Combining Data-Intense and Compute-Intense Methods for Fine-Grained Morphological Analyses

Petra Steiner
Friedrich Schiller University Jena
Jena, Germany

September 19, 2019

Outline

Characteristics of German Word-Formation I

- language with highly productive and complex processes of word formation
- most common: compounding and derivation
- long orthographical word forms, many combinatorially possible analyses, e.g. Arbeitsaufwand ‘work effort, expenditure of labor’

Figure 1: Ambiguous analysis of *Arbeitsaufwand* ‘expenditure of labor’

Characteristics of German Word-Formation I

- language with highly productive and complex processes of word formation
- most common: compounding and derivation
- long orthographical word forms, many combinatorially possible analyses, e.g. Arbeitsaufwand ‘work effort, expenditure of labor’

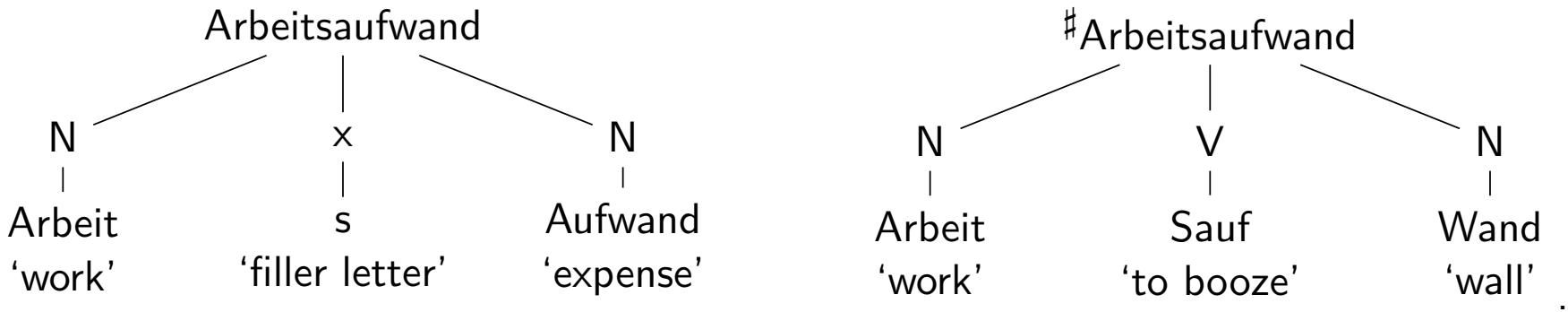


Figure 1: Ambiguous analysis of *Arbeitsaufwand* ‘expenditure of labor’

Characteristics of German Word-Formation II

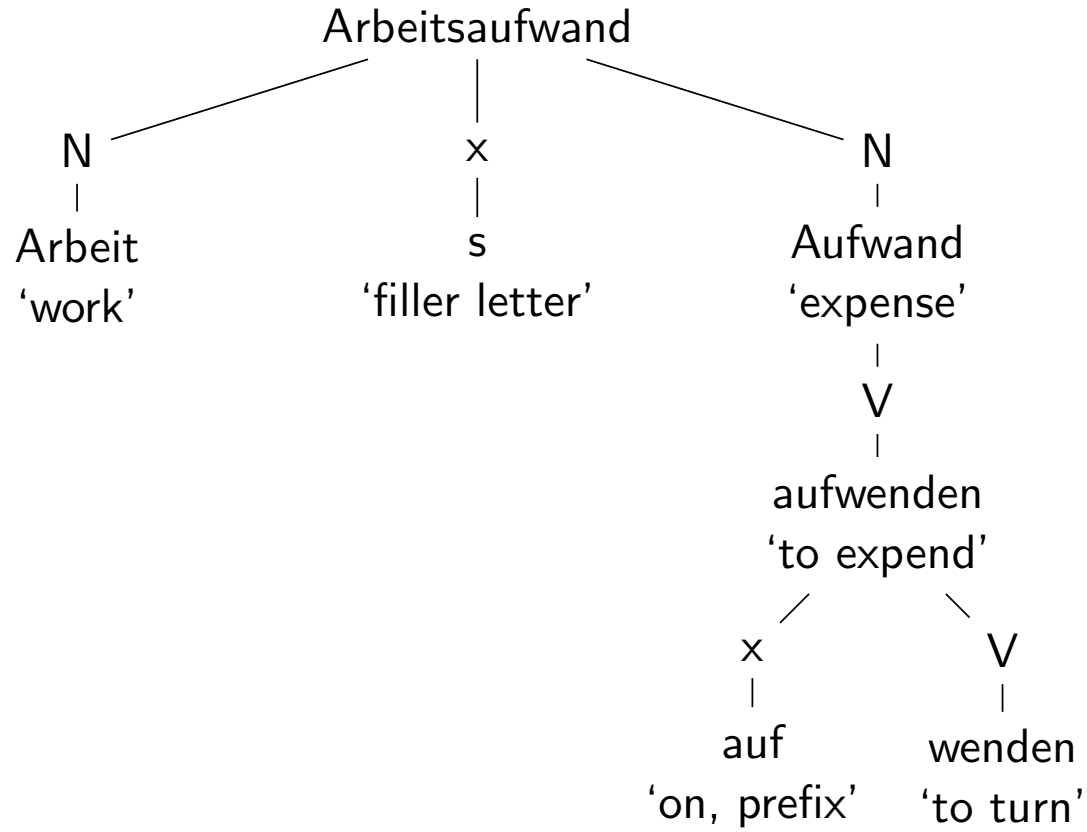


Figure 2: Deep analysis of *Arbeitsaufwand* ‘expenditure of labor’

Outline

Combining Data-Intense Methods with Contextual Retrieval

A hybrid approach for finding the correct splits of complex German words by using

- A formerly derived morphological trees database (Steiner, 2017)
- adjusted output of a morphological splitter
- co(n)texts from 1.8 Mio Wikipedia texts
- morphological segmentation based on corpus frequencies
- quantitative properties of German morpheme lengths

Combining Data-Intense Methods with Contextual Retrieval

A hybrid approach for finding the correct splits of complex German words by using

- A formerly derived morphological trees database (Steiner, 2017)
- adjusted output of a morphological splitter
- co(n)texts from 1.8 Mio Wikipedia texts
- morphological segmentation based on corpus frequencies
- quantitative properties of German morpheme lengths

Combining Data-Intense Methods with Contextual Retrieval

A hybrid approach for finding the correct splits of complex German words by using

- A formerly derived morphological trees database (Steiner, 2017)
- adjusted output of a morphological splitter
- co(n)texts from 1.8 Mio Wikipedia texts
- morphological segmentation based on corpus frequencies
- quantitative properties of German morpheme lengths

