Visual Exploration of Latin Derivational Morphology

Chris Culy, Eleonora Litta, Marco Passarotti
Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milano, Italy
chrisculy@mac.com, eleonoramaria.litta@unicatt.it, marco.passarotti@unicatt.it

Abstract

The *Word Formation Latin* project is developing a new lexicon of Latin based on derivational morphology, a branch of linguistics that is increasingly gaining interest in the area of NLP thanks to its connection with semantics. This paper describes an easy to use web application to access this resource, using a combination of queries and interactive visualisations.

Introduction

Up until very recently, in the area of Natural Language Processing (NLP), derivational morphology has always been neglected when compared to inflectional morphology, which, on the other hand, plays a central role in fundamental NLP tasks like PoS tagging, syntactic parsing and word sense disambiguation. Yet enhancing textual data with derivational morphology tagging promises to provide strong outcomes. First, it organises the lexicon at higher level than words, by building word formation based sets of lexical items sharing a common derivational ancestor. Secondly, derivational morphology acts like a kind of interface between morphology and semantics, as core semantic properties are shared at different extent by words built by a common word formation process.

In the past years, some lexical resources for derivational morphology of modern languages have been made available. Particularly worthy of notice are *Word Manager*, a system for morphological dictionaries available for German, English and Italian (Domenig and ten Hacken 1992) and the experimental studies that Hippisley, Tariq, and Chang (2001) made on an object-relational database tailored for representing hierarchical relationships, like those of relational morphology, using an existing Word-formation Dictionary of Russian.

More recently, there has been the production of the lexical network for Czech *DeriNet* (Žabokrtský et al. 2016), the derivational lexicon for German *DErivBASE* (Zeller, Snajder, and Pado 2013), that for Italian *derIvaTario* (Talamo, Celata, and Bertinetto 2016), and the derivational and morpho-semantic resource for French *Démonette* (Hathout and Namer 2014). Furthermore, stemming is a technique largely used for detecting word formation processes (Goldsmith 2001), and language independent NLP tools were trained to extract derivation information from inflectional lexica (Baranes and Sagot 2014).

Most of the resources mentioned above either have no public interface, or provide only a command-line access to data. *DerIvaTario* does have a public interface with many options, but no use of co-occurrence restrictions or of visualizations is made available (http://derivatario.sns.it/). *DeriNet* also has a public interface, but it is limited to searching (formative components of) lemmas, with no possibility of querying the resource by word formation rules or any co-occurrence restrictions (http://ufal.mff.cuni.cz/derinet).

On the ancient languages front, although there are numerous and varied resources and NLP tools for Ancient Greek and (ranging from digital libraries, treebanks and computational lexica to PoS taggers and syntactic parsers), no lexical resource for derivational morphology are available yet, where words are connected by word formation processes.

Word Formation Latin (WFL) is a derivational morphology resource for Latin that can also work as an NLP tool, thanks to its strict connection with a morphological analyser of Latin. The contents of WFL are lemmas analysed into their formative components, and relationships between them established on the basis of word formation rules (WFRs). For example, lemmas *amo* ("to love") and *amator* ("lover") are connected with a relationship that describes a change from a verb to a noun through the addition of a suffix that in itself bears semantic information: in this case, the suffix *–a–tor* characterises agentive and instrumental nouns.

The WFL project has received funding from the EU Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Individual Fellowship. The project is currently work-in-progress at the CIRCSE Research Centre of the Università Cattolica del Sacro Cuore of Milan (Italy) and it is due to be completed by October 2017.

Once completed, the lexicon will be made freely available for download, so that source data can be used for different purposes, ranging from running advanced queries to connecting WFL with other linguistic resources and exploiting its contents for various NLP tasks. Equally, the code for the interface will be distributed as open source. Lexical data will be released under CC-BY-NC-SA 4.0 licence (https://creativecommons.org/licenses/by-nc-sa/4.0/). The code for the interface will be distributed under the terms of the GNU General Public Licence (https://www.gnu.org/licenses/gpl-3.0.en.html).

Dealing with an ancient language, the aim of the project is also to make the resource easily accessible to a public larger than the community of computational linguists. In particular, typical users of WFL are expected to be theoretical linguists (e.g. for - comparative - studies on morphological productivity) and literary scholars working on Latin material (e.g., for retrieving in texts all the occurrences of words sharing the same derivational process), who are not always familiar with databases, and query languages. Thus, our challenge has been to build an intuitive and user-friendly web application supporting different kinds of queries to be run on WFL, which features a positive balance between potential of data extraction and simplicity, dynamism and interactivity.

