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 families (%)	Nonsingletons families (%)	Family size, mean (SD) 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.

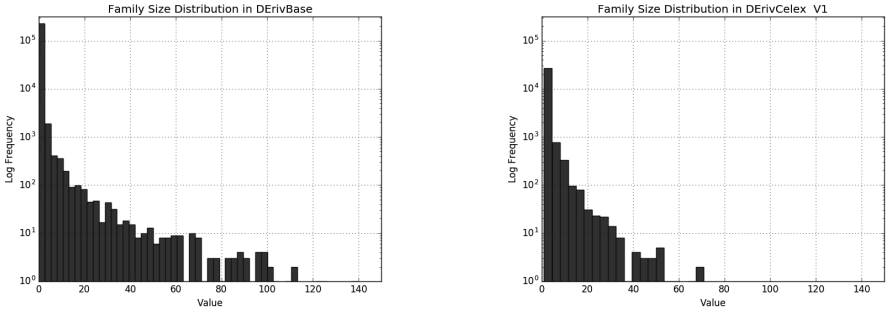

Figure 1: Family size distribution for DErivBase (left) and DErivCelex (right).

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]). Conse-

quently, *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 ⟵ ∅ ;                    /* stores a family ID for each lemma */
Headwords ⟵ ∅ ;                    /* 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.

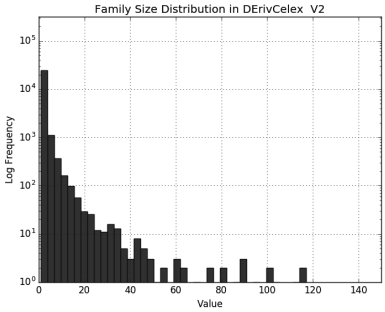

Figure 2: Family size for DErivCelex V2 derivational families

	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.