Combining Data-Intense Methods with Contextual Retrieval

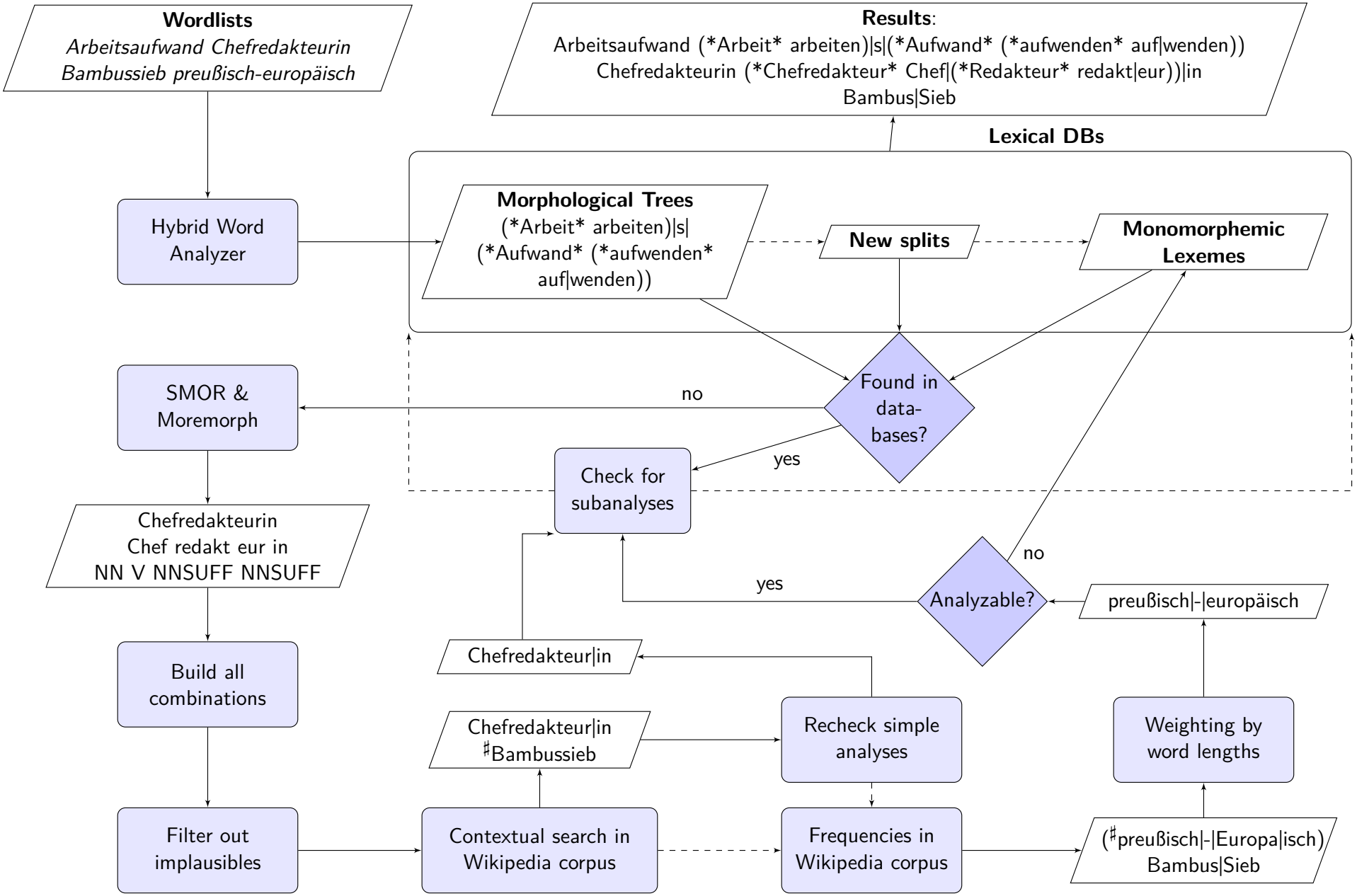
A hybrid approach for finding the correct splits of complex German words by using

- A formerly derived morphological trees database (Steiner, 2017)
- adjusted output of a morphological splitter
- co(n)texts from 1.8 Mio Wikipedia texts
- morphological segmentation based on corpus frequencies
- quantitative properties of German morpheme lengths

Combining Data-Intense Methods with Contextual Retrieval

A hybrid approach for finding the correct splits of complex German words by using

- A formerly derived morphological trees database (Steiner, 2017)
- adjusted output of a morphological splitter
- co(n)texts from 1.8 Mio Wikipedia texts
- morphological segmentation based on corpus frequencies
- quantitative properties of German morpheme lengths


Wordlists
Arbeitsaufwand Chefredakteurin
Bambussieb preußisch-europäisch
Results:
Arbeitsaufwand (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))
Chefredakteurin (*Chefredakteur* Chef|(*Redakteur* redakt|eur))|in
Bambus|Sieb
Lexical DBs
Hybrid Word Analyzer
Morphological Trees
(*Arbeit* arbeiten)|s|
(*Aufwand* (*aufwenden*
auf|wenden))
New splits
Monomorphemic Lexemes
Found in data-bases?
no
yes
SMOR & Moremorph
Check for subanalyses
Chefredakteurin
Chef redakt eur in
NN V NNSUFF NNSUFF
Analyzable?
no
yes
preußisch|-|europäisch
Chefredakteur|in
Build all combinations
Chefredakteur|in
♯Bambussieb
Recheck simple analyses
Weighting by word lengths
Filter out implausibles
Contextual search in Wikipedia corpus
Frequencies in Wikipedia corpus
(♯preußisch|-|Europa|isch)
Bambus|Sieb

Morphological Trees Database

- from CELEX, German part, and GermaNet database
- 101,588 entries
- Example: *Arbeitsaufwand*
  (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))

