

University of Stuttgart
Institut für Maschinelle Sprachverarbeitung

Derivational morphology: Data-driven and rule-based approaches

Sebastian Padó

Collaborators

My work on derivation was developed with three primary collaborators...

Britta Zeller



Jan Šnajder



Gabriella Lapesa

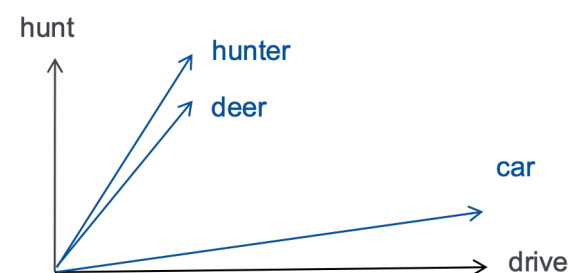


....plus numerous others: Marios Andreou, Aurélie Herbelot, Lea Kawaletz, Max Kisselew, Alexis Palmer, Sean Papay, Ingo Plag, Tillmann Pross, Laura Rimell, Antje Rossdeutscher.

All remaining errors are mine.

Derivational Morphology – What's in it for a computational semanticist?

- Characterizing word meaning is a never-ending task: even if we take a corpus-based approach



- Productivity
- Rarity (Zipfian distribution of words)
- But: morphological relatedness (via derivation) should imply (some degree of) semantic relatedness
 - Readers/listeners understand derivational neologisms:
Implies **(some) regularity of semantic interpretation**
- Vision: Better understanding of morphology-semantics interface can lead to better models of word meaning

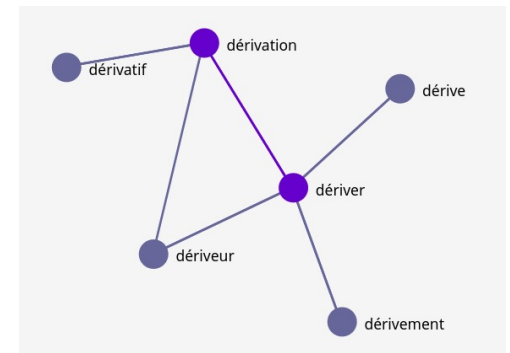
I set down the scrap of dolls dress, a **bedragglement** of loose lace hem.

Derivational Morphology – The Challenges

- Derivation is well known to be *semi regular*
 - Not every base can be combined with every pattern
 - The meaning of the derived word is only semi-predictable
- Insufficient to combine lexicon with a set of patterns: *we need a derivational lexicon* of attested forms
 - Minimally: set of *derivational families*
- What computational approach to use?
 - **Rule-based** vs. **data-driven** approach

die.v > *dier.n

bottle.n > bottle.v
card.n > card.v



Compact, interpretable, often too general

Learnable from data, flexible, often too specific

Derivational morphology resources can profit from both approaches!

Plan of the Talk

Part 1: Building Derivational Lexicons (DErivBase as an example)

Recurring motive: **Rules** + **Data**

- Induce Derivational Families
- Improve Consistency of Resource
- Determine Transparency

