpus frequency. In (7), *střetat* is considered the direct aspectual counterpart of *střetnout* because of its simpler affix (compared to the one in *střetávat*), although the latter is more frequent and has been attested by 9,336 hits in the SYNv5 corpus [9] (vs. 334 hits for *střetat*).[4] In (8), the perfective *hnout* was preferred to *hýbnout* with 35,230 vs. 409 hits in the SYNv5 corpus. Length was not considered a relevant criterion in this case because the difference between the two verbs is found in the stem, not in the affix. In both of these cases, the non-preferred verb was marked as a more complex variant of the preferred one, rather than a member of the aspectual pair.

(6) *dát*$_{Vpf}$ – *dávat*$_{Vimpf}$ – *dávávat*$_{Vimpf.iter}$ ‘to give’

(7) *střetnout*$_{Vpf}$ – *střetat*$_{Vimpf}$ – *střetávat*$_{Vimpf}$ ‘to clash’

(8) *hnout*$_{Vpf}$ – *hýbnout*$_{Vpf}$ – *hýbat*$_{Vimpf}$ ‘to move’

In total, 1,365 aspectual pairs of verbs were identified in the DeriNet database using the grouping of lexically related verbs from the VALLEX data.

4.2 Determining the base and derived verb in the aspectual pair

Since DeriNet data are organized into the parent-child pairs according to the direction of derivational relations, in each aspectual pair it must be determined which verb is to be represented as the base word (parent) and which of them as the derivative (child). The decision was intuitive with pairs in which one of the verbs has a more complex morpheme structure (an extra suffix) than the other member – the shorter verb is considered the base and the longer one as derived from it (i.e. *dát*→*dávat* in ex. (1)).

[4] The SYNv5 (corpus of the SYN series, version 5) is the largest currently available representative corpus of contemporary Czech. It contains 3.836 billion words.

Nevertheless, if the relation between the aspectual counterparts corresponded to resuffixation rather than suffixation and both verbs had a similar string length (see ex. (9)), speaker's intuition led to different results. Since theoretical linguistic literature provides no satisfactory answers to this issue (cf. [11, 15, 8]), we based the decision on the following criteria elaborated for this purpose:

- The general rule was that the derivational relation is oriented from the perfective verb (parent) to the imperfective one (child), referring to the aspect value assigned with the verbs in VALLEX; cf. ex. (10). We based this decision on the fact that perfective verbs seem to frequently be either shorter than their imperfective counterparts (and derivational parents tend to generally be shorter in various types of derivation) or are more often felt as unmarked.
- The direction was revised in pairs in which the child verb is shorter than the base one. For instance, in pairs containing a perfective with the suffix *-nou-* and an imperfective with *-a-*, the perfective expresses a punctual action and is considered the marked member of the aspectual pair. The direction of the relation (*Vpf*→*Vimpf*) thus has been inverted (to *Vimpf*→*Vpf*; ex. (11)).

(9) *skoč-i-t*$_{Vpf}$ → *skák-a-t*$_{Vimpf}$ 'to jump'

(10) *koup-i-t*$_{Vpf}$ → *kup-ova-t*$_{Vimpf}$'to buy'

(11) *štěk-a-t*$_{Vimpf}$ → *štěk-nou-t*$_{Vpf}$ 'to bark'

5 Finding aspectual pairs by string substitution rules

5.1 Patterns for substitution of final strings

In order to identify more aspectual pairs, we automatically extracted patterns for final string substitution from the VALLEX aspectual pairs and tried to apply them on verbs that have no parent in DeriNet 1.3. Technically, in most cases the final strings in the patterns did not correspond to suffixes in the linguistic sense: they usually contained also the infinitive ending *-t* (or, *-ci* in special cases) and sometimes one or more letters from the stem. If an alternation appeared in the stem, the alternating letters were included into the final string, too. This approach to alternations later proved successful, mainly because alternations are often influenced phonetically. Therefore, if a suffix is added to different words whose stems end similarly, it is likely to induce the same alternation, as *r*>*ř* in ex. (12) to (15). All these examples could be identified using the same substitution rule, namely *V-řít* → *V-írat*.

(12) *opřít*$_{Vpf}$ 'to lean' → *opírat*$_{Vimpf}$

(13) *utřít*$_{Vpf}$ 'to wipe' → *utírat*$_{Vimpf}$

(14) *umřít*$_{Vpf}$ 'to die' → *umírat*$_{Vimpf}$

(15) *zavřít*$_{Vpf}$ 'to close' → *zavírat*$_{Vimpf}$

The resulting list consisted of 183 substitution patterns. Some examples of such patterns are shown in ex. (16) and (17) together with candidate pairs of verbs which contain them. The pairs in (a) are correct, while those in (b) and (c) are ones which have been identified incorrectly. Different methods were used to eliminate such candidate pairs; see futher in this section.

(16) *V-it* → *V-nout*:
(a) $chladit_{Vimpf}$ 'to cool down' → $chladnout_{Vimpf}$ 'to cool down'
(b) *škrtit*$_{Vimpf}$ 'to strangle' $\not\rightarrow$ *škrtnout*$_{Vpf}$ 'to cross out'
(c) $rozhodit_{Vpf}$ 'to throw around' $\not\rightarrow$ $rozhodnout_{Vpf}$ 'to decide'

(17) *V-it* → *V-ovat*:
(a) $koupit_{Vpf}$ 'to buy' → $kupovat_{Vimpf}$ 'to buy'
(b) $radit_{Vimpf}$ 'to advise' $\not\rightarrow$ $radovat_{Vimpf}$ 'to rejoice'

5.2 Applying the patterns to the DeriNet data

The list of patterns was applied to the DeriNet data in order to find more aspectual pairs that were not covered by VALLEX; the search for candidate child verbs was limited only to those that had not been assigned a derivational ancestor before. This way, 5,578 candidate pairs of verbs related by aspect were identified.

The subsequent manual annotation was not carried out on all candidate pairs but on subsets identified by different criteria. First of all, the combination of aspect values in the candidate pairs was used mainly to filter out the pairs which should usually be correct (which followed the preferred *Vpf*→*Vimpf* pattern) and focus the attention on the candidates in which the aspectual characteristics of the parent and child were more unusual. Among such pairs there were those with the aspect pattern *Vimpf*→*Vpf* (ex. (16b)), *Vpf*→*Vpf* (ex. (16c)), or *Vimpf*→*Vimpf* (ex. (17b)). As has already been mentioned, all of the pairs in ex. (16) to (17) contain a combination of final strings which is very frequent in parent-child pairs, and thus were identified as derivational pair candidates when the substitution rules were applied. However, since those in (b) and (c) are not derivationally related, they had to be excluded in the manual annotation.

The criterion of frequency was also used to select unlikely candidate pairs. It was most helpful when identifying candidates in which the verbs followed one of the common final string patterns as well as a usual aspect pattern although each of them was a part of a different word formation family. This is a rather rare phenomenon in Czech and it occurs almost exclusively in cases where one of the words has a very low frequency in the language. An example of a correct pair is shown in ex. (18), while an example of an incorrect pair with the same final string pattern and aspectual properties follows in ex. (19). Note that the word *zpít* 'to get drunk' is not a frequent one in Czech, reaching only 1,004 corpus hits in the SYNv5 corpus [9], mainly because the same meaning is usually expressed by the derivationally related *opít* (19,976 hits). The suggested derived verb *zpívat* 'to sing' has 268,244 occurrences in SYNv5 corpus.

(18) *omdlít*$_{Vpf}$ 'to faint' → *omdlívat*$_{Vimpf}$ 'to faint'

(19) *zpít*$_{Vpf}$ 'to get drunk' ↛ *zpívat*$_{Vimpf}$ 'to sing'

The length criterion was only used to establish the direction of derivational edges, as in the aforementioned pairs of verbs with the suffixes *-nou-*$_{pf}$ and *-a-*$_{impf}$. However, it needs to be stressed that length was never the only criterion sufficient for the inversion of a derivational edge. While a longer word may contain multiple suffixes and therefore be a derivational child of a word containing less suffixes, it can also simply contain a single suffix which is longer than the suffix used in another word. In the latter case, which also includes the suffix pair *-nou-*$_{pf}$ and *-a-*$_{impf}$, other criteria like frequency or markedness were taken into consideration when deciding the direction of the derivational link.

5.3 Manual annotation

In the manual annotation, 968 pairs were marked as incorrect. The remaining 4,610 derivational edges were established in DeriNet data, in addition to the 1,365 edges between aspectual pairs found in VALLEX (Sect. 4). The edges between aspectual pairs of verbs are part of the recent release of the DeriNet database (DeriNet 1.4). Currently, almost 44,000 verb nodes are marked as children of other verb nodes in DeriNet 1.4, while in DeriNet 1.3 only about 24,000 verbs were identified as derived from verbs. The increase is caused by adding the links between aspectual counterparts (5,975 newly attached nodes) but also by connecting pairs derived by prefixation, which were however found using different methods not described in this paper.