This paper describes the WFL web application, currently hosted at http://wfl.marginalia.it/. The web application, as well as related work, can be characterized in terms of (a) the data it uses, (b) the model of derivational morphology that it incorporates, and (c) the interface developed to access the resource. The paper is organized accordingly. The Data and Model section presents the lexical basis of WFL and describes the types of WFRs available in the lexicon. The Interface section details the design and functionality of the web application. Finally, the Conclusion points out places for further work.

Data and Model

The lexical basis used for building WFL is the one provided by the morphological analyser for Latin Lemlat (Passarotti 2004, http://www.lemlat3.eu), which processes input word forms by providing them with out-of-context lemmatization and morphological features (PoS, gender, mood, tense, person, number etc.). The lexical basis counts 40,014 entries and 43,432 lemmas (as more than one lemma can be included into the same lexical entry), and it was recently enlarged by the addition of 26,250 onomastic lemmas (Budassi and Passarotti 2016).

The basic component of the lexical look-up table used by Lemlat for processing word forms is the so-called LES ("LExical Segment"), which roughly corresponds to the sequence of characters that remains the same across the inflectional paradigm of a lemma (thus not necessarily representing the word stem). For instance, *puell* is the LES for the lemma *puella* ("girl"), as it is the sequence of characters that does not change in the different forms of the lemma *puella*: *puell-a, puell-ae, puell-am, puell-ae, puell-arum, puell-as, puell-is*.

Building upon Lemlat, WFL connects its lexical items by WFRs. In WFL, there are two main types of WFRs: (a) derivation and (b) compounding. Derivation rules are further organised into two subcategories: (a) affixal, in its turn split into prefixal and suffixal, and (b) conversion, a derivation process that changes the PoS of the input word without affixation.

The WFL database is built in two steps. First, WFRs are found. Then, they are applied to lexical data. Affixal WFRs are found both according to previous literature on Latin derivational morphology (Jenks 1911; Fruyt 2011; Oniga 1988; Oniga 2007) and in a semi-automatic manner. The latter is performed by extracting from the list of lemmas of Lemlat the most frequent sequences of characters occurring on the left (prefixes) and on the right (suffixes) side of lemmas. The PoS for WFRs input and output lemmas as well as their inflectional category are manually assigned. Phonetic change is managed both automatically (e.g. when looking for verbs including prefix *con-* "together", a query will keep in account assimilation by searching at the same time for lemmas starting with *con*, *com*, *col*, *cor*, *co*, hence connecting *laboro* "to labor" with *collaboro* "to labor with") and manually: in case of apophony, e.g. *teneo* "to hold" > *detineo* "to hold off", the connection is hard coded. Further affixal WFRs are found by comparison with data. So far, we have detected 172 affixal WFRs: 72 prefixal and 100 suffixal.

Compounding and conversion WFRs are manually listed by considering all the possible combinations of main PoS (verbs, nouns, adjectives), regardless of their actual instantiations in the lexical basis. For instance, there are four possible types of conversion WFRs involving verbs: V-To-N (*claudo* → *clausa*; "to close" → "cell"), V-To-A (*eligo* → *elegans*; "to pick out" → "accustomed to select, tasteful", example of conversion with apophony), N-To-V (*magister* → *magistro*; "master" → "to rule"), A-To-V (*celer* → *celero*; "quick" → "to quicken"). Each compounding and conversion WFR type is further specified by the inflectional category of both input and output. For instance, A1-To-V1 is the conversion WFR that derives first conjugation verbs (V1) from first class adjectives (A1).

Applying WFRs to lexical data requires that each morphologically derived lemma is assigned a WFR and is paired with its base lemma. All those lemmas that share a common (not derived) ancestor belong to the same "morphological family". For instance, lemmas *amator* ("lover"), *amor* ("love") and *amabilis* ("lovable") all belong to the morphological family whose ancestor is the lemma *amo* ("to love").

