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.

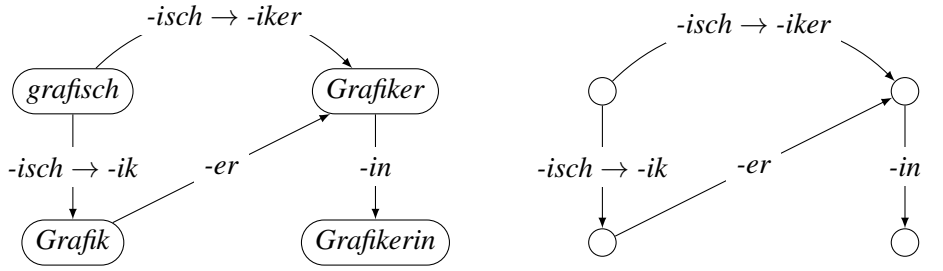

Figure 1: Illustration of a German derivational family (left) and its fingerprint (right)

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.

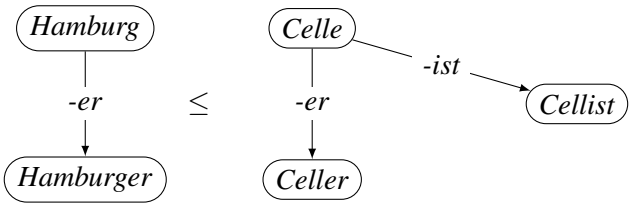

Figure 2: The left family is an induced subgraph of the right one: it is isomorphic to the right family sans *Cellist*.

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

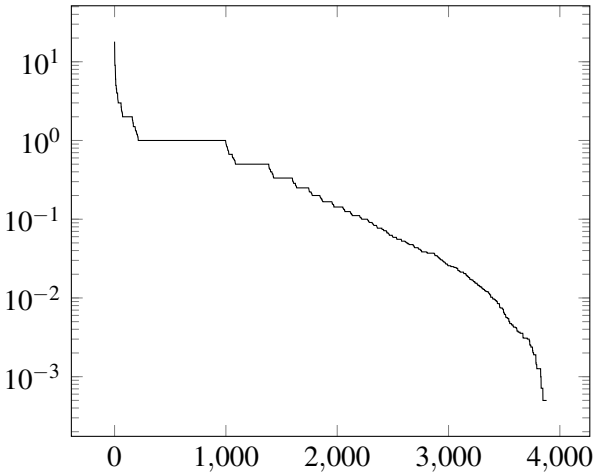

Figure 3: The ratio of the number of families for each error candidate, plotted by list index. Ratios are plotted on a logarithmic scale, so as to better illustrate differences in ratios which lie very close to zero.

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 Limbursk á. 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.