In the current version of DeriNet, 10,660 verbs are not marked as children of other nodes. There are various reasons why. In some cases, they are correctly marked as tree roots: they can be either unmotivated (e.g. the verb *bít* 'to hit') or loan words (e.g. *abstrahovat* 'to abstract'). Furthermore, some word formation processes have not yet been fully resolved in DeriNet, for example the derivation of verbs from non-verbs as in *zdravět* 'to get healthy', which is derived from the adjective *zdravý* 'healthy', or compounding (e.g. *zesteronásobnit* 'to multiply by hundred').

An example of several derivational links between verbs which have been newly established in DeriNet 1.4 is shown in Fig. 2. As opposed to the individual trees in Fig. 1, here multiple trees with a verbal root have been merged into one. The first highlighted link from *plivat*$_{Vimpf}$ 'to spit' toward *plivnout*$_{Vpf}$ 'to spit' has been found using the method described in the present paper and connects pure aspectual counterparts. The second relation, between *plivnout*$_{Vpf}$ 'to spit' and *odplivnout*$_{Vpf}$ 'to spit out' is that of derivation by prefixation with a change in meaning but without a change of aspect. Such relations have also been included in DeriNet 1.4 but the methods of their identification are beyond the scope of this paper.

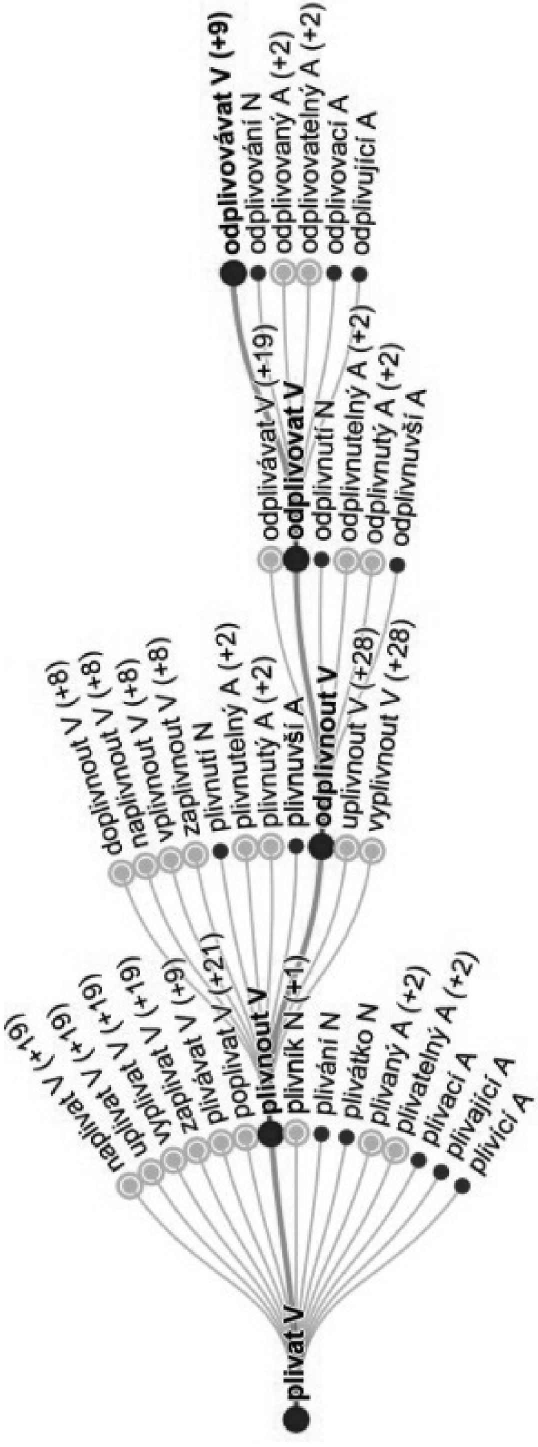


Figure 2: DeriNet 1.4: The derivational tree with the root verb *plivat*.impf 'to spit' to which other (both directly and indirectly) derivationally related verbs are linked

The next relation, from *odplivnout*$_{Vpf}$ ‘to spit out’ to *odplivovat*$_{Vimpf}$ ‘to spit out’ is that of secondary imperfectivization. Such pairs of aspectual counterparts have also been identified using the above described methods. The link between *odplivovat*$_{Vimpf}$ ‘to spit out’ and *odplivovávat*$_{Vimpf.iter}$ ‘to spit out’ has been preserved from DeriNet 1.3.

6 Conclusions

In the task of identifying aspectual pairs of verbs derived by suffixation, we started with the extraction of lexically related verbs with different aspect from an existing valency dictionary, VALLEX [10]. Because of its limited coverage, we used further semi-automatic procedures to identify more aspectual pairs. Our method was based on final string substitution patterns which were extracted from the pairs in VALLEX and subsequently sought for in the remaining part of DeriNet data. This way, we managed to establish almost 6,000 derivational links between aspectual pairs, which, by the way, makes DeriNet 1.4 probably the biggest freely available machine-tractable data resource on aspectual pairing in Czech.

Suffixation is not the only morpohological means for deriving verbs from verbs in Czech. Perhaps not even the dominating one, and also not the only one that possibly results in changed aspect. However, the presented topic is an important piece in the mosaic of verb-to-verb derivational morphology in Czech. After completing the mosaic with other shards, especially with those related to prefixation, we will hopefully arrive to the point in which our empirical evidence collected in DeriNet for studying verb derivations will be close to perfect.

7 Acknowledgement

This work was supported by the Grant No. GA16-18177S of the Grant Agency of the Czech Republic. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

[1] B. Havránek et al. *Slovník spisovného jazyka českého*. Academia, Praha, 1960–1971. Also accessible from http://ssjc.ujc.cas.cz.

[2] B. Comrie. *Aspect*. Cambridge University Press, Cambridge, 1976.

[3] D. Šlosar. *Slovotvorný vývoj českého slovesa*. UJEP, Brno, 1981.

[4] J. Hajič and J. Hlaváčová. MorfFlex CZ. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, 2013. http://hdl.handle.net/11858/00-097C-0000-0015-A780-9.

[5] D. Hlaváčková and A. Horák. Verbalex–new comprehensive lexicon of verb valencies for Czech. In *Proceedings of the Slovko Conference*, 2005.

[6] A. V. Isačenko. Slovesný vid, slovesná akce a obecný charakter slovesného děje. *Slovo a slovesnost*, 21:9–16, 1960.

[7] M. Komárek, J. Kořenský, J. Petr, and J. Veselková. *Mluvnice češtiny 2*. Academia, Prague, 1986.

[8] F. Kopečný. *Slovesný vid v češtině*. Nakladatelství ČSAV, Praha, 1962.

[9] M. Křen, V. Cvrček, T. Čapka, A. Čermáková, M. Hnátková, L. Chlumská, T. Jelínek, D. Kovářiková, V. Petkevič, P. Procházka, H. Skoumalová, M. Škrabal, P. Truneček, P. Vondřička, and A. Zasina. Corpus SYN, version 5 from April 24, 2017. UCNK FF UK, Prague, 2017. http://www.korpus.cz.

[10] M. Lopatková, V. Kettnerová, E. Bejček, A. Vernerová, and Z. Žabokrtský. *Valenční slovník českých sloves VALLEX*. Karolinum, Praha, 2016. Also accessible from http://ufal.mff.cuni.cz/vallex.

[11] M. Komárek. *Příspěvky k české morfologii*. Periplum, Olomouc, 2006.

[12] I. A. Mel'čuk. *Cours de morphologie générale. Vol. 2*. Presses de l'Université de Montréal, Montréal, 1976.

[13] K. Pala and P. Ševeček. Valence českých sloves. In *Proceedings FFBU*, pages 41–54, Brno, 1997.

[14] J. Panevová and P. Sgall. Slovesný vid v explicitním popisu jazyka. *Slovo a slovesnost*, 33:294–303, 1973.

[15] I. Poldauf. Souhrnný pohled na vid v nové češtině. *Slovo a slovesnost*, 25:46–56, 1964.

[16] R. Sedláček and P. Smrž. A New Czech Morphological Analyzer ajka. In *LNCS / Lecture Notes in Artificial Intelligence. Proceedings of the 4th International Conference Text, Speech and Dialogue (TSD 2001)*, pages 100–107, Berlin, 2001. Springer.

[17] N. Svozilová, H. Prouzová, and A. Jirsová. *Slovesa pro praxi*. Academia, Praha, 1997.

[18] M. Ševčíková and Z. Žabokrtský. Word-Formation Network for Czech. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 1087–1093, Reykjavik, Iceland, May 2014.