We recorded the rules in a table of an SQL relational database where each WFR is classified by type and it is assigned the required PoS, inflectional category and gender for its input and output. Together with the list of WFRs, the main tables of the database are the LES archive of Lemlat and the list of its lemmas (each assigned its PoS, inflectional category and, for nouns only, gender). Lemmas and WFRs are paired by using a number of *ad hoc* SQL queries providing the candidate lemmas for each WFR. So far, we have applied 151 WFRs, which build 5,044 morphological families and 20,356 input-output relations.

Evaluation is performed by calculating the precision rate (Van Rijsbergen 1979) of SQL queries, i.e. the percentage of the correct candidate input-output pairs that are automatically assigned to a WFR by a query. Precision for prefixal rules ranges between 0.95 and 0.8, as they imply fewer graphical mutations, while precision for suffixal rules can vary heavily, ranging from 0.75 to as little as 0.3. The recall of queries has to be calculated later in the project, when we will be able to verify how many derived lemmas are not automatically picked up.

For the website, the SQL database of WFL is effectively transformed via a Python module into a graph of the morphological families. In this graph, a node is a lemma, and an edge is the WFR used to derive the output lemma from the input one (or two, in the case of compounds), along with any affix used. The graph is represented as a collection of edges, and the set of morphological families is simply the set of weakly connected subgraphs, which is calculated by a series of breadth-first traversals of the whole graph. Examples are in Figures 4, 5, 6 and 7 below.

Interface

Design process

The website has been designed in an iterative collaborative process of conceptualization of the kinds of queries and results that a user would be interested in. This process selected four distinct perspectives to query the information included in the database. WFL can be browsed:

- by WFR – here the primary interest is the WFR itself. This view enables research questions on the behaviour of a specific WFR. For example, it is possible to view and download a list of all verbs that derive from a noun with a conversion derivation process (e.g. *radix* 'root' > *radicor* 'to grow roots');
- By affix – it acts similarly as above, but works more specifically on affixal behaviour. For example, this perspective enables to retrieve all masculine nouns featuring the suffix *–tor* and to verify how many of them correspond to a female equivalent ending in *–trix*;
- By PoS – the primary interest is the part of speech (PoS) of input and output lemmas. This view is useful for studies on macro-categories of morphological transformation, like nominalisation and verbalisation;
- By lemma – it focuses on both derived and non-derived lemmas. It supports studies on the productivity of one specific morphological family or a set of morphological families.

The results of these browsing options are of three types:

- lists of lemmas matching a query;
- derivational graphs. This type of graph represents the derivational chain (or cluster) for a specific lemma, which includes all the lemmas derived from the lemma selected, as well as all those the lemma is derived from;
- a summary of the application of a given WFR to different PoS and the resulting lemmas.

An important design aspect of the WFL web application is the goal of limiting queries that produce no results. Queries could produce no results if they search either for unattested WFRs (for example, no prefixal WFR alters the PoS of the lemma), or for WFRs not yet included in WFL. In addition, users who use the application as the database grows will see that growth as the number of options increases.

To achieve this goal, the database is analysed for which affixes and PoS are associated with which WFR(s) both as input and output, and other similar co-occurrence restrictions. These restrictions are dynamically reflected in the interface, limiting the user's query choices to those that will produce one or more results. We made this design decision following standard best practices in user interface design in order to meet the need of building a powerful, yet simple interface for querying WFL.

There are a couple important points to be made here. The first one is that by calculating the co-occurrence restrictions and representing them in the interface, we have essentially done the equivalent of hundreds of queries (for the combinations), the results of which can easily be seen at glance from the interface. For instance, the suffix *–il–* is available only for denominal adjectives (N-To-A), which means that it is not available for any other possible combinations of input/output PoS, which we can easily see from the menus in the interface. Thus, we have reduced all the database queries to a glance at the interface (a "visual query" in information visualization terms).

A second point is that that not all unproductive queries are prevented, only those involving co-occurrence restrictions. It is possible, via the "Explore by lemma" section, to search for any string as a lemma. Of course, if there is no result, the user can only be sure that the form searched for is not in the database. But this is part of the more general problem of unattested forms. For forms that scholars believe could have occurred but are accidentally missing from the record, an * is typically used in the literature. We are in fact exploring ways to incorporate this kind of information in the database and preliminary support is already in the tool. Needless to say, adding likely but unattested forms to the database is another tremendous effort and presents additional challenges.