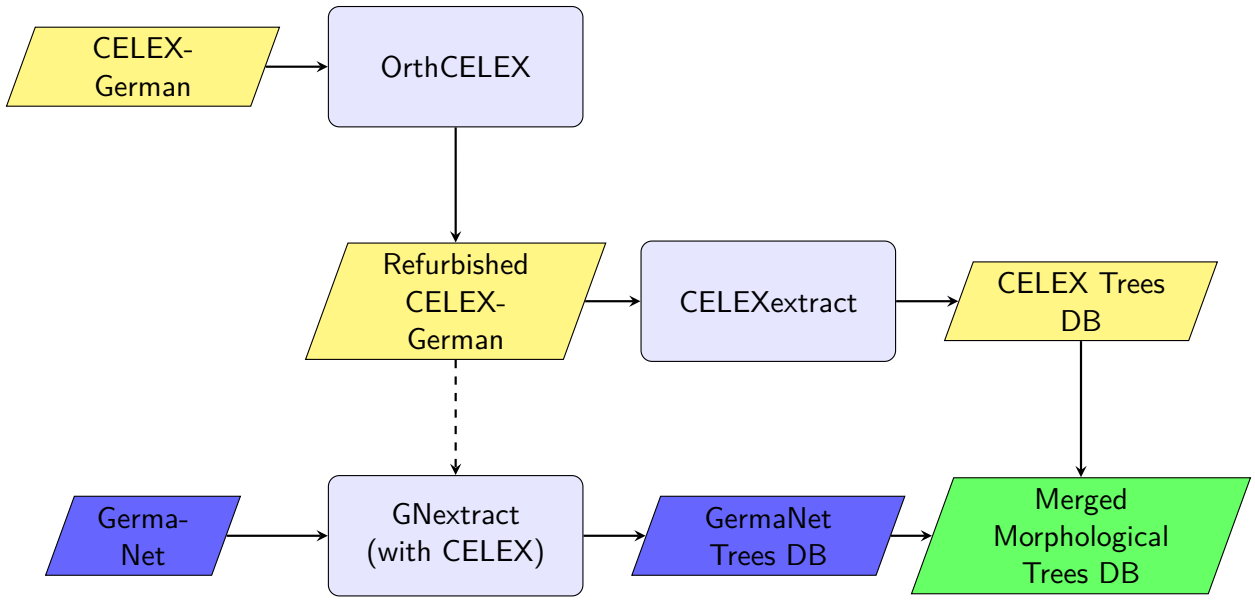
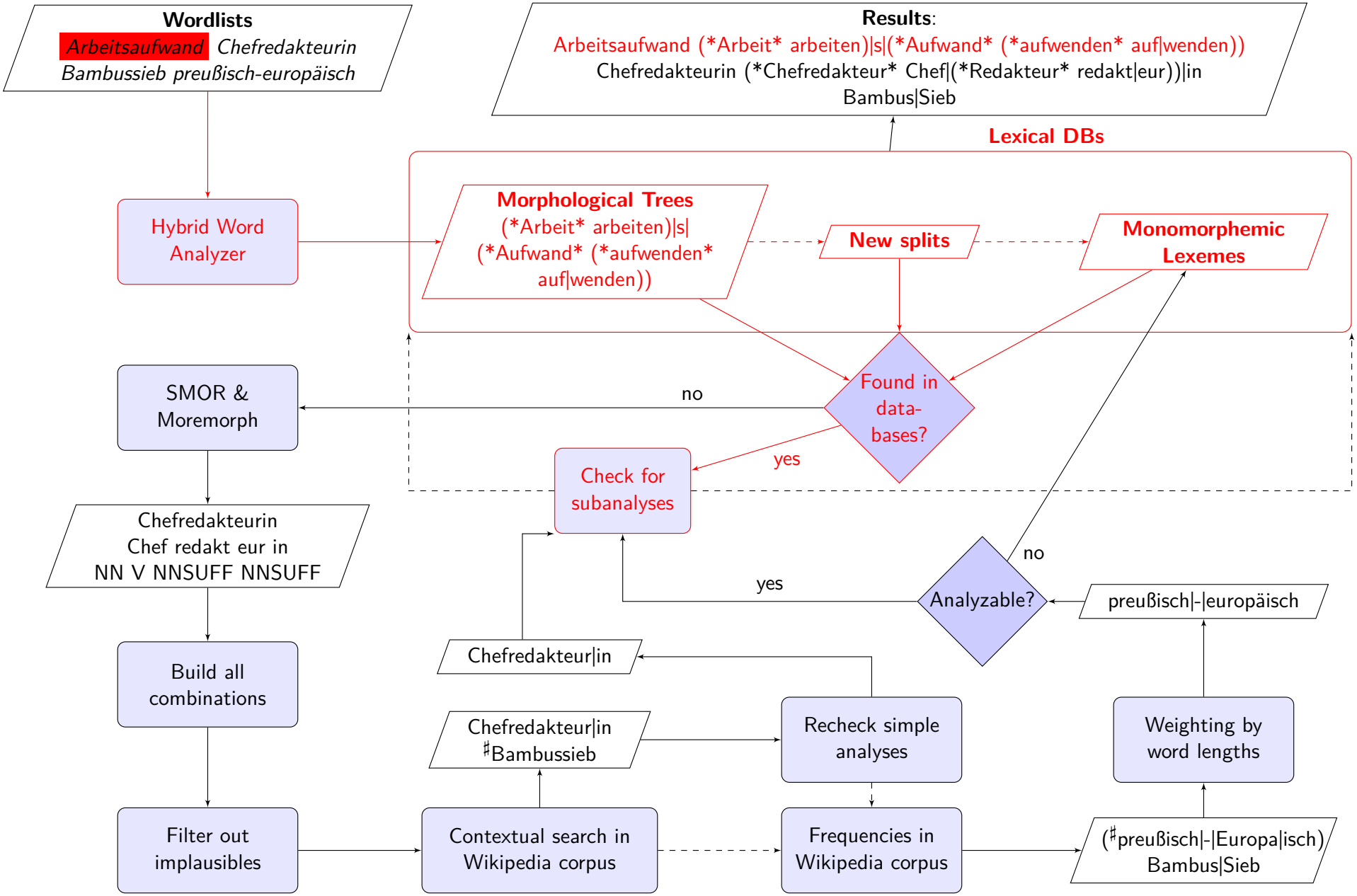
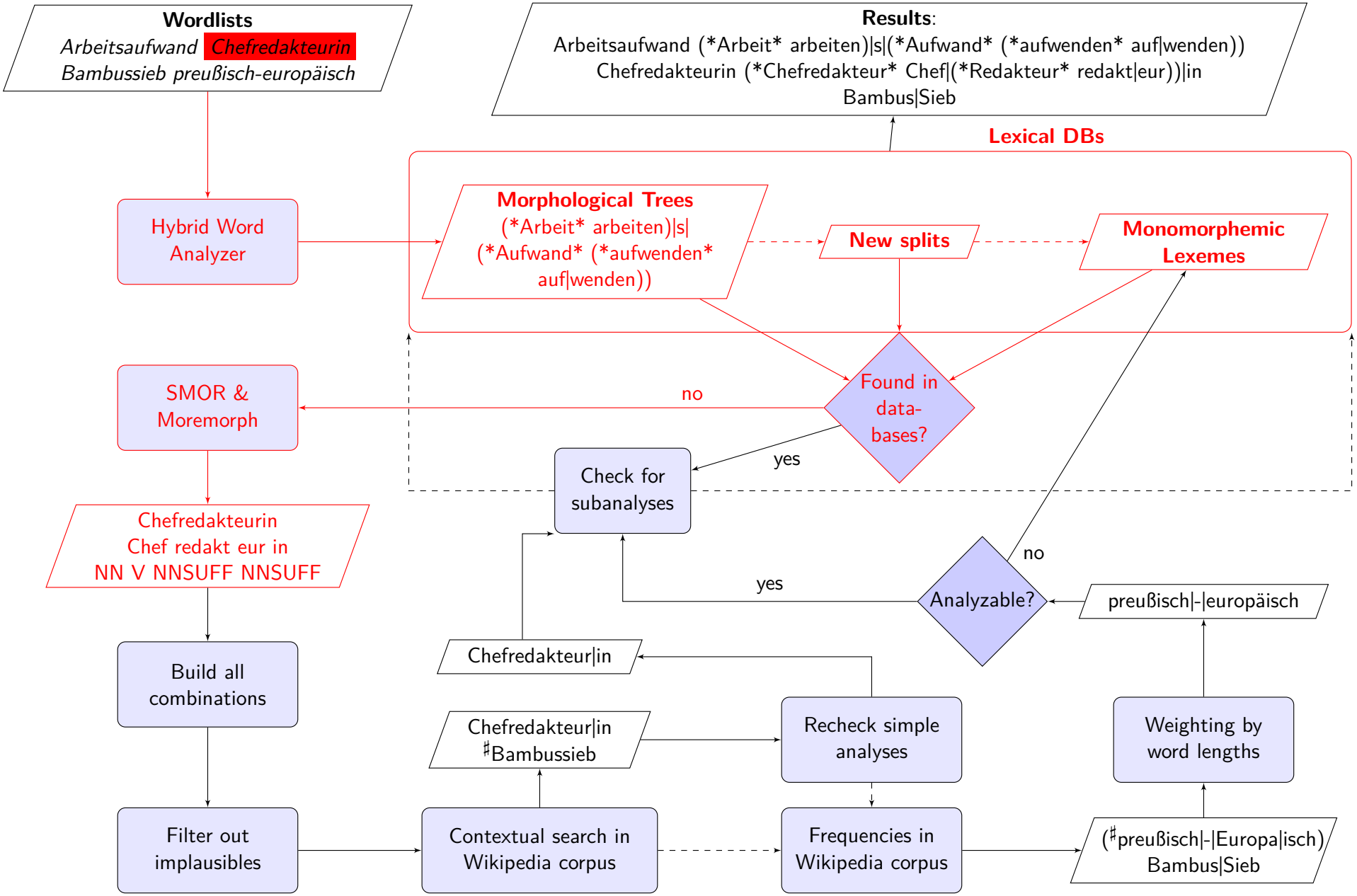


Figure 3: Extracting and Merging Morphological Trees


Wordlists
Arbeitsaufwand Chefredakteurin
Bambussieb preußisch-europäisch
Results:
Arbeitsaufwand (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))
Chefredakteurin (*Chefredakteur* Chef|(*Redakteur* redakt|eur))|in
Bambus|Sieb
Lexical DBs
Hybrid Word Analyzer
Morphological Trees
(*Arbeit* arbeiten)|s|
(*Aufwand* (*aufwenden*
auf|wenden))
New splits
Monomorphemic Lexemes
Found in data-bases?
no
yes
SMOR & Moremorph
Check for subanalyses
Chefredakteurin
Chef redakt eur in
NN V NNSUFF NNSUFF
Analyzable?
no
yes
preußisch|-|europäisch
Chefredakteur|in
Build all combinations
Chefredakteur|in
♯Bambussieb
Recheck simple analyses
Weighting by word lengths
Filter out implausibles
Contextual search in Wikipedia corpus
Frequencies in Wikipedia corpus
(♯preußisch|-|Europa|isch)
Bambus|Sieb