DErivCelex: Development and Evaluation of a German Derivational Morphology Lexicon based on CELEX

Elnaz Shafaei Bajestan, Diego Frassinelli,
Gabriella Lapesa and Sebastian Padó
Institute for Natural Language Processing
University of Stuttgart
E-mail: shafaeez,frassido,lapesaga,pado@ims.uni-stuttgart.de

Abstract

Derivational lexicons group words into derivational families, that is, equivalence classes of derivationally related words, and play an important prerequisite in computational studies of derivational morphology. While several such lexicons exist for a number of languages, they lack in comparability. We present an algorithm that extracts such lexicons from the German morphological layer of CELEX, a lexical database that is available for English, Dutch, and German, thus making a step towards the creation of more comparable derivational lexicons at least for these languages. We evaluate the result, DErivCelex, against DErivBase, a large derivational lexicon created semi-automatically. We find that DErivCelex excels in precision, but lacks in recall. Further analysis shows that a substantial part of the recall gap is due to different assumptions about the limits of what can be considered a derivational relationship. We conclude by presenting a refined version of DErivCelex that builds on a more liberal definition of derivation and improves recall.

1 Introduction

Processing of morphological information is a well established task in computational linguistics, often constituting the first step in an NLP pipeline. The earliest focus of the research community was dealing with inflection in the form of lemmatization or stemming (Porter [13]). In recent years, computational semantics research has shown more interest in the NLP aspects of derivation (Padó et al. [10], Cotterell et al. [2]).

Such research requires *derivational lexicons* that minimally group together derivationally related words into *derivational families*. There are two main families of approaches to create such lexicons as clusters of derivationally related lemmas, e.g., {*ask_V asker_N, asking_N, asking_A*}. The first one is to exploit existing dictionaries or other lexical resources. Examples are CatVar (Habash and Dorr

[6]) for English, Démonette (Hathout and Namer [7]) for French, and DeriNet (Žabokrtský et al. [16]) for Czech. The second approach is to acquire derivational lexicons from corpora. Examples of this approach are DErivBase for German (Zeller et al. [17]) and DErivBase.HR for Croatian (Šnajder [15]): hand-written derivational rules are employed to map base words into potential derived words, and corpus information is used as a filter (if the potential derived word is attested in the reference corpus, it is added to the resource).

A problem that all previous studies share is that the proposed methods are to a large extent *language-specific*: resource-based approaches have to build on whatever (typically idiosyncratic) resources there are for a given language. Corpus-based approaches are not only reliant on language-specific corpora but also involve manual rule creation, which is hard to standardize. Consequently, in the present state of affairs, it is very difficult to make *valid cross-lingual comparisons* on the basis of these lexicons, for example regarding derivational factors like productivity (Plag [12]) or psycholinguistic phenomena like morphological priming (Kempley and Morton [8]).

In this paper, we present a first step towards a greater degree of cross-lingual comparability of derivational lexicons. Our approach is to automatically extract derivational lexicons from a *multilingual family of dictionaries*, namely CELEX (Baayen et al. [1]). CELEX is a psycholinguistic lexical database available for English, German, and Dutch that was carefully verified by experts and is widely used in psycholinguistics. CELEX, however, does not explicitly contain derivational families and has a limited lemma coverage. Our contributions in this paper are: (a), we present an algorithm that automatically extracts derivational families from CELEX; (b), we evaluate the result for German, which we call DErivCelex, against the existing German DErivBase derivational lexicon to better understand the size–quality trade-off.

2 Extracting Derivational Families from CELEX

As mentioned above, CELEX provides an array of information about lexical units at different linguistic levels. Four fields in the morphological section are relevant for grouping lemmas into derivational families:

1. `Head`: the canonical form of a stem.
2. `MorphStatus`: the morphological category of a stem. The stem can either be monomorphemic, complex, a zero derivation, a lexicalized flection, undetermined, or irrelevant.
3. `ImmClass`: the word class labels for the elements identified in the stem's immediate segmentation.
4. `StrucLab`: the complete hierarchical segmentation of the stem. For example, the segmentation of the noun *Tagelöhner (day laborer)* is: `(((Tag)[N],(e)[N|N.N],(Lohn)[N])[N],(er)[N|N.])[N].`

The exact procedure followed to populate the derivational families is described in

```
   input  : The lemma lexicon file (gml.cd) from the German morphology section of CELEX2
   output : derivational families of DErivCelex
1  FamilyIDs ⟵ ∅ ;                    /* stores a family ID for each lemma */
2  Headwords ⟵ ∅ ;                    /* stores a headword for each lemma */
3  foreach line in gml.cd do
4      /* If lemma is Monomorph or Compound or Derivational compound,
           create a new derivational family                          */
5      if MorphStatus = 'M' or ImmClass has the pattern of a Compound or ImmClass has the
        pattern of a Derivational Compound then
6          FamilyIDs [ StrucLab ] ← new family ID;
7          Headwords [ StrucLab ] ← Head + '_' + GetPOS(StrucLab);
8      end
9      /* If lemma is a zero or normal derivation, traverse tree    */
10     else if MorphStatus = 'Z' or ImmClass has the pattern of a Derivation then
11         Stem ← StrucLab;
12         while Stem is a result of a zero derivation or a derivation do
13             FamilyIDs [ Stem ] ← new family ID;
14             Base ← GetBase(Stem);
15             POS ← GetPOS(Stem);
16             Headwords [ Stem ] ← Base + '_' + POS;
17             MergeFamilies(FamilyIDs [Stem ], FamilyIDs [Base ]);
18             Stem ← Base
19         end
20     end
21 end
```

Algorithm 1: Extract derivational families from CELEX.

Algorithm 1. The idea behind the method is that all words that share the same head of the same part of speech (lines 5-8) are grouped into the same family. However, since compounding is very productive in Dutch and German, we need to ensure that the lemmas in each family are a) the result of a derivational process or a chain of derivations applied to a monomorph (the head) or b) they are the result of a derivation or a chain of derivation applied to a compound. As a result, each derivational family in DErivCelex can be headed by either a monomorph or a compound, but not both. For example, German *Bürger (citizen), bürgerlich (civic)* will end up the same family since they share the head *Bürger*. The corresponding *Grossbürger, grossbürgerlich (bourgeois)* will be grouped in another family, headed by *Grossbürger*.

To tease apart compounding and non compounding processes, we rely on the CELEX definitions of compounds (i.e., the joining of two stems into one new stem either with or without a link morpheme) and derivational compounds (i.e., new compound formation in combination with a derivational affix either as a triform or a quaternary split), as opposed to derivations (i.e., forming a new stem through prefixation, circumfixation, postfixation with one affix, and postfixation with two affixes). To distinguish these cases, the extraction algorithm needs to examine the morphological structure recursively (lines 10-20).

3 Comparing DErivBase and DErivCelex

We applied Algorithm 1 to the German CELEX, resulting in a derivational lexicon that we call *DErivCelex*. We now compare DErivCelex with DErivBase ver. 1.4.1 (Zeller et al. [17]), the largest derivational lexicon for German. DErivBase was developed on the basis of a very large set of lemmas, covering all content words in SdeWaC (Faaß and Eckart [4]) with frequency above 4. At the same time, the DErivBase construction method was semi-automatic, and the resource is known to contain errors. The goal of this section is to compare and contrast the properties of DErivBase and DErivCelex.

Resource sizes and structures. Overall, DErivCelex contains 46,667 lemmas grouped into 27,859 families, in contrast to the 280,336 lemmas in DErivBase, grouped into 228,213 families. The two resources share 36,867 lemmas (79% of the coverage of DErivCelex). The upper part of Table 1 reports statistics on the family sizes of the two resources. Although DErivCelex has a significantly smaller coverage, the percentage of non-singleton families[1] is three times larger than for DErivBase which captures the "long tail" from the corpus. Thus, the numbers of lemmas with non-trivial derivational information are closer: 65K for DErivBase vs. 16K for DErivCelex. As the statistics on family size and the plots in Figure 1 show, the distributions over family sizes are roughly in line. We see this convergence as a good sign.

To compare the two resources on a more equal footing, we also analysed their intersection, which can be defined on various levels. We focus on the family level by defining the concept of *corresponding families* as follows: If the head of a family f in DErivCelex also exists in DErivBase as a member of family f', then f and f' are corresponding families. We consider the union of all corresponding families in the two resources, respectively. Note that this definition covers families including lemmas that are not present in the other resource.

We found 19,277 such families on the DErivCelex side and 17,126 on the DErivBase side – note that the number is smaller for DErivBase because according to our definition of derivational family, multiple DErivCelex families can correspond to the same DErivBase family (cf. the *ziehen* example below). Their statistics are shown in the lower half of Table 1. As expected, the "shared" families in DErivBase are substantially larger: it is indeed the "long tail" of the DErivBase singleton families that DErivCelex does not capture. The numbers of DErivCelex also go up, but only a little. The numbers show that the DErivBase families are substantially larger than the DErivCelex families. This is supported by the examples for corresponding families in Table 2: The family for the adjective *weitschweifig (prolix)* contains the same lemmas which are in both resources; similarly for the noun *Weitsicht (far-sightedness)*. On the other hand, the families of the *Werk (factory/creation)* and *unterziehen (to undergo)* are very much larger in DErivBase.

[1] Singleton families are those containing only one lemma.