One additional design aspect is pre-computation of much of the information. The SQL database is accessed only when the web application is started. At this point, the conversion to a graph is done, as are the computations for the derivational graphs, as well as the co-occurrences analyses mentioned above. Since the SQL database is only accessed at start-up, this allows the database to be developed separately from the web application.

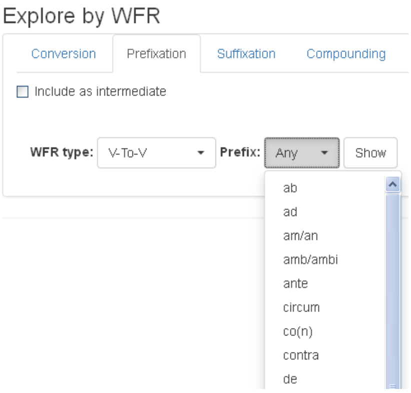
One final notable design aspect is the attempt to keep as much information in the client as possible in order to reduce the number of interactions with the server-side application and thus have better responsiveness. All of the co-occurrence information, as well as the basic lemmas information, is passed to the web client when it connects to the server so that the only interaction with the server-side application is to retrieve the morphological families and the lemmas associated with a given WFR. The site is thus a type of "single page web application".

Functionality

In the web application, the four perspectives on queries mentioned above are implemented as four different screens, accessed via a top-level menu.

For WFRs and affixes, the basic type (e.g. "Prefixation" for WFRs, or "Prefixes" for Affixes) is chosen via tab buttons, and for all perspectives the finer grained choices are specified via dropdown menus, which encode the co-occurrence restrictions. The difference between querying the database by WFRs and by affixes is reflected in the priority of dropdown menus. For WFRs, first a WFR type (or types) is chosen (e.g. V-to-V for deverbal verbs), and then any desired affixes. The choice of the WFR type updates the second dropdown menu to restrict the affixes to just the ones that occur with the selected WFR type. A similar interaction holds for affixes.

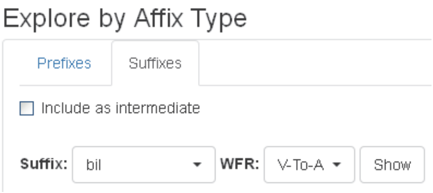
An example of exploration of WFL by WFR is shown in Figure 1, where we can see in the dropdown menu a selection of the prefixes available for the formative process that involves prefixation of verbs.



WFR type: V-To-V
Prefix: Any

Figure 1. Querying WFL by WFR

Figure 2 shows the browsing option enabling to query WFL by affixes. In particular, the query in Figure 2 searches for deverbal adjectives (V-To-A) formed with suffix *–bil*.



Suffix: bil
WFR: V-To-A

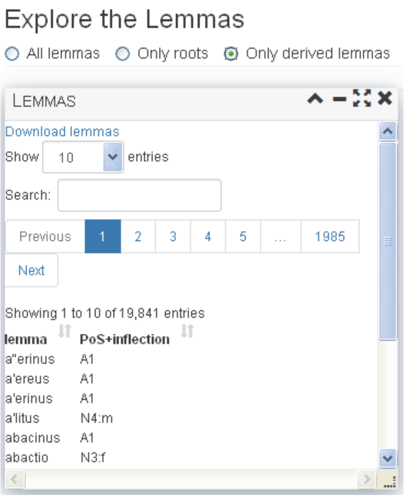
Figure 2. Querying WFL by affix

The PoS-based query option does not have the intermediate level of selection; rather it is all done with a series of dropdown menus, similar in concept to the ones for WFR and for affixes. For each possible item involved in a WFR (one or two base input lemmas - the latter for compounds - and the output), there is the choice of PoS, and then refinements of that PoS: these are inflectional categories for all PoS (declension for nouns, classes for adjectives and conjugation for verbs), as well as gender for nouns. Again, the options for the inflectional categories are limited to those appropriate for the PoS chosen.

Querying WFL by lemma is the simplest type: radio buttons allow for the selection of all lemmas, only roots of derivational clusters (not derived lemmas) or only derived lemmas. The list of lemmas with their PoS (and gender, for nouns) is shown in a paged, filterable list.

The three types of query results are visualized in distinct ways in separate floating windows, with interactions enabling to explore across the result types.

To provide an example, we can start from the list of morphologically derived lemmas available from the "by lemma" browsing option (Figure 3).