Wordlists
Arbeitsaufwand Chefredakteurin
Bambussieb preußisch-europäisch
Results:
Arbeitsaufwand (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))
Chefredakteurin (*Chefredakteur* Chef|(*Redakteur* redakt|eur))|in
Bambus|Sieb
Lexical DBs
Hybrid Word Analyzer
Morphological Trees
(*Arbeit* arbeiten)|s|
(*Aufwand* (*aufwenden*
auf|wenden))
New splits
Monomorphemic Lexemes
Found in data-bases?
no
yes
SMOR & Moremorph
Check for subanalyses
Chefredakteurin
Chef redakt eur in
NN V NNSUFF NNSUFF
Analyzable?
no
yes
preußisch|-|europäisch
Chefredakteur|in
Build all combinations
Chefredakteur|in
♯Bambussieb
Recheck simple analyses
Weighting by word lengths
Filter out implausibles
Contextual search in Wikipedia corpus
Frequencies in Wikipedia corpus
(♯preußisch|-|Europa|isch)
Bambus|Sieb

SMOR: A Morphological Tool for German

- Stuttgarter Morphological Analysis Tool, adjusted by the add-on Moremorph
- Main lexicon with 42,205 entries, proper name lexicons with 16,718 entries and different datasets with other morphological information

(1) Example output for *Chefredakteurin* 'editor-in-chief$_{female}$':
Chef R:redakteur in
Chef rede:<>n:<> A:akte U:urin
Chef rede:<>n:<> A:akteur in
Chef rede:<>n:<> A:akt e U:urin
Chef rede:<>n:<> akt eur in
Chef rede:<>n:<> A:akte U:urin
Chef rede:<>n:<> A:akteur in
Chef rede:<>n:<> A:akt e U:urin
Chef rede:<>n:<> akt eur in

Chef redakt eur in

(2) a. [[NN,NN],[NNSUFF]] Chefredakteur|in
b. ♯[[NN],[NN, NNSUFF]] Chef|redakteurin
c. ♯[[NN, NN, NNSUFF]] Chefredakteurin

SMOR: A Morphological Tool for German

- Stuttgarter Morphological Analysis Tool, adjusted by the add-on Moremorph
- Main lexicon with 42,205 entries, proper name lexicons with 16,718 entries and different datasets with other morphological information

(1) Example output for *Chefredakteurin* 'editor-in-chief$_{female}$':
Chef R:redakteur in
Chef rede:<>n:<> A:akte U:urin
Chef rede:<>n:<> A:akteur in
Chef rede:<>n:<> A:akt e U:urin
Chef rede:<>n:<> akt eur in
Chef rede:<>n:<> A:akte U:urin
Chef rede:<>n:<> A:akteur in
Chef rede:<>n:<> A:akt e U:urin
Chef rede:<>n:<> akt eur in

Chef redakt eur in

(2) a. [[NN,NN],[NNSUFF]] Chefredakteur|in
b. #[[NN],[NN, NNSUFF]] Chef|redakteurin
c. #[[NN, NN, NNSUFF]] Chefredakteurin

SMOR: A Morphological Tool for German

- Stuttgarter Morphological Analysis Tool, adjusted by the add-on Moremorph
- Main lexicon with 42,205 entries, proper name lexicons with 16,718 entries and different datasets with other morphological information

(1) Example output for *Chefredakteurin* 'editor-in-chief$_{female}$':
Chef R:redakteur in
Chef rede:<>n:<> A:akte U:urin
Chef rede:<>n:<> A:akteur in
Chef rede:<>n:<> A:akt e U:urin
Chef rede:<>n:<> akt eur in
Chef rede:<>n:<> A:akte U:urin
Chef rede:<>n:<> A:akteur in
Chef rede:<>n:<> A:akt e U:urin
Chef rede:<>n:<> akt eur in

Chef redakt eur in

(2) a. [[NN,NN],[NNSUFF]] Chefredakteur|in
b. #[[NN],[NN, NNSUFF]] Chef|redakteurin
c. #[[NN, NN, NNSUFF]] Chefredakteurin

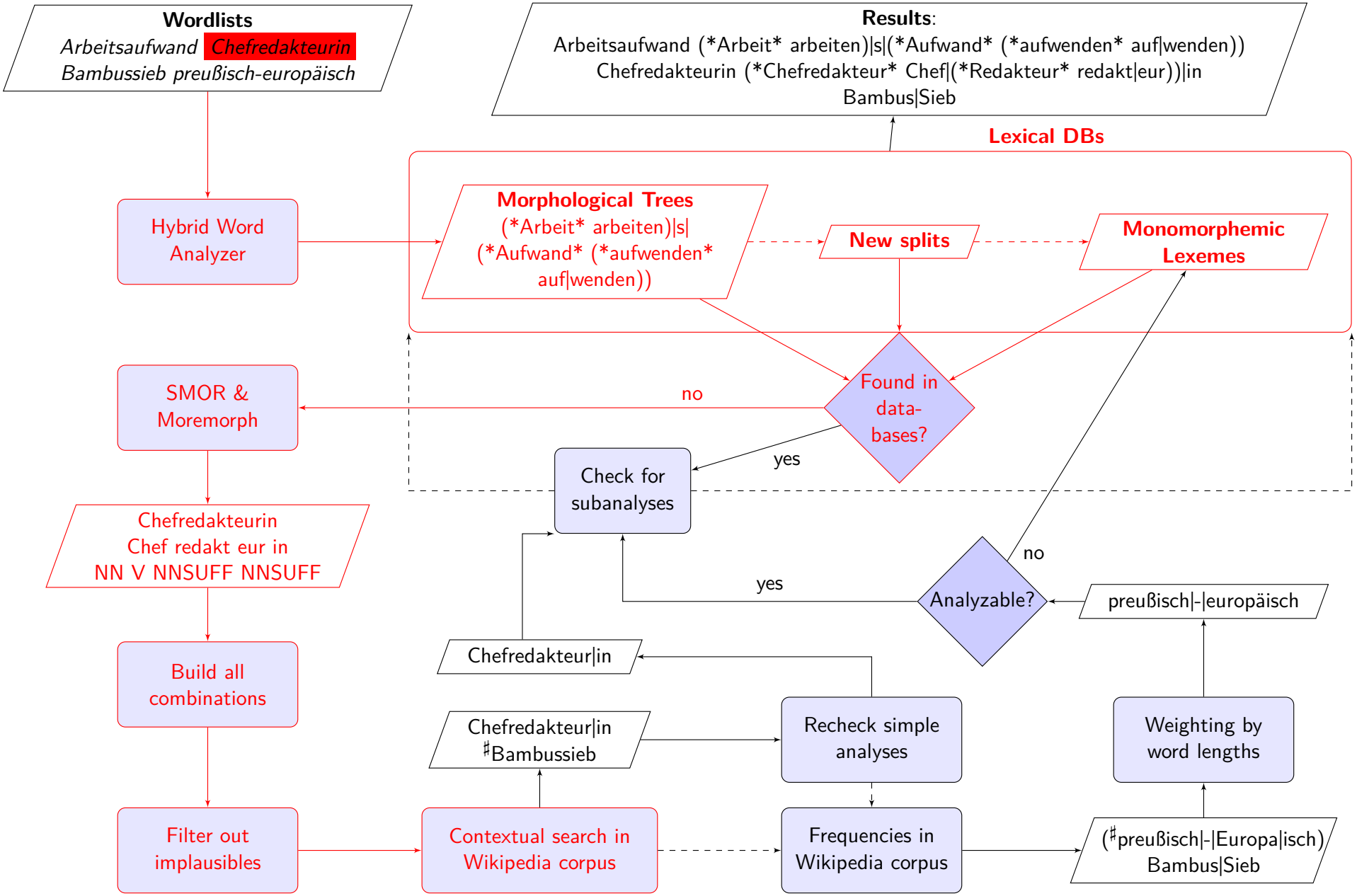
SMOR: A Morphological Tool for German

- Stuttgarter Morphological Analysis Tool, adjusted by the add-on Moremorph
- Main lexicon with 42,205 entries, proper name lexicons with 16,718 entries and different datasets with other morphological information