Resource	Singletons	Nonsingletons	Family size, mean (SD)	
	families (%)	families (%)	with singletons	without singletons
DErivBase (n = 228,213)	92	8	1.23 (2.23)	4.01 (7.57)
DErivCelex (n = 27,859)	79	21	1.68 (2.56)	4.22 (4.80)
DErivBase (n = 17,126)	54	46	7.50 (20.41)	17.06 (29.60)
DErivCelex (n = 19,277)	78	22	1.79 (2.94)	4.69 (5.44)

Table 1: Number and size of families in DErivBase and DErivCelex. Above: Complete resources. Below: Corresponding families.

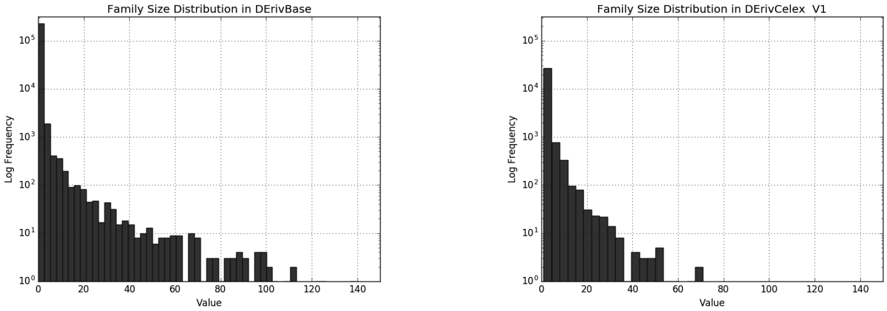


Figure 1: Family size distribution for DErivBase (left) and DErivCelex (right).

Shared lemma	DErivCelex	DErivBase	Overlap size
weitschweifig_A (prolix)	2	2	2
Weitsicht_N (far-sightedness)	3	4	3
Werk_N (factory/creation)	8	79	4
unterziehen_V (undergo)	1	97	1

Table 2: Examples of corresponding families between DErivBase and DErivCelex

These differences arise from fundamentally different assumptions about what constitutes morphological derivation, and reflect the ongoing discussion about the definition of the notion derivation (Olsen [9]). CELEX, and thus DErivCelex, follows a tradition in German linguistics that treats prefixation as a word formation process distinct from derivation (Fleischer [5]). As a result, the derivational families extracted from CELEX tend to be *more cautious*. For example, *unterziehen* is analysed as a compound and ends up in a derivational singleton family. In contrast, DErivBase includes prefixation in derivation (Erben [3], Smolka et al. [14]). Consequently, *unterziehen* is analysed as a prefix derivation with *unter-* and becomes part of the huge *ziehen* derivational family. Similar, but less clear, differences exist with regard to the analysis of stem changes: In DErivBase, *Werk (work/opus)* shares a broad family with lemmas like *wirken (to effect), Wirkung (effect/impact)*, while the DErivCelex family is considerably more narrow. In section 4, we will reconsider the definition of derivation assumed by CELEX and DErivCelex.

Correctness of DErivCelex To evaluate DErivCelex, we employed the same evaluation framework developed for DErivBase by Zeller et al. [17]. The evaluation involves two gold standard samples, targeting different aspects of the performance of a derivational lexicon: its coverage (*recall sample*) , and the correctness of the information it contains (*precision sample*).

Coverage is quantified based on a *recall sample*, which consists of 2000 lemma pairs. For each lemma pair $\{w_1, w_2\}$ in the sample, w_1 is a member of a non-singleton DErivBase family and w_2 is drawn from a set of potentially derivationally related words as computed by a string similarity measure. The pairs were manually annotated as derivationally related or unrelated, and the sample was used to compute *recall* (i.e., what percentage of all valid derivational relationships are represented in DErivBase).

Correctness is quantified based on a *precision sample*. It consists of 2000 lemma pairs of which w_1 and w_2 are members of the same DErivBase family (i.e., have been classified as derivationally related in DErivBase). Each pair was manually annotated as derivationally related or unrelated. This annotation was used to compute *precision* (i.e., what percentage of the pairs predicted to be derivationally related by DErivBase are actually correct).[2]

We evaluate DErivCelex on the same data. Note, however, that this puts DErivCelex at a disadvantage vis-à-vis DErivBase, since both samples are constructed to focus on lemmas covered by DErivBase and therefore contain lemmas from the "long tail". In fact, DErivCelex has coverage only for 1523 of the 4000 lemmas. For this reason, we additionally report *relative recall*, i.e.,'recall relative to coverage on the sample'.

The results are shown in Table 3. The precision of DErivCelex is very high at 0.93, higher than for the standard version of DErivBase and comparable to a high-precision variant reported in Zeller et al. [17]. We believe that this is quite a good result. Conversely, however, the recall of DErivCelex on the whole sample is very low, at 22%. Relative recall, which removes lemma coverage from the picture, is 43% – considerably higher than 22% but still far below DErivBase's 71%. We believe that a substantial part of the gap is due to the less restricted notion of derivation adopted by DErivBase compared to CELEX, which of course is also reflected in the gold standard.

[2] The need to draw two separate samples is that the number of actual derivational relations among all candidates for such relations is very small. Thus, any sampling technique that considers all candidates (which is necessary to compute recall) will, assuming reasonable sample sizes, contain so

	Coverage (# pairs)	Precision	Recall	Relative Recall
DErivBase	4000	0.83	0.71	0.71
DErivCelex	1523	0.93	0.22	0.43

Table 3: Evaluation against the DErivBase gold standard

4 Including Prefixation in Derivation: DErivCelex V2

As discussed in the previous section, CELEX treats prefix verbs (e.g., *unterziehen*, *vorgreifen*) as compounds. As a consequence, they are treated as heads of new derivational families and represented separately from their heads (e.g., *ziehen* and *greifen*). Since there are no striking linguistic reasons to keep prefixation and derivation separate, and it makes sense from a computational point of view to provide a unified treatment, we created a new version of DErivCelex that treats prefixation as a type of derivation (but abstained from touching the less clear cut field of stem changes). This involved changing the extraction procedure to reinterpret specific cases of composition (namely prefix verbs) as derivations, shown in Algorithm 2. For the purpose of this procedure, we defined prefix verbs as compositions of verbal bases with prefixes that are prepositions, adverbs, or adjectives. This covers 1,784 prefix verbs.

The output is a derivational morphology resource for German, called DErivCelex V2, with 46,667 lemmas and 26,196 families. The overall statistics for the number of families and the (non-)singleton percentages are presented for DErivCelex V2, compared to DErivBase, in the upper part of Table 4. Naturally, the number of lemmas in DErivCelex V2 remains at 46,667, unchanged from V1. The number of families has however decreased from 27,859 to 26,196, which leads to somewhat larger families (1.78 in V2 vs. 1.68 in V1). DErivCelex V2, with or without singletons included, has still larger families than DErivBase. There is no significant difference in the percentage of non-singleton families between DErivCelex V2 and DErivCelex V1. These findings are also evident in a longer tail in the Zipfian distribution of family size for DErivCelex V2 (figure 2) compared to the distribution of family size in DErivCelex V1.

We compute *corresponding families* between DErivBase and DErivCelex V2 as above. We found 17,867 corresponding families in DErivCelex and 16,316 in DErivBase. The lower part of table 4 looks into singleton and nonsingleton corresponding families. Regarding the percentage of non-singleton families, the difference between DErivCelex V2 and DErivBase is smaller than the difference between DErivCelex V1 and DErivBase. Furthermore, the average size of non-singleton families for DErivCelex V2 is closer to that of the DErivBase, compared to the same statistics for DErivCelex V1 and DErivBase. The corresponding families share, on average, 1.6 lemmas (min = 1, max = 68).

few true positives that it will only yield very rough estimates of precision, and vice versa.

```
input  : The lemma lexicon file (gml.cd) from the German morphology section of CELEX2
output : derivational families of DErivCelex V2
FamilyIDs ⟵ $\emptyset$ ;                /* stores a family ID for each lemma */
Headwords ⟵ $\emptyset$ ;                /* stores a headword for each lemma */
foreach line in gml.cd do
    /* If lemma is Monomorph, Compound, or Derivational compound,
       but not a Prefix Verb, create a new derivational family   */
    if (MorphStatus = 'M' or ImmClass has the pattern of a Compound or ImmClass has the
      pattern of a Derivational Compound) and (ImmClass does not have the pattern of a
      Prefix Verb) then
        FamilyIDs [ StrucLab ] ← new family ID;
        Headwords [ StrucLab ] ← Head + '_' + GetPOS(StrucLab);
    end
    /* If lemma is a Zero Derivation or a Derivation or a Prefix
       Verb, traverse the tree downwards                         */
    else if MorphStatus = 'Z' or ImmClass has the pattern of a Derivation or ImmClass has
      the pattern of a Prefix Verb then
        Stem ← StrucLab;
        while Stem is a result of a zero derivation or a derivation or a prefix verb do
            FamilyIDs [ Stem ] ← new family ID;
            Base ← GetBase(Stem);
            POS ← GetPOS(Stem);
            Headwords [ Stem ] ← Base + '_' + POS;
            MergeFamilies(FamilyIDs [Stem ], FamilyIDs [Base ]);
            Stem ← Base
        end
    end
end
```

Algorithm 2: Extract DErivCelex V2 from CELEX, treating prefix verbs as cases of derivations (changes shown in blue)

Resource	Singletons families (%)	Nonsingletons families (%)	Family size, mean (SD) with singletons	Family size, mean (SD) without singletons
DErivBase (n = 228,213)	92	8	1.23 (2.23)	4.01 (7.57)
DErivCelex V2 (n = 26,196)	79	21	1.78 (3.61)	4.78 (7.20)
DErivBase (n = 16,316)	59	41	5.70 (16.10)	13.63 (24.39)
DErivCelex V2 (n = 17,867)	79	21	1.94 (4.24)	5.55 (8.40)

Table 4: Number and size of families in DErivBase and DErivCelex V2. Above: Complete resources. Below: Corresponding families.