Explore the Lemmas

○ All lemmas ○ Only roots ● Only derived lemmas

LEMMAS

Download lemmas

Show 10 entries

Search:

Previous 1 2 3 4 5 … 1985

Next

Showing 1 to 10 of 19,841 entries

lemma	PoS+inflection
a"erinus	A1
a'ereus	A1
a'erinus	A1
a'litus	N4:m
abacinus	A1
abactio	N3:f

Figure 3. List of derived lemmas

Clicking on a lemma opens its derivational graph in a separate window as a "centrifugal" layout of the multitree graph (Furnas and Zacks 1994). Figure 4 shows the derivational graph for the verb *proclamo* ("to cry out").

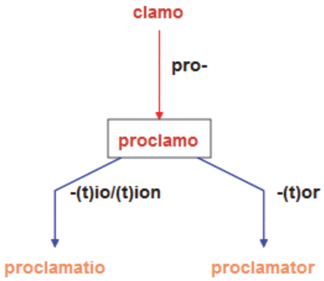


Figure 4. Derivational graph of proclamo

In the graph of Figure 4, nodes are filled with lemmas and edges are labelled with affixes or input-output PoS (the latter in the case of compounds and conversion WFRs). Lemmas are colour-coded by PoS, while edges are colour-coded for whether they connect a predecessor of the clicked lemma or a successor. The selected lemma is shown inside a box. Clicking on any lemma in the graph replaces the current derivational graph with the one for the clicked lemma.

Clicking on an edge label in the graph opens a new window which provides a visualization summarizing the application of the corresponding WFR by PoS. Figure 5 shows the visualization for the V-to-V WFR with prefix *pro–*. The window providing this visualization is opened by clicking on the edge connecting the node for *clamo* ("to call") with that for *proclamo* in the graph of Figure 5.

The visualization in Figure 5 is a left-rooted tree, with the name of the affix as the root (first level of the tree).

The second level of the tree reports all the combinations of the input and output PoS with their refinements (e.g. conjugation for verbs). The width of each branch indicates the relative frequency of that subset of applications of the WFR, while a tooltip (not shown in the Figure) gives the precise count.

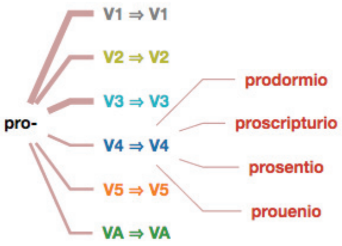


Figure 5. Summary of V-To-V pro–

The third level consists of the output lemmas for each input-output combination. The graph is collapsible so the user can focus on certain subsets only. In Figure 5, only the branch showing V4 (fourth conjugation verbs) is expanded. As the subsets change, the list of lemmas is updated to reflect just the subsets that are selected. Clicking on a lemma in these trees shows its derivational graph.

An additional feature of querying the database by WFRs and affixes is to search across the full derivational path of lemmas, thus providing results that go beyond the "outermost" WFR. By selecting this option (referred to as "include as intermediate") one can search not only for all the lemmas derived by a specific WFR but also for those that include one such lemma along their derivational path. For instance, with this option selected, among the results of a query that searches for deverbal adjectives formed with suffix *–bil* is not only the adjective *affabilis* ("that can be easily spoken", "courteous", derived from the verb *affor*, "to speak to"), but also the noun *affabilitas* ("courtesy") which has a deverbal adjective formed with suffix *–bil* along its derivational path namely *affabilis*.

Yet another feature is the ability to download any of the query results. The list of lemmas can be downloaded as a tab-delimited text file, while the derivation graphs and WFR trees can be downloaded as images.

WFL does not only include derivational word formation, but also compounds. Special provisions are made to accommodate them in the graph model. In considering derivational graphs, compounds pose an analytical issue, namely whether visualizing or not both the roots of compounds, thus resulting in a multitree graph rather than a simple tree. The website does not take a position in this question, but rather it leaves this decision up to the user, allowing for either alternative to be displayed. This feature is unique to the WFL web application, as none of the existing systems include compounds.

Consider the nominal compound *agricola* ("farmer"), formed from the noun *ager* ("field", "farm") and the verb *colo* (“to till”, “to cultivate”). The morphological family of the (non-derived) lemma *ager* is in Figure 6.