(1) Example output for *Chefredakteurin* 'editor-in-chief$_{female}$':
Chef R:redakteur in
Chef rede:<>n:<> A:akte U:urin
Chef rede:<>n:<> A:akteur in
Chef rede:<>n:<> A:akt e U:urin
Chef rede:<>n:<> akt eur in
Chef rede:<>n:<> A:akte U:urin
Chef rede:<>n:<> A:akteur in
Chef rede:<>n:<> A:akt e U:urin
Chef rede:<>n:<> akt eur in

Chef redakt eur in

(2) a. [[NN,NN],[NNSUFF]] Chefredakteur|in
b. #[[NN],[NN, NNSUFF]] Chef|redakteurin
c. #[[NN, NN, NNSUFF]] Chefredakteurin


Wordlists
Arbeitsaufwand Chefredakteurin
Bambussieb preußisch-europäisch
Results:
Arbeitsaufwand (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))
Chefredakteurin (*Chefredakteur* Chef|(*Redakteur* redakt|eur))|in
Bambus|Sieb
Lexical DBs
Hybrid Word Analyzer
Morphological Trees
(*Arbeit* arbeiten)|s|
(*Aufwand* (*aufwenden*
auf|wenden))
New splits
Monomorphemic Lexemes
Found in data-bases?
no
yes
SMOR & Moremorph
Check for subanalyses
Chefredakteurin
Chef redakt eur in
NN V NNSUFF NNSUFF
Analyzable?
no
yes
preußisch|-|europäisch
Build all combinations
Chefredakteur|in
Chefredakteur|in
♯Bambussieb
Recheck simple analyses
Weighting by word lengths
Filter out implausibles
Contextual search in Wikipedia corpus
Frequencies in Wikipedia corpus
(♯preußisch|-|Europa|isch)
Bambus|Sieb

Idea: For splits of unknown compounds, each immediate constituent should be found within the context at least somewhere inside a large corpus. For derivatives, this holds only for hypothetical constituents which are free morphs or lexemes.

- Contexts: the texts of a corpus in which the respective analyzed word form occurs. Corpus: 1.8 million articles of the annotated German Wikipedia Korpus of 2015 (Margaretha and Lüngen, 2014)
- Tokenizer: a modified version of the tool from Dipper (2016); lemmatizer: TreeTagger (Schmid, 1999)
- Text indices: for the tokenized and lemmatized forms.
- For each text containing the input word form W_{wf}, the document frequencies $(df_1 \dots df_m)$ of the free hypothetical immediate constituents $(c_{wf,s,1} \dots c_{wf,s,n})$ are being retrieved and summarized. This yields a text frequency score $(S_{wf,s,t})$ for each text and split of n constituents.

$$S_{wf,s,t} = \sum_{c=1}^{n} df_i \tag{1}$$

Of all morphological analyses for W_{wf}, the one with the largest score is processed for the storage. A missing hypothetical constituent inside a text containing W_{wf} leads to a document frequency of 0 for this constituent, which can be compensated by the frequencies of the other constituents of the split sequence.

Idea: For splits of unknown compounds, each immediate constituent should be found within the context at least somewhere inside a large corpus. For derivatives, this holds only for hypothetical constituents which are free morphs or lexemes.

- Contexts: the texts of a corpus in which the respective analyzed word form occurs. Corpus: 1.8 million articles of the annotated German Wikipedia Korpus of 2015 (Margaretha and Lüngen, 2014)
- Tokenizer: a modified version of the tool from Dipper (2016); lemmatizer: TreeTagger (Schmid, 1999)
- Text indices: for the tokenized and lemmatized forms.
- For each text containing the input word form W_{wf}, the document frequencies $(df_1 \dots df_m)$ of the free hypothetical immediate constituents $(c_{wf,s,1} \dots c_{wf,s,n})$ are being retrieved and summarized. This yields a text frequency score $(S_{wf,s,t})$ for each text and split of n constituents.

$$S_{wf,s,t} = \sum_{c=1}^{n} df_i \tag{1}$$

Of all morphological analyses for W_{wf}, the one with the largest score is processed for the storage. A missing hypothetical constituent inside a text containing W_{wf} leads to a document frequency of 0 for this constituent, which can be compensated by the frequencies of the other constituents of the split sequence.

Idea: For splits of unknown compounds, each immediate constituent should be found within the context at least somewhere inside a large corpus. For derivatives, this holds only for hypothetical constituents which are free morphs or lexemes.

- Contexts: the texts of a corpus in which the respective analyzed word form occurs. Corpus: 1.8 million articles of the annotated German Wikipedia Korpus of 2015 (Margaretha and Lüngen, 2014)
- Tokenizer: a modified version of the tool from Dipper (2016); lemmatizer: TreeTagger (Schmid, 1999)
- Text indices: for the tokenized and lemmatized forms.
- For each text containing the input word form W_{wf}, the document frequencies $(df_1 \dots df_m)$ of the free hypothetical immediate constituents $(c_{wf,s,1} \dots c_{wf,s,n})$ are being retrieved and summarized. This yields a text frequency score $(S_{wf,s,t})$ for each text and split of n constituents.

$$S_{wf,s,t} = \sum_{c=1}^{n} df_i \tag{1}$$

Of all morphological analyses for W_{wf}, the one with the largest score is processed for the storage. A missing hypothetical constituent inside a text containing W_{wf} leads to a document frequency of 0 for this constituent, which can be compensated by the frequencies of the other constituents of the split sequence.

Idea: For splits of unknown compounds, each immediate constituent should be found within the context at least somewhere inside a large corpus. For derivatives, this holds only for hypothetical constituents which are free morphs or lexemes.

- Contexts: the texts of a corpus in which the respective analyzed word form occurs. Corpus: 1.8 million articles of the annotated German Wikipedia Korpus of 2015 (Margaretha and Lüngen, 2014)
- Tokenizer: a modified version of the tool from Dipper (2016); lemmatizer: TreeTagger (Schmid, 1999)
- Text indices: for the tokenized and lemmatized forms.
- For each text containing the input word form W_{wf}, the document frequencies $(df_1 \dots df_m)$ of the free hypothetical immediate constituents $(c_{wf,s,1} \dots c_{wf,s,n})$ are being retrieved and summarized. This yields a text frequency score $(S_{wf,s,t})$ for each text and split of n constituents.