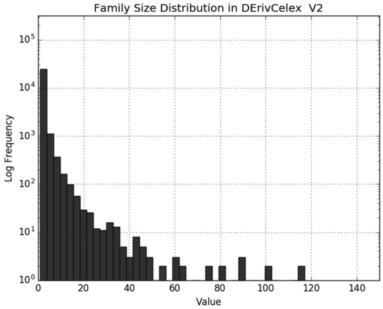


Figure 2: Family size for DErivCelex V2 derivational families

	Coverage	Precision	Recall	Relative Recall
DErivBase	4000	0.83	0.71	0.71
DErivCelex V1	1523	0.93	0.22	0.43
DErivCelex V2	1523	0.93	0.22	0.45

Table 5: Evaluation against the DErivBase gold standard

Taken together, the overall structure of DErivCelex V2 has changed from DErivCelex V1 towards DErivBase, having more populated families and compensating for the missing long tail of DErivBase in DErivCelex V1 to some extent. Naturally, DErivCelex V2 still has a much shorter tail than DErivBase as a result of its lexicon-based, as opposed to a corpus-based, methodology.

Has DErivCelex V2 also changed with regard to quantitative evaluation? The results are shown in Table 5. The precision has not changed from V1, which shows that the extension did not introduce wrong derivational relations. Unfortunately, the effect on the recall is also rather small. It is not visible at two significant digits in recall and only amounts to 2% in relative recall (up to 45%): prefix verbs, even though conceptually prominent, are quantiatively a relatively small part of German derivational morphology. Thus, the substantial recall gap compared to DErivBase remains. At this point, we cannot distinguish between the two salient interpretations, namely (a) that it is due to the resource-based methodology of creating DErivCelex, and (b) that it is due to the DErivBase-friendly sampling bias in the gold standard. This would require the creation of a new, resource-independent gold standard.

5 Discussion and Conclusion

In this paper, we have considered the task of creating derivational lexicons, and have argued that existing resources crucially lack in cross-lingual comparability. We have presented an algorithm that extracts such lexicons from the German morphological layer of CELEX, a lexical database that is available for multiple languages, and have evaluated the result, DErivCelex, against the German DErivBase resource. We found that (a) DErivCelex misses the “long tail” of lemmas that DErivBase covers; (b) has an extremely high precision; (c) inherits a more restrictive definition of derivation from CELEX than DErivBase adopts. In our estimation, (a) is not a deal-breaker for applications unless they deal with very low-frequency lemmas: DErivCelex does provide nontrivial derivational information for over 16K lemmas. The most interesting and unexpected finding is (c). Its consequences for applications, such as psycholinguistic modeling of morphological priming (Padó et al. [11]), remain to be explored in future work. Another direction that we will follow is the creation and evaluation of corresponding derivational lexicons derived from the Dutch and English versions of CELEX.

Acknowledgments

We gratefully acknowledge funding of our research by the DFG, SFB 732 (project B9: Lapesa and Padó; project D10: Frassinelli and Padó).

References

[1] Harald Baayen, Richard Piepenbrock, and Léon Gulikers. CELEX2 (LDC96L14). *Philadelphia: Linguistic Data Consortium*, 1995.

[2] Ryan Cotterell and Hinrich Schütze. Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics*, 2017.

[3] Johannes Erben. *Einführung in die deutsche Wortbildungslehre*. Erich Schmidt, 1975.

[4] Gertrud Faaß and Kerstin Eckart. Sdewac – a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg, 2013.

[5] Wolfgang Fleischer. *Wortbildung der deutschen Gegenwartssprache*. VEB Bibliographisches Institut Leipzig, 1969.

[6] Nizar Habash and Bonnie Dorr. A categorial variation database for English. In *Proceedings of NAACL-HLT*, pages 17–23, Edmonton, AL, 2003.

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

[8] Steve T. Kempley and John Morton. The effects of priming with regularly and irregularly related words in auditory word recognition. *British Journal of Psychology*, pages 441–445, 1982.

[9] Susan Olsen. Delineating derivation and compounding. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, pages 26–49. Oxford University Press, 2014.

[10] Sebastian Padó, Jan Šnajder, and Britta D. Zeller. Derivational smoothing for syntactic distributional semantics. In *Proceedings of ACL*, pages 731–735, Sofia, Bulgaria, 2013.

[11] Sebastian Padó, Britta Zeller, and Jan Šnajder. Morphological priming in German: The word is not enough (or is it?). In *Proceedings of NetWords*, pages 42–45, Pisa, Italy, 2015.

[12] Ingo Plag. *Word-formation in English*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2003.

[13] Martin Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[14] Eva Smolka, Katrin H. Preller, and Carsten Eulitz. 'verstehen' ('understand') primes 'stehen' ('stand'): Morphological structure overrides semantic compositionality in the lexical representation of german complex verbs. *Journal of Memory and Language*, 72:16–36, 2014.

[15] Jan Šnajder. DErivBase.HR: a high-coverage derivational morphology resource for croatian. In *Proceedings of the LREC*, Reykjavík, Iceland, 2014.

[16] Zdeněk Žabokrtský, Magda Sevcikova, Milan Straka, Jonáš Vidra, and Adéla Limburská. Merging data resources for inflectional and derivational morphology in Czech. In *Proceedings of LREC*, pages 23–28, Portoroz, Slovenia, 2016.

[17] Britta Zeller, Jan Šnajder, and Sebastian Padó. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. *Proceedings of ACL*, pages 1201–1211, 2013.

Online Software Components for Accessing Derivational Networks

Jonáš Vidra, Zdeněk Žabokrtský

Institute of Formal and Applied Linguistics
Charles University, Czech Republic
E-mail: {vidra, zabokrtsky}@ufal.mff.cuni.cz

1 Introduction

This paper presents two studies on software tools developed for lexical derivational databases such as DeriNet [9]. DeriNet is a network of lexical derivations in the Czech language. Because of the size of the network, programs for searching and visualizing it are needed. In the current version 1.4 it contains 1 011 965 lexemes connected with 773 363 edges, constituting 238 602 derivational clusters (nests). Each lexeme is annotated with a unique identifier, the lemma it represents, its part-of-speech tag and in case it is derived, then also the identifier of the lexeme it is derived from (its derivational parent). Currently, the annotation in DeriNet only allows a single parent for each lexeme and thus the clusters are tree-shaped.

In section 2 we introduce a domain-specific query language implemented by a search engine named DeriSearch. Although it was originally developed for use with DeriNet, the tool is sufficiently general to be used with other similar databases. To show this portability, we have imported the Word Formation Latin resource [3] into the search engine and present examples from both Czech and Latin.

Section 3 contains information about our recent experiments with visualization methods for derivational trees. Some of the visualization methods are available in DeriSearch and other in DeriNet Viewer introduced in [9]. Both applications are accessible at https://ufal.mff.cuni.cz/derinet.

2 Querying derivational databases

A browser of derivational clusters called DeriNet Viewer introduced in [9] shows the derivational tree for any lemma specified by the user. However, such a limited search is obviously insufficient for revealing specific errors and edge cases in the data. This was our initial motivation for looking for a tool that would allow us to find derivational trees by specifying more detailed conditions, such as combinations of part-of-speech tags or constraints on the tree shape. For instance, the system should be able to find "A noun ending with *-tel* (prototypically 'doer' lexemes, such as *ředitel* ('director') or *učitel* ('teacher')) not derived from a verb."

We decided to design a query language with which it would be possible to express such queries, and a search engine that would process this language.

When designing a new formal language, a tradeoff between expressive power and simplicity has to be made. We are aware of a number of already existing tree query languages (and their user interfaces and search engine back-ends), especially from the treebanking world. They range from relatively limited languages tailored for a specific kind of data resource, such as Netgraph [4], through more complex query languages such as that of TIGERSearch [2], to highly elaborated, very expressive and general-purpose query languages such as PML-TQ [8]. However, our long-term experiences with PML-TQ show that most queries made by real users are very simple and only utilize a fraction of the capabilities of that language. Most of the language features are seldom used, but they still complicate the grammar and thus make learning and remembering the language more difficult.