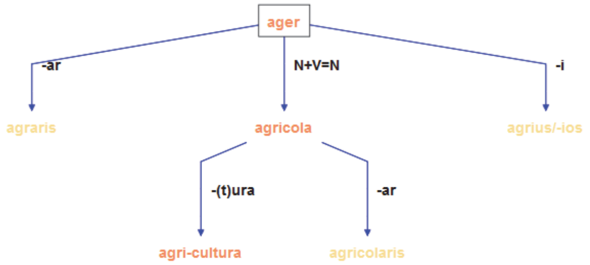


Figure 6. Simple morphological family of ager

In the graph of Figure 6, the node for *agricola* depends on that for *ager*, thus missing to represent that *agricola* is derived from two base lemmas (*ager* and *colo*). However, the WFL web application enables the user to change the visualization mode of derivational graphs by showing full compound derivations, as in Figure 7.

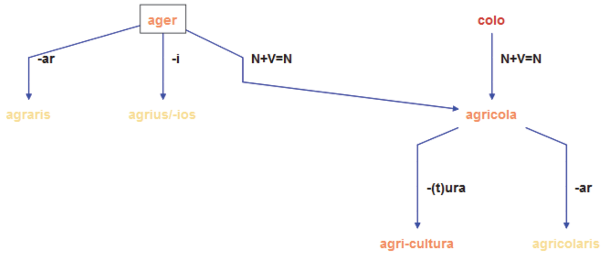


Figure 7. Full morphological family of ager

Conclusion

The current state of the web application provides an easy to use and innovative interface for experts and non-experts alike. However, there is room for improvement in each of the areas of data, model, and interface.

For the data, we are constantly expanding WFL, which is supposed to be completed by the end of 2017. For the model, we intend to make the graph model more explicit in the programming, and eventually to add more elaborated information along the lines of that in *Démonette*.

For the interface, we intend to allow querying of more than WFR at a time. The current Python module on the server side provides some summary statistics about the derivational families, but only on the server. We intend to incorporate that information, as well as other graph-based analysis measures, in the spirit of Hippisley, Tariq, and Chang (2001).

Finally, we would like to make both the resource and the interface available in the infrastructure CLARIN (https://www.clarin.eu/), thus making it possible for other similar resources to be accessed and queried via one common interface specifically suited for derivational morphology data.

References

Baranes, M., and Sagot, B. 2014. A Language Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 2793–2799. Reykjavik, Iceland: ELRA.

Budassi, M., and Passarotti, M. 2016. 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)*, 90–94. Berlin: The Association for Computational Linguistics.

Domenig, M., and ten Hacken, P. 1992. *Word Manager: A system for morphological dictionaries*. Hildesheim: Olms.

Fruyt, M. 2011. Word Formation in Classical Latin. In Clackson, J. ed., *A companion to the Latin language*, 157–175. Chichester: John Wiley & Sons.

Furnas, G.W., and Zacks, J. 1994. Multitrees: Enriching and Reusing Hierarchical Structure. In *Proceedings of the ACM CHI 94 Human Factors in Computing Systems Conference*, 330–336. Boston (Mass.): ACM Press.

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2): 153–198.

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

Hippisley, A.R., Tariq, M., and Chang, D. 2001. Hierarchical data and the derivational relationship between words. In *Proceedings of the Institute for Research into Cognitive Sciences Workshop on Linguistic Databases*, 125–133. Philadelphia (PA): Penn University.

Jenks, P.R. 1911. *A Manual of Latin Word Formation for Secondary Schools*. Lexington (Mass.): DC Heath & Company.

Oniga, R. 1988. *I composti nominali latini: una morfologia generativa*. Bologna: Pàtron.

Oniga, R. 2007. *Il latino: breve introduzione linguistica*. Milano: Franco Angeli.

Passarotti, M. 2004. Development and perspectives of the Latin morphological analyzer LEMLAT. *Linguistica Computazionale* XX-XXI: 397–414.

Talamo, L., Celata, C., and Bertinetto, P.M. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure* 9(1): 72–102.

Van Rijsbergen, C.J. 1979. *Information retrieval*. London: Butterworths, 2nd edition.

Žabokrtský, Z., Ševčíková, M., Straka, M., Vidra, J., and Limburská, A. 2016. Merging Data Resources for Inflectional and Derivational Morphology in Czech. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 1307–1314. Portorož, Slovenia: ELRA.

Zeller, B.D., Snajder, J., and Pado, Ś. 2013. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1201–1211. Sofia, Bulgaria: ACL.