$$S_{wf,s,t} = \sum_{c=1}^{n} df_i \tag{1}$$

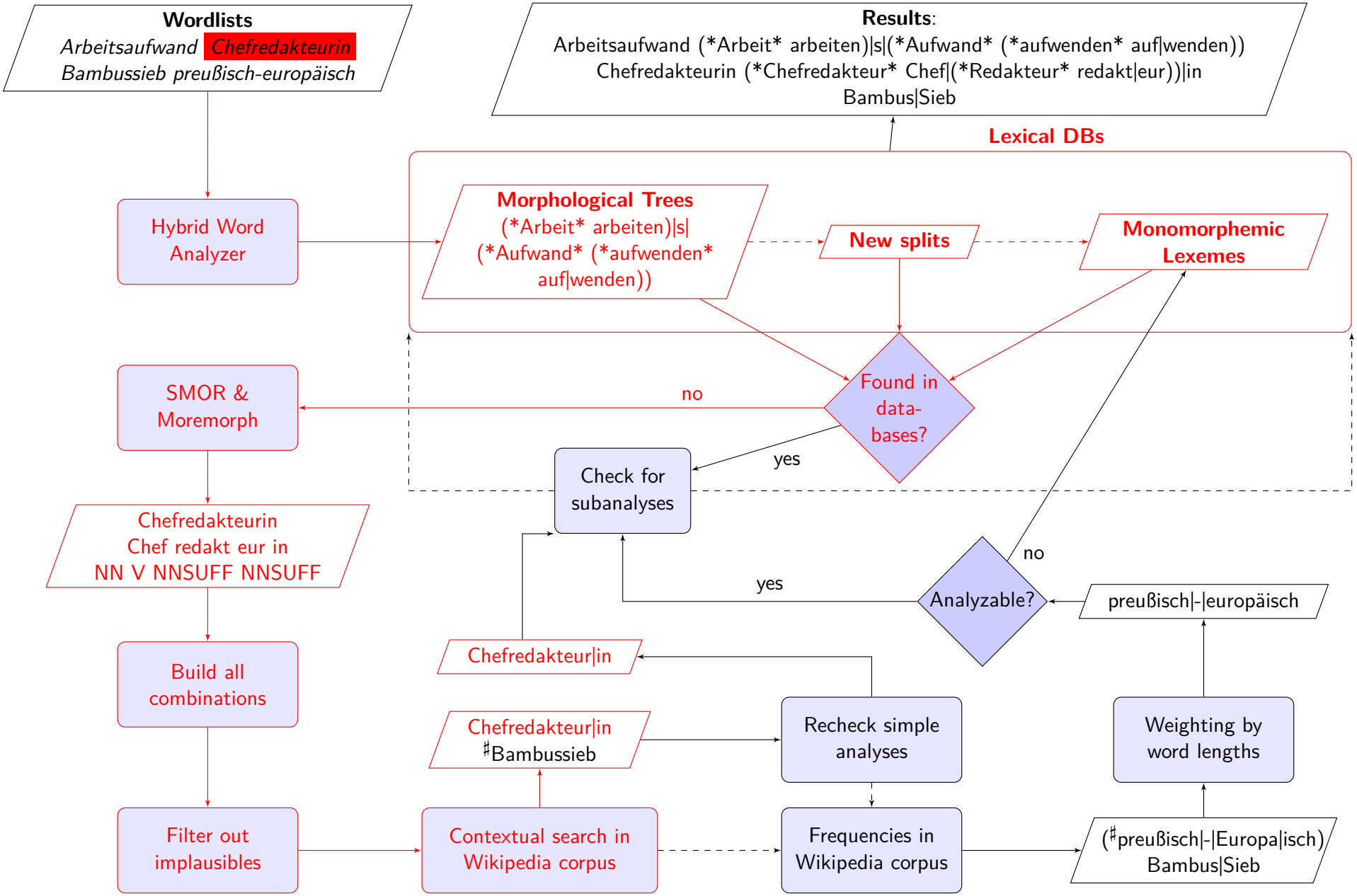
Of all morphological analyses for W_{wf}, the one with the largest score is processed for the storage. A missing hypothetical constituent inside a text containing W_{wf} leads to a document frequency of 0 for this constituent, which can be compensated by the frequencies of the other constituents of the split sequence.

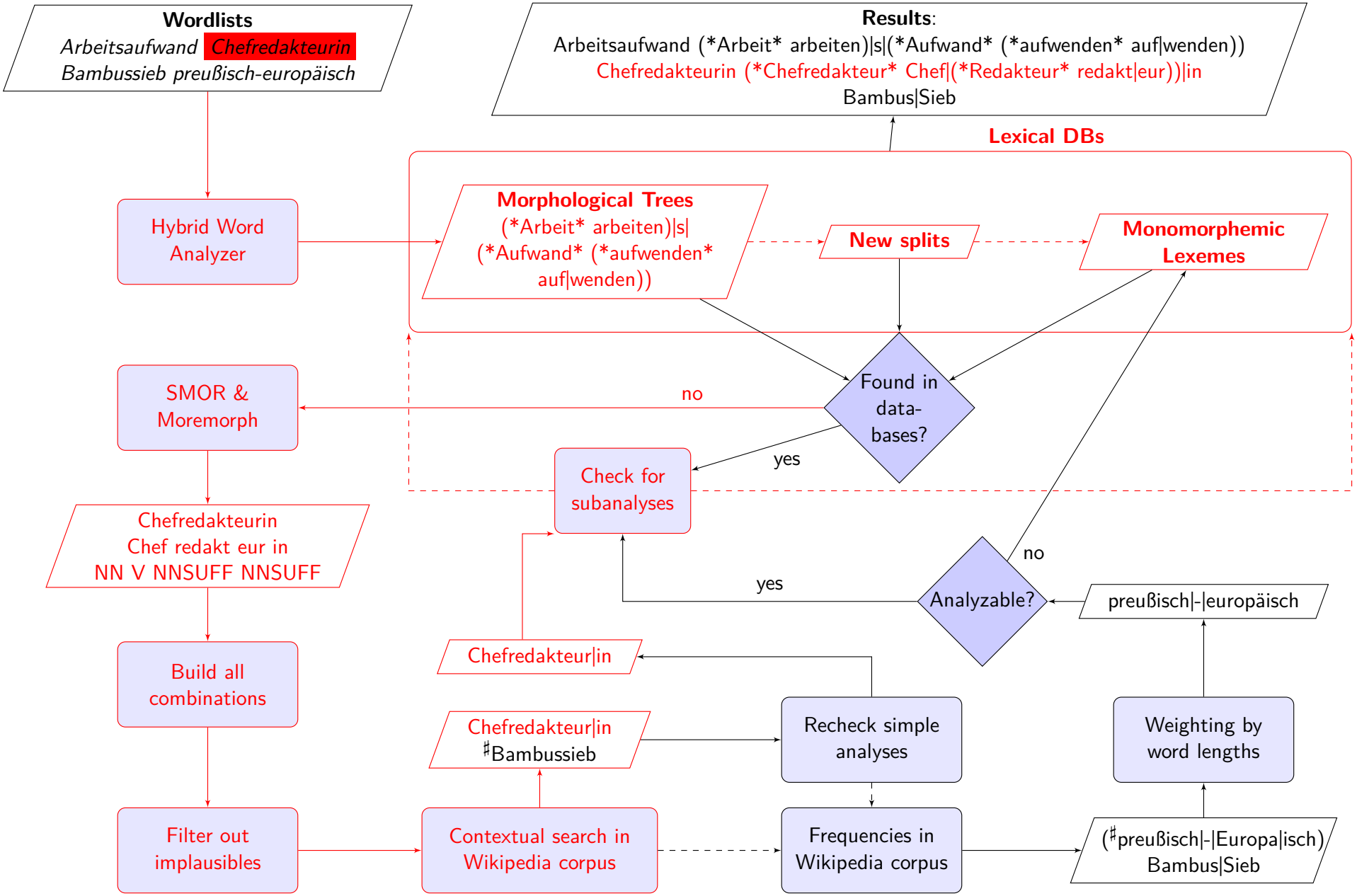
Idea: For splits of unknown compounds, each immediate constituent should be found within the context at least somewhere inside a large corpus. For derivatives, this holds only for hypothetical constituents which are free morphs or lexemes.