Last but not least, there are also software engineering aspects. The PML-TQ search engine, which would undoubtedly cover all our query needs already now (when speaking about query language expressivity), is a large software project that requires major effort to support and maintain. However, at this moment we prefer rather a technologically lightweight solution that can be easily arranged towards the contemporary WWW technologies and flexibly changed when we gather more empirical evidence about real users' query needs.

A major design decision therefore was that simple queries should be simple to write, preferably without consulting the manual, even if it meant limiting the expressive power of the language. Ideally, it should be possible to search for a derivational cluster around a particular word by simply entering that word.

When designing the query language, we drew inspiration from the Corpus Query Language (CQL) [1], which is very popular in the linguistics community. The language had to be extended to support querying tree structures, as CQL was designed for searching only in linear sequences of tokens, not in trees.

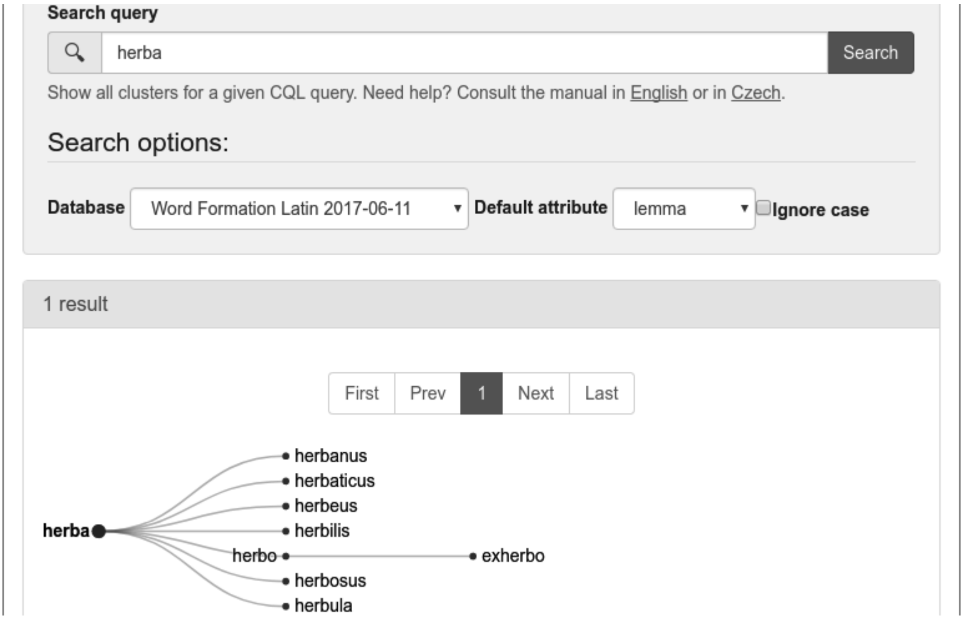
2.1 Specifying constraints on lexemes

The simplest query is a single lemma. If the user types `herba` ("grass") into the search box, the tree around the lemma *herba* is shown – see Figure 1.

Attributes other than lemmas can be queried by changing the desired options in the panel beneath the search box, which contains the following search settings:

Database selects the database to search. Currently, several versions of DeriNet [9] and Word Formation Latin [3] databases can be searched.

Default attribute allows users to match search terms against different attributes of the lexemes. By default, lemmas are matched, but this can be changed to e.g. part-of-speech tags. Then, the query `N1` would match all nouns in the first


Search query

herba

Show all clusters for a given CQL query. Need help? Consult the manual in English or in Czech.

Search options:

Database: Word Formation Latin 2017-06-11

Default attribute: lemma

□ Ignore case

1 result

Figure 1: DeriSearch user interface showing the derivational tree for the lemma *herba* from the Word Formation Latin data.

declension and `N` would match all nouns without declension information. Different corpora may provide different attributes in the list.

Ignore case makes the queries case insensitive.

By default, words typed directly into the search box are matched against lemmas by exact string comparison. If the user wants to search for substrings, alternatives or text patterns, they can do so by using JavaScript-compatible regular expressions. These are written by enclosing them in double quotes and allow expressing queries such as “All lemmas ending with *-os*”: `"os$"` The regular expressions match substrings, so if a whole-word match is desired, it has to be anchored on both ends: “Match lemmas ‘curo’ and ‘caro’”: `"^c[au]ro$"`.

Multiple lexeme attributes can be conditioned at the same time by enclosing conditions `attribute="value"` inside square brackets. By using this form, the user can search for example for “All adjectives beginning with *gra-*”: `[pos="A" lemma="^gra"]`. This also enables users to do an unconstrained query `[]` that matches all nodes, or to search using attributes not selected as the **Default attribute**, because the attributes are explicitly listed in the square brackets.

2.2 Specifying structural constraints

Parent-child relations in a tree are expressed by chaining multiple lexeme expressions together, e.g. “All nouns derived from verbs in the third declension”:

`[pos="V3"] [pos="N"]`. The expression on the left matches the parent and the one on the right matches any of its immediate children. Longer chains are also possible, for example "All lemmas ending in *-a* derived from a verb which is derived from a verb": `[pos="V"] [pos="V"] [lemma="a$"]`.

Multiple children are expressed by putting them into a comma-separated list in parentheses: `"us$" ("itas$", "us$" "itas$")`.

3 How to visualize derivational trees

3.1 Inspiration from dependency treebanking

Visualizing derivations is similar to drawing dependency trees, which have a much longer tradition in computational linguistics, but there are important differences too. First, nodes of dependency trees are ordered with two relations: a partial order expressing the parent-child relations and a linear order expressing textual precedence. In derivational trees siblings are unordered. Second, derivational trees have a larger node count and degree. For example, the trees containing the Czech lemmas *dát* ("to give") or *trhat* ("to rip") have hundreds of nodes. An out-degree (count of immediate children) of over 20 is common and some lexemes, such as the Latin *fero*, have a degree of over 100.

The most popular ways of dependency tree visualization are the following:

(A) as a sentence written on a single line, with links expressed as arcs above or below it. It has the advantage of presenting an easily readable sentence together with the parse tree, but this offers no benefit to derivational trees, which are not created by annotating sentences. Such trees are produced e.g. by the `brat` annotation tool [7]; see Figure 2 for an example,

(B) in a two-dimensional shape with the tree root at the top and its descendants on levels below, produced for example by `Tred` [5]. See Figure 4 for an example output from `Tred` [5].

(C) in 2D, this time with the root on the left, such as in Figure 3 produced by `Udapi` [6].

3.2 Requirements on derivational tree visualization

The visualization must express the derivational links between nodes and if possible, should show the overall structure of derivational clusters. It must also convey information about individual lexemes, especially the lemmas and parts-of-speech, but optionally also other information included in the database.

Our tool used the top-down approach (B) in its previous version, but as the DeriNet project progressed, it became unsuitable due to increasing tree size. With larger out-degree, the displayed trees grew in size horizontally, overflowing from the screen. On the other hand, depth of the trees doesn't increase as much. Currently, the deepest trees have 10 levels in Czech and 6 in Latin, and although they will grow in future versions, we don't expect large increases in depth.

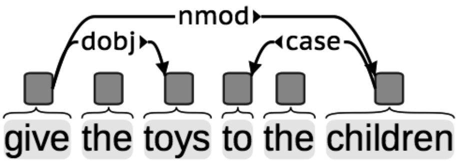


Figure 2: A tree from brat [7]

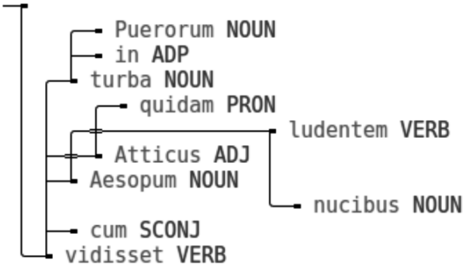


Figure 3: A tree from Udapi [6]

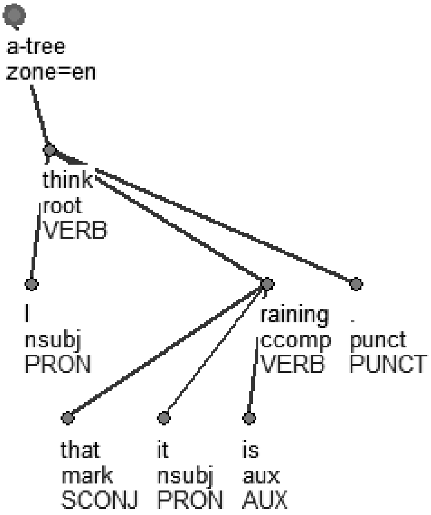


Figure 4: A tree from Tred [5]

Another important point to consider is that the nodes are annotated with textual data, which is written horizontally. Even if every datum is written on a new line, the nodes are rectangles that are longer along the horizontal axis. Nodes stacked side-by-side thus need a larger screen to display than nodes stacked on top of one another, even though most screens are also wide rather than tall.

