DErivBase: A derivational morphology resource for German

Britta D. Zeller*, Jan Šnajder†, Sebastian Padó*

*Institute of Computational Linguistics, Heidelberg University
†Faculty of Electrical Engineering and Computing, University of Zagreb

The 51st Annual Meeting of the Association for Computational Linguistics
August 6, 2013
A derivational resource – what is that?

- Derivation: a morphological process of word formation
- Derivational resource groups content words into derivational families: 
  \( to\ sleep_V - sleepy_A - sleepless_A - sleep_N \ldots \) 
  \( \Rightarrow \) Concept for a set of morphologically related words across POSes
- Resource provides information of morphological relatedness
  \( \leftrightarrow \) frequently implies semantic relatedness
- Degree of similarity depends on idiosyncrasies: 
  \( book_N - bookish_A \)
- Most previous research in computational morphology is about inflection normalisation, although derivational information is valuable
A derivational resource – what for?

Accounts for semantic relationships across POS boundaries:

- Extension of semantic roles resources [Green et al., 2004]: Extend lexical unit inventory of FrameNet [Baker et al., 1998]:
  \( \text{to ornament}_V \rightarrow \text{ornamentation}_N \)

- Improvement of text fluency:
  Reformulation in Natural Language Generation [Thadani and McKeown, 2011]:
  \( \text{Ferrero is mainly a candy producer}_N. \rightarrow \text{Ferrero produces}_V \text{ candies} \).

- Textual Entailment [Szpektor and Dagan, 2008]:
  Knowledge of derivations provides information for inference rules, e.g. noun modifiers which act as predicate:
  \( \text{the running}_A X \leftrightarrow X \text{ runs}_V \)
Related Work

- Manually constructed morphological analyzers: two-level approach, replacement rules in finite state technology [Koskenniemi, 1983], [Karttunen and Beesley, 2005]
  - No distinction between different morphological processes
  - We aim at more fine-grained control over precision and recall
- Derivational resource for English: CatVar [Habash and Dorr, 2003]
  - Builds on resources available only for English
Morphology for German

- Related resources and their shortcomings:
  - Celex [Baayen et al., 1996]: Limited coverage
  - IMSLex [Fitschen, 2004]: Not publicly available
  - Smor [Schmid et al., 2004], Morphix [Finkler and Neumann, 1988]: No distinction between inflection, compounding, and derivation

- DERivBase:
  - Publicly available
  - Contains morphologically related derivational families from a corpus
  - Covers over 280,000 German verbs, nouns, and adjectives
  - Rule-based approach → high precision
A rule-based approach

Motivation:

- German derivational processes are quite regular
- Small number of generic processes; can be freely combined
- Rules based on preexisting linguistic knowledge

Examples for derivational processes:

- **Suffix derivation:** *to edit*$_V$ $-$ *edition*$_N$
  
  “append ‘ion’ to the end of the stem”

- **Stem change:** *to sing*$_V$ $-$ *song*$_N$
  
  “replace ‘i’ by ‘o’ ”

- **Combinations:** *to perceive*$_V$ $-$ *perception*$_N$
  
  “alter stem ‘eive’ into ‘ept’, append ‘ion’ to the end of the stem”
Application of rule-based framework

German derivation rules
Application of rule-based framework

List of German verbs, nouns, and adjectives

German derivation rules
Application of rule-based framework

SdeWaC corpus → Lemma extraction → List of German verbs, nouns, and adjectives

German derivation rules
Application of rule-based framework

SdeWaC corpus

Lemma extraction

List of German verbs, nouns, and adjectives

German derivation rules
Application of rule-based framework

1. SdeWaC corpus
2. Lemma extraction
3. List of German verbs, nouns, and adjectives
4. German derivation rules
5. Derivation generation
Application of rule-based framework

SdeWaC corpus → Lemma extraction → List of German verbs, nouns, and adjectives → Filtering on lemma list → German derivation rules → Derivation generation → Derivation relations
Application of rule-based framework

- SdeWaC corpus
  - Lemma extraction
  - List of German verbs, nouns, and adjectives
    - Filtering on lemma list
    - German derivation rules
    - Derivation generation
    - Derivation relations
    - Derivational families
Definition of rule-based framework

- Modeling framework by Šnajder and Dalbelo Bašić, 2010
- Core of the framework:
  - Transformation function $t$: Maps a basis lemma into a derived lemma:
    - Input: to manage$_V$
    - Function: sfx(‘ment‘)
    - Output: management$_N$
  - Inflectional paradigms $\mathcal{P}_1$, $\mathcal{P}_2$: POS and gender information for basis/derived lemma
  - Derivational rules $d$: Derivation of derived lemma from basis lemma

\[
d = (t, \mathcal{P}_1, \mathcal{P}_2)
\]
Transformation functions

- Atomic string edit operations, e.g., \( sfx('ment') \)
- Can be composed into higher-order functions:

\[
d = ((sfx('ness') \circ try(rsfx('y', 'i'))), \mathcal{A}, \mathcal{N})
\]

\[
\rightarrow \text{kind}_A - \text{kindness}_N
\]

\[
\rightarrow \text{happy}_A - \text{happiness}_N
\]