- Contexts: the texts of a corpus in which the respective analyzed word form occurs. Corpus: 1.8 million articles of the annotated German Wikipedia Korpus of 2015 (Margaretha and Lüngen, 2014)
- Tokenizer: a modified version of the tool from Dipper (2016); lemmatizer: TreeTagger (Schmid, 1999)
- Text indices: for the tokenized and lemmatized forms.
- For each text containing the input word form W_{wf}, the document frequencies $(df_1 \dots df_m)$ of the free hypothetical immediate constituents $(c_{wf,s,1} \dots c_{wf,s,n})$ are being retrieved and summarized. This yields a text frequency score $(S_{wf,s,t})$ for each text and split of n constituents.

$$S_{wf,s,t} = \sum_{c=1}^{n} df_i \tag{1}$$

Of all morphological analyses for W_{wf}, the one with the largest score is processed for the storage. A missing hypothetical constituent inside a text containing W_{wf} leads to a document frequency of 0 for this constituent, which can be compensated by the frequencies of the other constituents of the split sequence.


Wordlists
Arbeitsaufwand Chefredakteurin
Bambussieb preußisch-europäisch
Results:
Arbeitsaufwand (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))
Chefredakteurin (*Chefredakteur* Chef|(*Redakteur* redakt|eur))|in
Bambus|Sieb
Lexical DBs
Hybrid Word Analyzer
Morphological Trees
(*Arbeit* arbeiten)|s|
(*Aufwand* (*aufwenden*
auf|wenden))
New splits
Monomorphemic Lexemes
Found in data-bases?
no
yes
SMOR & Moremorph
Check for subanalyses
Chefredakteurin
Chef redakt eur in
NN V NNSUFF NNSUFF
Analyzable?
no
yes
preußisch|-|europäisch
Build all combinations
Chefredakteur|in
Chefredakteur|in
♯Bambussieb
Recheck simple analyses
Weighting by word lengths
Filter out implausibles
Contextual search in Wikipedia corpus
Frequencies in Wikipedia corpus
(♯preußisch|-|Europa|isch)
Bambus|Sieb


Wordlists
Arbeitsaufwand Chefredakteurin
Bambussieb preußisch-europäisch
Results:
Arbeitsaufwand (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))
Chefredakteurin (*Chefredakteur* Chef|(*Redakteur* redakt|eur))|in
Bambus|Sieb
Lexical DBs
Hybrid Word Analyzer
Morphological Trees
(*Arbeit* arbeiten)|s|
(*Aufwand* (*aufwenden*
auf|wenden))
New splits
Monomorphemic Lexemes
Found in data-bases?
no
yes
SMOR & Moremorph
Check for subanalyses
Chefredakteurin
Chef redakt eur in
NN V NNSUFF NNSUFF
Analyzable?
no
yes
preußisch|-|europäisch
Chefredakteur|in
Build all combinations
Chefredakteur|in
♯Bambussieb
Recheck simple analyses
Weighting by word lengths
Filter out implausibles
Contextual search in Wikipedia corpus
Frequencies in Wikipedia corpus
(♯preußisch|-|Europa|isch)
Bambus|Sieb

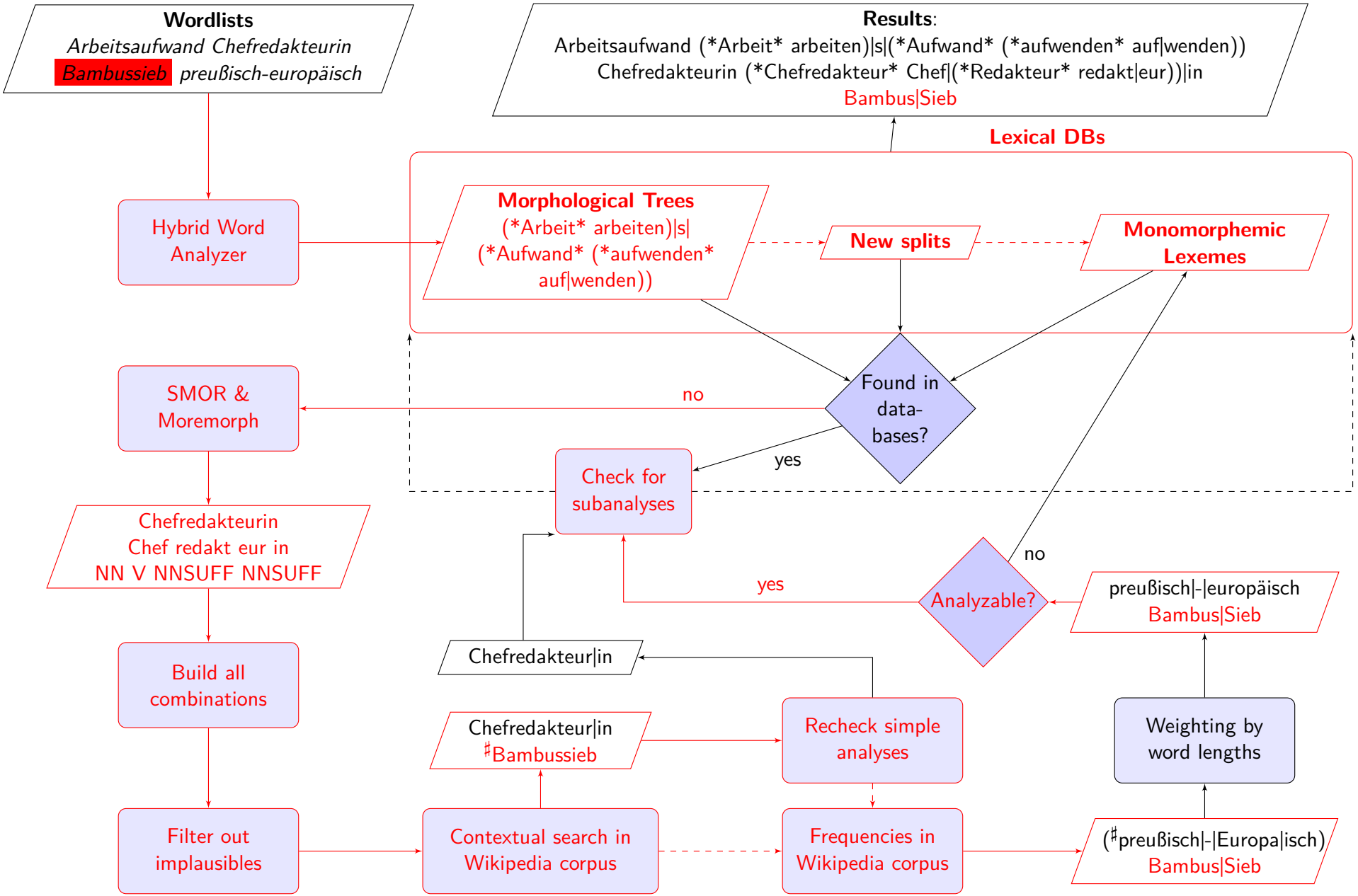
Corpus Frequencies

The corpus itself is considered as a context in the widest sense if

- no text contains the word form W_{wf}
- a double check for longer word forms is advisable

(3) *Bambussieb* 'bamboo screen'
a. [[NN],[NN]] Bambus|Sieb
b. ♯[[NN, NN]] Bambussieb

- Investigations on the lengths of German morphs show that German simplex lexemes rarely possess more than 7 phonemes (98.41%) (Menzerath, 1954; Gerlach, 1982). The number of graphemes is proportional and slightly larger (Krott, 1996).
- Check: all word forms with more than 8 characters if a. the contextual search found only splits comprising just one constituent but b. hypothetical splits with more than one constituent do exist.