3.3 Optimal usage of available space

From a user-interface viewpoint, it is best when all important information is shown on a single screen. When that's not possible and the user has to scroll the content, it is better if they scroll only vertically, because then they can use the mousewheel. Since the trees are quite shallow, but may have many nodes the same level, it is better to display the tree with the root on the left and its children to the right, as in the LEFT-TO-RIGHT method in Figure 6, rather than top-down, as in the TOP-DOWN method in Figure 5.

It is also possible to position the nodes in a non-linear fashion. We've experimented with a CIRCULAR style of display, where the tree root is in the center and its children are arranged in a circle around it. On lower levels, children of internal nodes fill fan-shaped areas, as illustrated in Figure 7. Since the circumference of a circle grows with its radius and the trees generally have more nodes on levels farther away from the root, this arrangement fills the given area more efficiently than linear trees, where nodes on the upper levels are pushed apart by their children. This method works best for small to medium-sized trees, because once the whole perimeter of a particular level becomes too crowded, the nodes either start overlapping or the diameter has to be increased to accommodate them.

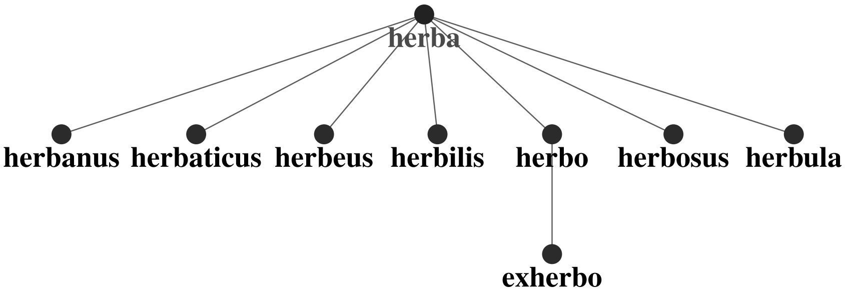


Figure 5: A derivational tree for the lemma *herba*, drawn TOP-DOWN.

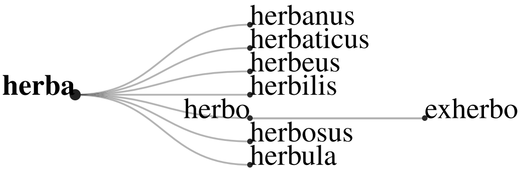


Figure 6: A derivational tree for the lemma *herba*, drawn LEFT-TO-RIGHT.

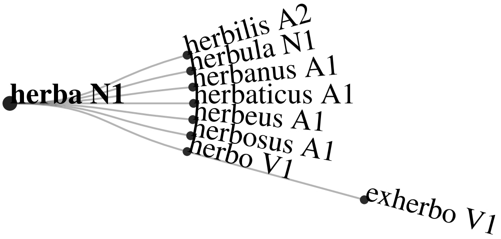


Figure 7: A derivational tree for the lemma *herba*, drawn in a CIRCULAR way.

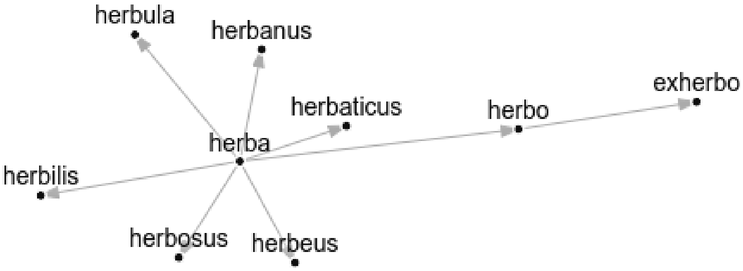


Figure 8: A derivational tree for the lemma *herba*, drawn with STRETCHy edges.

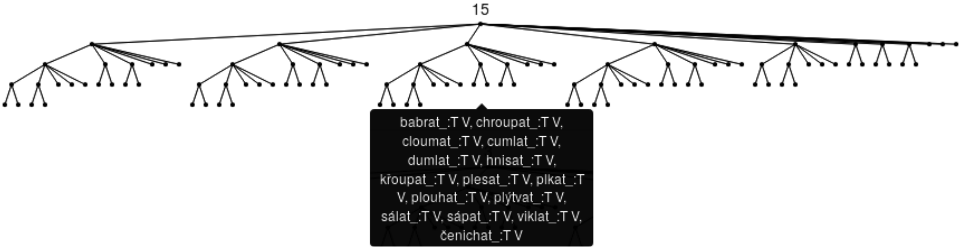
Another possibility is to display the trees dynamically. The previous examples have shown statically drawn trees that didn't change when the page was scrolled or zoomed. These are suitable for printing, but computer visualization allows us to change the shape and amount of displayed information in reaction to user actions.

We've created a visualization STRETCH, which positions the tree using a physics simulation with attractive and repellent forces. The edges between nodes act as springs and nodes repel one another as if by being charged with static electricity. Optimal placement of nodes in this layout is hard to determine automatically, but the user can reposition nodes using their mouse. An example can be seen in Figure 8.

3.4 Abstracting information from complex structures

Another method of fitting trees into limited space is to not show all parts of the tree in full. We can either hide some nodes or subtrees entirely or distill the important information into a more compact representation by e.g. hiding the labels with lemmas and other texts.

DeriNet Viewer does this by offering derivational tree statistics. This display only shows the tree shape, a count of trees with such shape in the database and the attributes of their roots. It omits attributes of the internal nodes and leaves. For an example, see Figure 9.



babrat_:T V, chroupat_:T V, cloumat_:T V, cumlat_:T V, dumlat_:T V, hnisat_:T V, křoupat_:T V, plesat_:T V, plkat_:T V, plouhat_:T V, plýtvat_:T V, sálat_:T V, sápat_:T V, viklat_:T V, čenichat_:T V

Figure 9: DeriNet Viewer showing a list of lemmas of root nodes whose trees have the same shape.

Instead of displaying an overview of the whole tree, we can also display a detailed view of just a small part of the tree. DeriSearch gives users the ability to collapse a subtree into a single node, hiding its contents, by clicking on a node. It also has a built-in “importance heuristic” that hides certain subtrees automatically, with a user-defined threshold. Currently, the heuristic is based on the frequency of the subtree’s shape and parts-of-speech in the data. More frequent (and therefore regular and expected) subtree types are being hidden more aggressively.

3.5 User evaluation

To evaluate the relative quality of different visualization methods, we’ve created four variants: TOP-DOWN, LEFT-TO-RIGHT, CIRCULAR and STRETCH and made them available in DeriSearch. For a visual overview, see Figure 10. We then asked 13 users to evaluate them while displaying the same data. We showed the users 16 different trees, each of them visualized in 4 different ways, and asked them to subjectively rate each visualization on a five point absolute scale. The answers were averaged and rescaled to a 0-100 range with higher values being better.

The results are summarized in Table 1 and Figure 11 shows the score of each method for different tree sizes. The plot shows that no single method is perfect for every tree.

The scores and subsequent discussion with the evaluators confirmed that a subjectively important factor is whether the whole derivational cluster fits on a single screen. The LEFT-TO-RIGHT and TOP-DOWN methods get good scores for as long

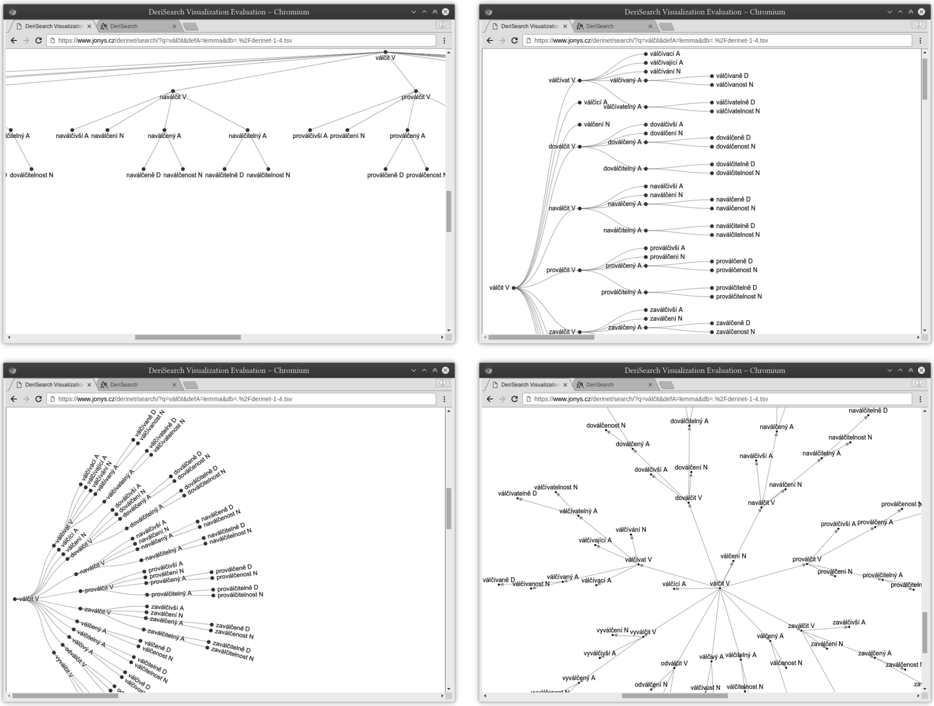


Figure 10: A comparison of the four user-evaluated visualization methods side by side; clockwise from top-left: TOP-DOWN, LEFT-TO-RIGHT, STRETCH and CIRCULAR. All visualizations are rendered with identical font size settings to a window with outer dimensions of 1024 × 768.