Part 2: Analysis of Derivation

- Effect of Derivation on Information Content
- Effect of Derivation on Valence

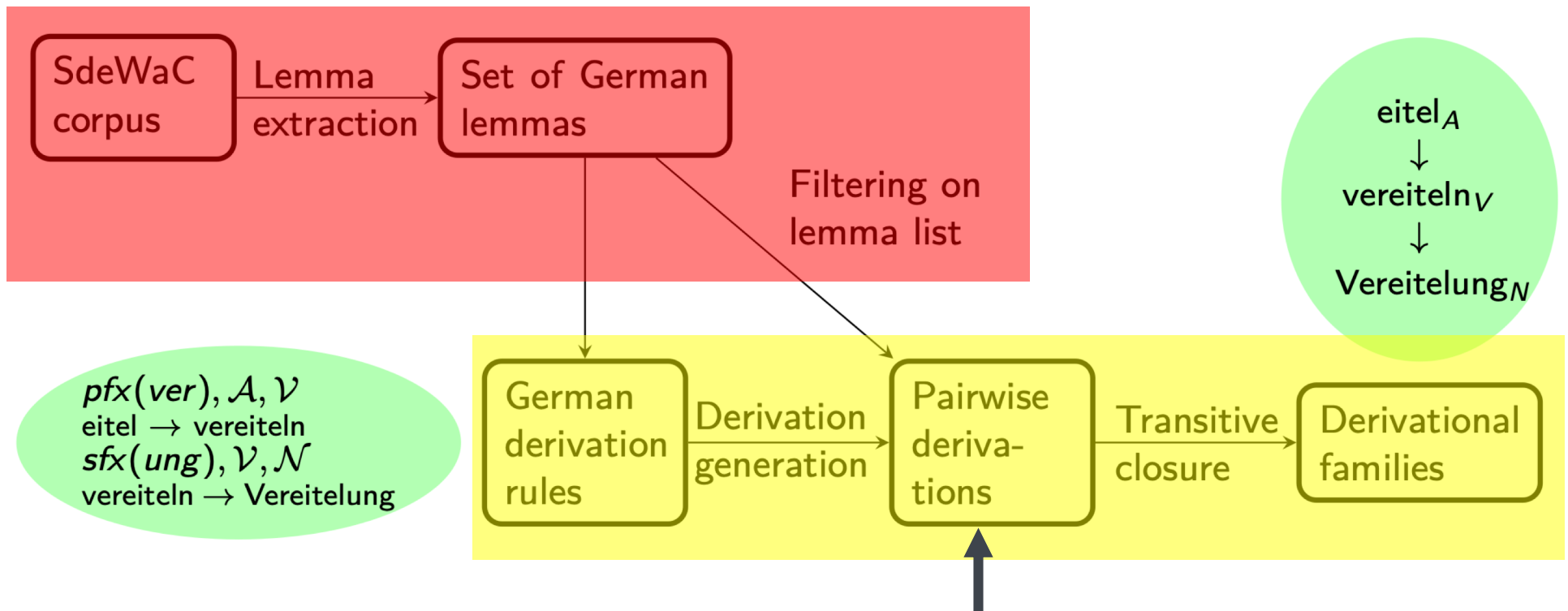
Part 1: Building Derivational Lexicons

Study 1: Derivational Lexicon Induction

[Zeller, Šnajder, Padó ACL 2013]

- **Goal:** Collect set of *derivational families*
 - **Approach:** Manually define derivation patterns as string transformations
 - Transformations map base lemma string onto derived lemma
(Šnajder & Dalbelo Bašić 2010)
- $$d = ((sfx('ness') \circ try(rsfx('y', 'i'))), \mathcal{A}, \mathcal{N})$$
- Inventory: suffix, infix, prefix addition/deletion/replacement; optionality
 - Source: German “school grammars”
 - Result: 158 rules (1 person week of work)
 - NB. No constraints on production – overgenerate!
- **Corpus:** SDeWaC, German web corpus (Faaß et al. 2010, ~900M words)

Lexicon Induction



Frequency threshold: exclude hapax legomena (*respectation.N)

Lexicon Induction: Result

- Result: DErivBase 1.4, groups 280k lemmas into 240k families
 - Non-singleton families: 18k families with 59k lemmas
- Evaluation: are word pairs morphologically related? (Manual annotation)

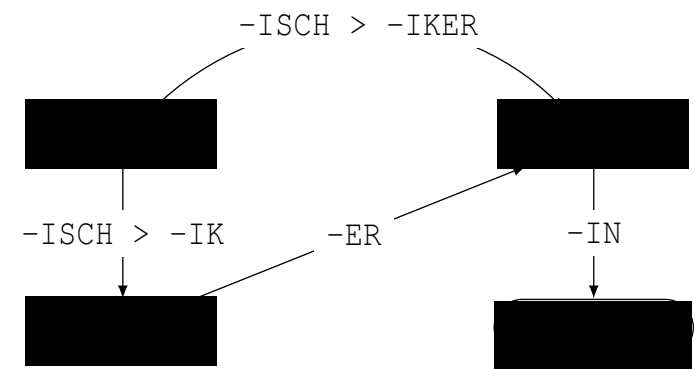
Method	Precision	Recall
DErivBase	0.83	0.71
Stemming	0.66	0.07
String distance	0.36	0.20