- Rule induction: Derivation rules in traditional grammar books
  - Total implemented rules: 158
  - Amount of work: \( \sim 22 \) person-hours
Induction of derivational families

- Input: Set $\mathcal{L}$ of lemma-paradigm pairs $l-p$ from lemmatised, POS-tagged SdeWaC with gender information
  
  [Schmid, 1994, Faaß et al., 2010, Bohnet, 2010]:
  
  to respect-$V$

- Generate possible derivations with derivational rules $d$:
  
  respect-$N$, to disrespect-$V$, respected-$A$

- Avoid overgeneration: Remove derivations which occur less than 3 times in $\mathcal{L}$:
  
  * respectation-$N$

- Building the derivational family:
  
  Transitive closure of all pairs connected by derivation relations
Evaluation setting

- Induction of derivational families: **clustering problem**
- Similar to semantic class induction [im Walde and Brew, 2002] or coreference resolution [Cardie and Wagstaff, 1999]
  - Several evaluation techniques proposed
  - Our choice: Evaluation of Precision and Recall for **pairs of lemmas**
Evaluation sampling

- Skewed class distribution: Almost all pairs in \( L \) derivationally unrelated → Random sampling of pairs is problematic
- Preselection through String Similarity clustering based on Levenshtein distance ↔ Baseline
- Assumption: Preselection contains all true positive lemma pairs (all lemmas of derivational families): 
  cut\(_N\), to cut\(_V\), cutting\(_A\), cutlery\(_N\), cuttlefish\(_N\), cute\(_A\) ...
- Sampling: Draw a pair of lemmas from the same cluster, and compute Precision and Recall
  Total: 2,000 pairs

N.B.: Due to methodological caution, we carried out a more complex sampling; details in the paper
Sample annotation

- Binary annotation for each lemma pair: derivationally related or not?
  - Positive annotations: semantically and/or morphologically related
  - Negative annotations: no morphological relation, lemmatization errors, compound words
- Inter-Annotator Agreement:
  - Agreement: 0.85
  - Cohen’s $\kappa$: 0.79
## Results I

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>DErivBase</td>
<td>0.83</td>
<td>0.71</td>
</tr>
<tr>
<td>Stemming</td>
<td>0.66</td>
<td>0.07</td>
</tr>
<tr>
<td>String distance</td>
<td>0.36</td>
<td>0.20</td>
</tr>
</tbody>
</table>

- DErivBase achieves good precision and substantial recall
- Stemming leads to overclustering $\rightarrow$ low recall
- String similarity achieves more balanced but still poor results
Results II

- Manual analysis: Reliability of the derivational rules
- Three groups of rules:
  - L3 – very reliable
  - L2 – generally reliable
  - L1 – less reliable

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>DErivBase-L123</td>
<td>0.83</td>
<td>0.71</td>
</tr>
<tr>
<td>DErivBase-L23</td>
<td>0.88</td>
<td>0.61</td>
</tr>
<tr>
<td>DErivBase-L3</td>
<td>0.93</td>
<td>0.35</td>
</tr>
</tbody>
</table>
Conclusions

- Derivational resources provide **knowledge across POSes** which is helpful for various NLP tasks
- DErivBase is the first broad-coverage **German derivational resource**, and publicly available
- Combination of rule-based framework and corpus evidence allows for **high accuracy** and **solid coverage**
Thank you for your attention.

**Download** DErivBase from:
http://www.cl.uni-heidelberg.de/~zeller/res/derivbase/

Don’t miss our talk **today at 17:05 in Hall 7:**
Application of **DErivBase for smoothing distributional semantics**


Noun phrase coreference as clustering.

Morphemes as necessary concept for structures discovery from untagged corpora.
In Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, pages 295–298, Sydney, Australia.

Design and application of a gold standard for morphological analysis: SMOR in validation.
Morphix - a fast realization of a classification-based approach to morphology.

Ein computerlinguistisches Lexikon als komplexes System.

Inducing frame semantic verb classes from wordnet and Idoce.

A categorial variation database for English.
Unsupervised learning of morphology.

Inducing german semantic verb classes from purely syntactic
subcategorisation information.

Twenty-five years of finite-state morphology.
*In Arppe, A., Carlson, L., Lindén, K., Piitulainen, J., Suominen, M.,
Vainio, M., Westerlund, H., and Yli-Jyr, A., editors, Inquiries into Words,


A computational model of Croatian derivational morphology.
In Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Balkan Languages, pages 109–118, Dubrovnik, Croatia.

Learning Entailment Rules for Unary Templates.
In Proceedings of the 22nd International Conference on Computational Linguistics, pages 849–856, Manchester, UK.

Towards strict sentence intersection: Decoding and evaluation strategies.
Addendum

Statistics of the implemented rules

- 79 noun derivations, 33 verb derivations, 46 adjective derivations
- 6 zero derivations, 106 suffixations, 35 prefixations, 9 stem changes, 2 circumfixations
Addendum

Statistics of the performance of DErivBase

- Total coverage: 280,336 lemmas
- Grouped into 239,680 derivational families:
  - 17,799 non-singletons covering 58,455 lemmas
  - Many singletons are compound nouns
- Biggest 100% precision family: 40 lemmas