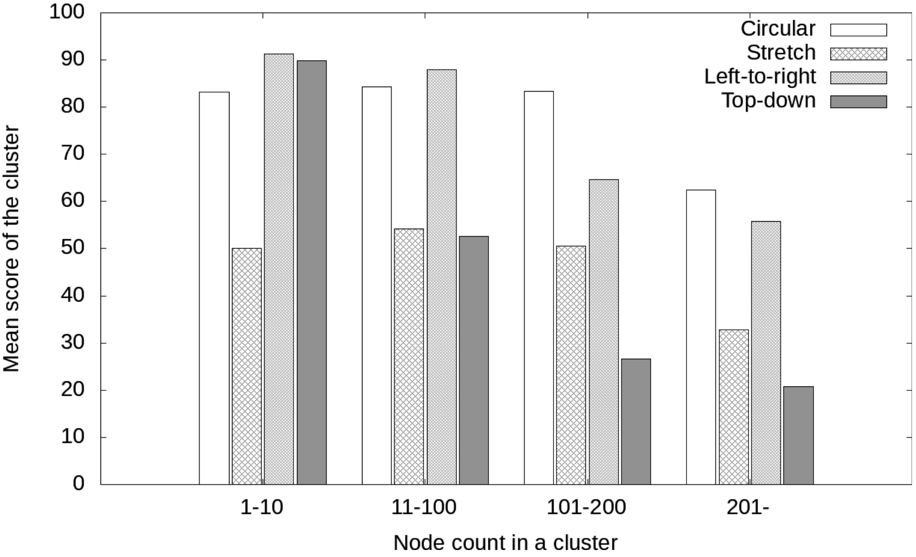


Figure 11: Results of the usability study. The x-axis shows clusters bucketed by their node count. Each bucket contains 4 clusters. The y-axis shows mean score achieved on that bucket on a scale of 0-100. Higher scores are better.

Table 1: Results of the usability study. Scores are on a scale of 0-100 with higher values being better. Each score is a mean of 13 annotations, SD is its standard deviation.

	CIRCULAR		STRETCH		LEFT-TO-RIGHT		TOP-DOWN	
size	score	SD	score	SD	score	SD	score	SD
5	78	24.7	53	39.3	96	9.4	80	23.2
5	83	12.3	47	37.6	91	16.3	93	11.3
6	85	12.9	45	38.2	87	19.9	93	11.3
6	85	12.9	52	34.5	89	16.7	91	16.3
20	88	13.1	45	36.8	90	16.9	79	21.8
40	85	16.7	58	41.7	89	16.7	47	31.0
44	75	27.0	50	27.0	86	16.5	48	36.0
70	89	12.9	62	36.1	85	16.7	37	22.6
118	88	13.1	50	35.4	68	22.6	22	17.5
160	83	19.5	43	37.1	62	25.0	25	18.5
179	85	16.7	52	36.1	62	25.0	29	27.9
195	76	25.9	55	41.0	65	24.0	28	28.6
225	79	14.4	47	37.6	62	25.0	31	28.5
453	77	19.8	37	37.7	58	26.8	16	24.6
784	54	27.9	18	26.4	52	32.8	16	22.2
1073	39	31.0	27	32.8	50	35.4	18	26.4

as they don't overflow the user's display. As soon as they do, the scores worsen. For the TOP-DOWN trees, this happens sooner. To illustrate this factor, Figure 11 shows the score of each method for different tree sizes. The point at which the trees overflow a single screen depends on the screen resolution, font rendering settings, zoom level and the exact tree shape, but for TOP-DOWN trees, it generally occurs at around 40 nodes, while the LEFT-TO-RIGHT method can fit around 65 nodes into a typical maximized browser window at a 1920 × 1080, 96 dots-per-inch screen.

Other important factor that influences scores is text legibility: The CIRCULAR method displays texts sideways and the orientation flips near the top and the bottom. In addition, texts can sometimes overlap in the LEFT-TO-RIGHT, CIRCULAR and STRETCH methods.

The STRETCH method is the only one that got lower scores on small trees than on mid-sized ones. Users reported that with small trees, it is not immediately obvious where the root of the tree is and the user has to explicitly follow the arrows. In bigger trees, a radial pattern emerges with the root near the center.

The STRETCH method also has larger standard deviations, suggesting that it is controversial, with some users liking it and others disliking it. Users have reported that the LEFT-TO-RIGHT and TOP-DOWN methods feel "more natural" and "less complex" than the STRETCH method, which is why they've received the best scores on the smallest trees, where their other disadvantages don't play a role.

4 Conclusions and future work

In this paper, we have described two parts of the software ecosystem developed to support the DeriNet project: a query language used for searching derivational trees and various visualization options available in the tools DeriNet Viewer and DeriSearch. These applications are publicly available as online services at `https://ufal.mff.cuni.cz/derinet`.

We plan to explore more visualization options in the future and to optimize the available ones. We feel that although clear, intuitive visualization of large trees is an important prerequisite for developing derivational resources, the ways of achieving it have not been sufficiently explored yet.

By importing foreign data into DeriSearch, we show that it is extensible and general enough to accommodate the needs of projects other than DeriNet. We hope that this search tool can and will be used by other projects working with derivational trees.

The applications will have to be updated in the near future to handle non-tree structures. DeriNet will gain support for lexical composition in version 2.0 and Word Formation Latin already has the parents of compositional words annotated. This information is currently ignored when importing Word Formation Latin into DeriSearch.

Acknowledgement

This work was supported by the Grant No. GA16-18177S of the Grant Agency of the Czech Republic. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

[1] O. Christ, B. M. Schulze, A. Hofmann, and E. König. *The IMS Corpus Workbench: Corpus Query Processor (CQP) User's Manual.* University of Stuttgart, Germany, August 1999.

[2] Wolfgang Lezius. TIGERSearch – Ein Suchwerkzeug für Baumbanken. In *Proceedings of the 6. Konferenz zur Verarbeitung naturlicher Sprache (6th Conference on Natural Language Processing, KONVENS 2002), Saarbrucken, Germany*, 2002.

[3] E. Litta, M. Passarotti, and C. Culy. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of CLiC-it 2016 & EVALITA 2016*, Napoli, Italy, December 2016.

[4] Jiří Mírovský. Netgraph - making searching in treebanks easy. In *IJCNLP 2008 Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 945–950, Hyderabad, India, 2008. Asian Federation of Natural Language Processing, International Institute of Information Technology.

[5] P. Pajas and J. Štěpánek. Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 673–680, Manchester, UK, 2008.

[6] M. Popel, Z. Žabokrtský, and M. Vojtek. Udapi: Universal API for universal dependencies. In *NoDaLiDa 2017 Workshop on Universal Dependencies*, pages 96–101, Göteborg, Sweden, 2017. Göteborgs universitet.

[7] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. BRAT: A Web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, pages 102–107, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[8] Jan Štěpánek and Petr Pajas. Querying diverse treebanks in a uniform way. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1828–1835, Valletta, Malta, 2010. European Language Resources Association.

[9] Z. Žabokrtský, M. Ševčíková, M. Straka, J. Vidra, and A. Limburská. Merging Data Resources for Inflectional and Derivational Morphology in Czech. In *Proceedings of LREC'16*, pages 1307–1314, Portoroz, Slovenia, 2016.

Questo volume è stato stampato
nel mese di settembre 2017
su materiali e con tecnologie ecocompatibili
presso la Litografia Solari
Peschiera Borromeo (MI)

EDUCatt
Ente per il Diritto allo Studio Universitario dell'Università Cattolica
Largo Gemelli 1, 20123 Milano - tel. 02.7234.22.35 - fax 02.80.53.215
e-mail: editoriale.dsu@educatt.it (produzione); librario.dsu@educatt.it (distribuzione)
web: www.educatt.it/libri

ISBN: 978-88-9335-225-3

Recent years have seen a growing interest in research aimed at building new linguistic resources and Natural Language Processing (NLP) tools for derivational morphology. The current increased interest in both the theoretical and applicative aspects of word formation is strictly connected to the large need for automatic semantic processing of linguistic data.

The Word Formation Latin project received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 658332-WFL. It ran from November 2015 to October 2017 and resulted in a word formation based lexicon and tool for Latin. The work was carried out at the CIRCSE Research Centre of Università Cattolica del Sacro Cuore in Milan.

The first Workshop on Resources and Tools for Derivational Morphology (DeriMo), whose contributions are collected in these proceedings, was organised to celebrate the end of the project and to consider the current status of research in the field.

13,50 euro

9 788893 352253