- **Precision errors:** Too general rules, lemmatization errors, NEs
 - Can be further increased by focusing on reliable rules – at cost of recall
- **Recall errors:** Frequency threshold, idiosyncratic cases (e.g. diachronic)
 - Hard to improve *ad hoc*

Study 2: Improving Recall

[Papay, Lapesa, Padó, DeriMo 2017]

- **Goal:** Further improve recall
- **Approach:** Derivational rules are *correlated*
 - If ISCH-IK + IK-IKER exists, then ISCH-IKER should exist as well
 - **Rule-based approach**
- Define fingerprint as *structure of delexicalized graph*
 - Captures relevant structure of derivational family:
 - 20k non-singleton families in DErivBase: 4539 fingerprints
- Hypothesis: If two families share *almost* the same fingerprint, (at least) one of them is incorrect



Mining fingerprint differences

- We focus on pairs of fingerprints that differ in **exactly one edge**
- Q: Which fingerprint is likely to be the correct one?
- A: The one that is instantiated by **more derivational families**
 - Larger fingerprint more frequent \rightarrow missing edge is recall error (*false negative*)
 - Smaller fingerprint more frequent \rightarrow extra edge is precision error (*false positive*)
- Experimental setup: We select 250 pairs of fingerprints each with the largest and smallest $f(\text{fingerprint 1})/f(\text{fingerprint 2})$
 - Manually annotate and evaluate

Evaluation

	FN	FP	OK	Other Errors
percentage in top 250	78.8	1.2	14.4	5.6
percentage in bottom 250	8.0	4.4	78.8	8.8

- Top 250 fingerprint pairs (likely false negatives): **Indeed FNs**

Ehrenbürger / Ehrenbürgerin (honorary citizen m./f.)
Einzeltäter / --- (lone offender m./ ---)

- Method can fill gaps!

- Bottom 250 fingerprint pairs (likely false positives): **Not false positives, but OK**

stöpseln / einstöpseln (plug / plug in)
→ errechnen / --- (calculate / ---)

- Rare but valid cases of derivation (e.g., specific (morpho)-semantic classes)

- **Take-home:** Missing and extra edges are not alike

- Top pairs serve to increase recall; bottom pairs not as useful

Study 3: Transparency

[Zeller, Šnajder, Padó, COLING 2014]

- DErivBase as described so far only captures morphological relatedness
 - Transparency relevant for many studies!
- **Goal:** include information about *semantic relatedness*
- **Approach:** Classify if lemma pairs of family are *semantically related*
 - Binary classification based on **word usage** and **rules**
 - **Hypothesis 1:** High distributional similarity → More likely to be related
 - **Hypothesis 2:** Derivation patterns differ a priori in their transparency
 - Diminutive (more transparent) `dog.n > doggie.n`
 - vs. conversion (less transparent) `card.n > card.v`
- **Method:** Supervised learning based on hypothesis-driven features

Transparency: Results

Method	Precision	Recall	F ₁
Majority baseline (sem. related)	72.6	100	84.1
Classifier, <i>distributional group</i>	80.5	96.6	87.8
Classifier, <i>rule-based group</i>	82.7	93.1	87.6
Classifier, <i>hybrid group</i>	80.4	95.3	87.2
Classifier, all features	86.2	93.9	89.9

- 73% of lemma pairs are semantically related: **majority baseline F=84.1**
- Can improve precision of class “related” to **86%**, at **cost of 6% recall**
 - Feature groups are complementary
 - Information available in DERivBase 2.0: Removal of intransparent edges
 - 18k families → 20k families

Part I: Summary, Assessment

- DErivBase: Derivational resource for German
 - Aims at high precision & high recall; distinguishes morphological and semantic relatedness
 - Lightweight: uses only grammar books and corpus
 - Methodology (somewhat) transferable to other languages (DErivBase.HR, .RU)
- Combination of data-driven and rule-based processes
 - Rules overgenerate – so analyses filtered against corpus data
 - Consistency criterion (fingerprints) can improve in particular recall
 - In principle applicable to other resources (probably requires typed edges!)
- Not manually validated: Contains errors Biss/bisschen (bite/a bit)

Part II: Analyses

- Most work on meaning effects of derivational morphology either
 - takes a psycholinguistic approach transparency
 - Takes a formal semantics approach +in → +FEMALE
- Comparatively little on *usage-based characteristics* of derivation
 - Next: two studies on this topic

Application 1: Information Content and Derivation

[Padó, Palmer, Kisselew, Šnajder, IWCS WS on Distributional Semantics 2015]

- **Question:** Does derivation systematically change *information content*?
 - Laca (2001): “a derived lexeme presupposes the lexeme it is derived from”
 - Hypothesis: Derived words are *more specific* / have *higher information content*
 - Specific meaning → restricted contexts → low entropy (Santus et al. 2014)

$$\text{seller}(x) = \exists e. \text{agent}(e,x) \wedge \text{sell}(e)$$

- **Method:**

- Compute SDeWAC distributional representations (“count vectors”)
- Measure entropy of count vectors
- Six DERivBase patterns (80 pairs each)
 - Two prefix, four suffix

	cry	smile	people
happy	2	20	4
unhappy	4	0	4

FEMALE	+in	NEGATIVE	un+
DIMINUTIVE	+chen	DIRECTED	an+
OPPOSITION	anti+	TRAVERSE	durch+

Results

Pattern	Base	Derived	Sample word pair	English translation	Entropy
<i>un-</i>	adj	adj	<i>sagbar</i> → <i>unsagbar</i>	<i>sayable</i> → <i>unspeakable</i>	60/20
<i>anti-</i>	adj	adj	<i>religiös</i> → <i>antireligiös</i>	<i>religious</i> → <i>antireligious</i>	78/2
<i>-in</i>	noun	noun	<i>Bäcker</i> → <i>Bäckerin</i>	<i>baker</i> → <i>female baker</i>	76/4
<i>-chen</i>	noun	noun	<i>Schiff</i> → <i>Schiffchen</i>	<i>ship</i> → <i>small ship</i>	74/6
<i>an-</i>	verb	verb	<i>backen</i> → <i>anbacken</i>	<i>to bake</i> → <i>to stick, burn</i>	71/9
<i>durch-</i>	verb	verb	<i>atmen</i> → <i>durchatmen</i>	<i>to breathe</i> → <i>to breathe deeply</i>	76/4

- All patterns: most derived words are more specific (90%)
 - **Most:** *anti-*: derived words properly ‘derivative’ political / antipolitical
 - **Least:** *un-*: derived words take on ‘life of their own’ erhört / unerhört
(heard / outrageous)
- Exceptions with lower specificity: *higher frequency* (Hay 2001)
 - 15/20 exceptions for *un-*, both exceptions for *anti-*
 - Makes cognitive sense: need either generalization from base, or exposure
- **Transparency, relative frequency, and information content correlate**

Application 2: Effects of Derivation on Valence

[Lapesa, Padó, Pross, Rosseitscher, IWCS 2017]

- **Question:** Does derivation affect specific *semantic aspects*?

- *Valence*: positive vs. negative associations

++ pet street sickness --

- **Method:**

- Re-use six derivation patterns
- Obtain 5-point valence scores from German Affective Norms (Schulte im Wald & Köper 2014)
- Regression analysis:

$\Delta \text{valence} \sim \text{pattern} + \text{concreteness} + \text{pattern} * \text{concreteness}$

$$0.5 = a_{\text{IN}} + a_{\text{conc}} * 4 + a_{\text{IN:conc}} * 4$$

Bäcker: val = 3.5, concr = 4

Bäckerin: val = 4, concr = 4

Results

- **Derived words have higher valence**
 - Mean valence 2.47 vs. 2.74
 - Not a frequency effect (present as covariates)
- Main **negative effect of Concreteness...**
- ...but **interactions with patterns anti+, +chen**
 - Valence *increases more* for concrete words
- Main **positive effect of +in (female)**
 - “Appreciative contexts”: implicit gender bias..
- Lots of potential for further investigations!

Predictor	Shift
Intercept	.27 ***
Concreteness	-.03 *
ANTI-: Concreteness	.05 **
-CHEN: Concreteness	.06 *
-IN	.06 *

Abstract: Deckmantel (smoke screen) /
Deckmäntel**chen** +0.15

Concrete: Bauernhaus (farm house) /
Bauernhäus**chen** +0.65

Eine gefragte Dekorateurin gibt Tips
a popular decorator (f.) gives tips

Conclusion

- Building derivational lexicons is a great research
- **During construction:** computational research questions on rule-based vs. data-driven approaches
 - Complementarity accommodates semi-regular nature of derivation
- **The result** can use it as basis for further research
 - Linguistics: Usage-based characterization of derivation
 - Conversion (Kisselew et al. 2016), Contrastive analysis of competing derivation patterns (Varvara et al. 2021), etc.
 - NLP: Derivational smoothing (Šnajder et al. 2013, Luang et al. 2014), Paraphrasing (Cotterell & Schütze 2018), etc.

Outlook: Two Avenues for Research

- **Avenue 1:** Can we carry out this type of analysis cross-linguistically?
 - Obviously, morphological systems differ
 - Is there a sense in which we can set up “parallel” derivational lexicons?
Can we align them in a meaningful way?
 - Vision: account, e.g., for cross-lingual differences in morphological priming
- **Avenue 2:** Consequences of developments in word embeddings
 - Milestone 1: FastText (Bojanowski et al. 2017), embeddings at subword level
 - Milestone 2: Transformers (Vaswani et al. 2017), “contextualized” embeddings with instance-level disambiguation
 - Both important in derivation, but not investigated in depth (as far as I know)

Thank you!

- Time for Questions!

- Acknowledgment to Deutsche Forschungsgemeinschaft for funding my research (project B9 / SFB 732)



Literature

- DERivBase: Inducing and Evaluating a Derivational Morphology Resource for German. Proceedings of ACL, pp. 1201-1211, 2013. B. Zeller, J. Šnajder and S. Padó.
- Towards Semantic Validation of a Derivational Lexicon. Proceedings of COLING, pp. 1728-1739, 2014. B. Zeller, S. Padó and J. Šnajder.
- Predictability of Distributional Semantics in Derivational Word Formation. Proceedings of COLING, pp. 1285-1296, 2016. S. Padó, A. Herbelot, M. Kisselew and J. Šnajder.
- Evaluating and Improving a Derivational Lexicon with Graph-theoretical Methods. Proceedings of DeriMo, 2017. S. Papay, G. Lapesa and S. Padó.
- Are doggies cuter than dogs? Emotional valence and concreteness in German derivational morphology. Proceedings of IWCS, 2017. G. Lapesa, S. Padó, T. Pross and A. Rossdeutscher.

Abstract

A central characteristic of derivational morphology is its (semi)regularity, i.e., the existence of regular patterns which are however subject to exceptions, gaps and subregularities. In the first part of my talk, I will argue that the creation of derivational resources can profit from the combination of theory-driven and data-driven methods, and will present evidence for this claim from the construction of DERivBase, a derivational dictionary for German, which combines hand written rules, distributional data, and graph theoretic methods [1, 2, 3].

In the second part, I will move to the exploitation of such resources and discuss the tension between how the semantic effects of derivation are captured on the theoretical side (transparency, specificity) and how they are captured on the distributional side [4, 5, 6].”