Wordlists
Arbeitsaufwand Chefredakteurin
Bambussieb preußisch-europäisch
Results:
Arbeitsaufwand (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))
Chefredakteurin (*Chefredakteur* Chef|(*Redakteur* redakt|eur))|in
Bambus|Sieb
Lexical DBs
Hybrid Word Analyzer
Morphological Trees
(*Arbeit* arbeiten)|s|
(*Aufwand* (*aufwenden*
auf|wenden))
New splits
Monomorphemic Lexemes
Found in data-bases?
no
yes
SMOR & Moremorph
Check for subanalyses
Chefredakteurin
Chef redakt eur in
NN V NNSUFF NNSUFF
Analyzable?
no
yes
preußisch|-|europäisch
Bambus|Sieb
Chefredakteur|in
Build all combinations
Chefredakteur|in
♯Bambussieb
Recheck simple analyses
Weighting by word lengths
Filter out implausibles
Contextual search in Wikipedia corpus
Frequencies in Wikipedia corpus
(♯preußisch|-|Europa|isch)
Bambus|Sieb

The frequency-based weighting has a bias towards constructions with small constituents.

(4) a. ♯Figur|Kombi|Nation 'figure|combi (short form of combination)|nation'
b. Figur|Kombination 'figure|combination'
c. Figur|(*Kombination* kombin|ation)

- The functional dependency between morph/lexeme frequency and length is mutual (Köhler (1986), Krott (2004)) and influenced by other factors such as the age of words and lexicon size.
- For each constituent with a length of l characters, the frequency of its word length class L_l is used as an inverse proportional factor for the document frequencies (2).

$$WeightedS_{wf,s,t} = \sum_{c=1}^{n} \frac{df_i}{freq(L_{l(c)})} \quad (2)$$

The frequency-based weighting has a bias towards constructions with small constituents.

(4) a. ♯Figur|Kombi|Nation 'figure|combi (short form of combination)|nation'
b. Figur|Kombination 'figure|combination'
c. Figur|(*Kombination* kombin|ation)

- The functional dependency between morph/lexeme frequency and length is mutual (Köhler (1986), Krott (2004)) and influenced by other factors such as the age of words and lexicon size.
- For each constituent with a length of l characters, the frequency of its word length class L_l is used as an inverse proportional factor for the document frequencies (2).

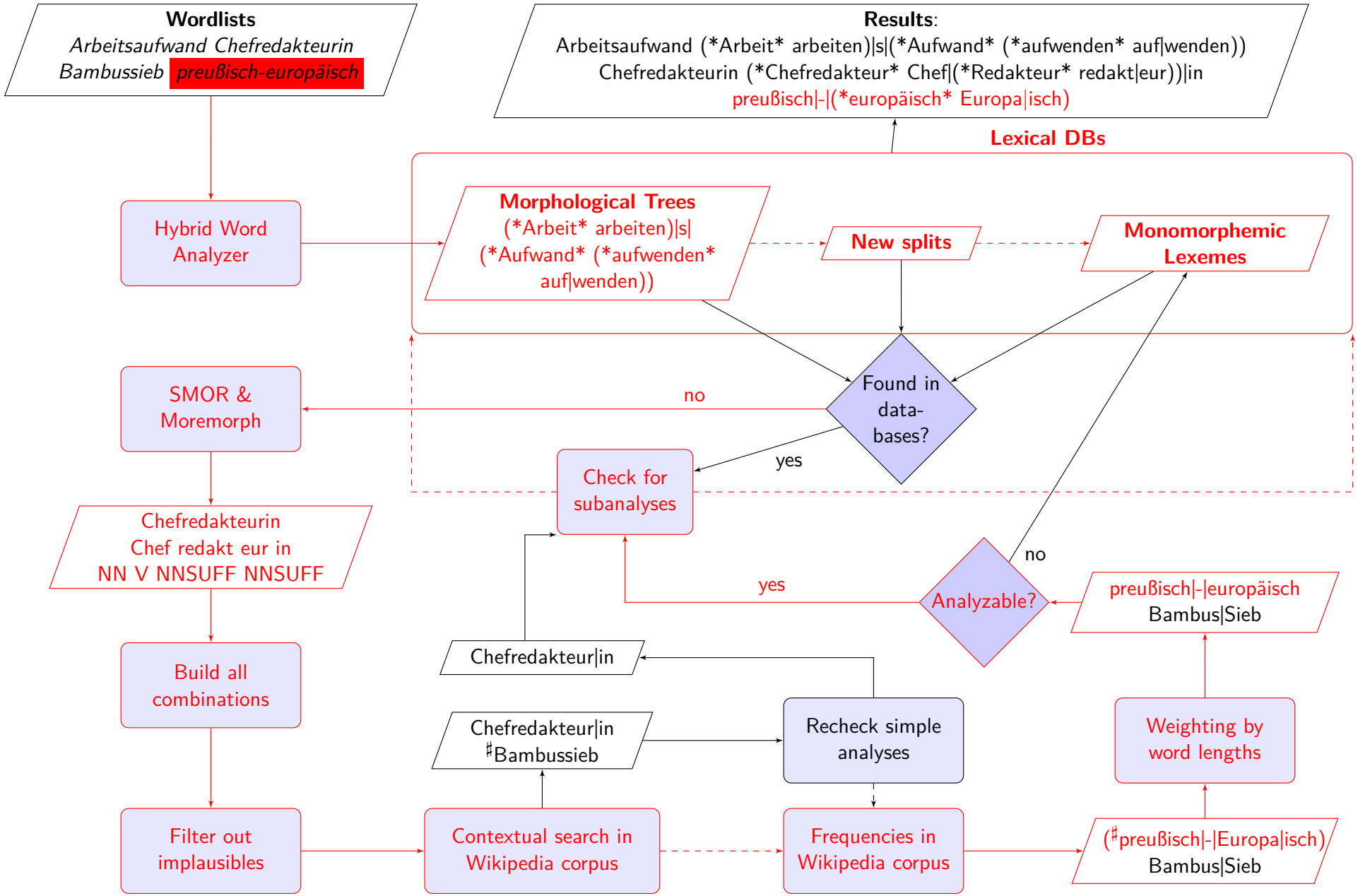
$$WeightedS_{wf,s,t} = \sum_{c=1}^{n} \frac{df_i}{freq(L_{l(c)})} \quad (2)$$

The frequency-based weighting has a bias towards constructions with small constituents.

(4) a. ♯Figur|Kombi|Nation 'figure|combi (short form of combination)|nation'
b. Figur|Kombination 'figure|combination'
c. Figur|(*Kombination* kombin|ation)

- The functional dependency between morph/lexeme frequency and length is mutual (Köhler (1986), Krott (2004)) and influenced by other factors such as the age of words and lexicon size.
- For each constituent with a length of l characters, the frequency of its word length class L_l is used as an inverse proportional factor for the document frequencies (2).

$$WeightedS_{wf,s,t} = \sum_{c=1}^{n} \frac{df_i}{freq(L_{l(c)})} \tag{2}$$


Wordlists
Arbeitsaufwand Chefredakteurin
Bambussieb preußisch-europäisch
Results:
Arbeitsaufwand (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))
Chefredakteurin (*Chefredakteur* Chef|(*Redakteur* redakt|eur))|in
preußisch|-|(*europäisch* Europa|isch)
Lexical DBs
Hybrid Word Analyzer
Morphological Trees
(*Arbeit* arbeiten)|s|
(*Aufwand* (*aufwenden*
auf|wenden))
New splits
Monomorphemic Lexemes
Found in data-bases?
no
yes
SMOR & Moremorph
Check for subanalyses
Chefredakteurin
Chef redakt eur in
NN V NNSUFF NNSUFF
Analyzable?
no
yes
preußisch|-|europäisch
Bambus|Sieb
Chefredakteur|in
Build all combinations
Chefredakteur|in
♯Bambussieb
Recheck simple analyses
Weighting by word lengths
Filter out implausibles
Contextual search in Wikipedia corpus
Frequencies in Wikipedia corpus
(♯preußisch|-|Europa|isch)
Bambus|Sieb

Outline

- Corpus: *Korpus Magazin Lufthansa Bordbuch (MLD)*, part of the DeReKo-2016-I (Institut für Deutsche Sprache 2016) corpus (see Kupietz et al. 2010), an in-flight magazine with articles on traveling, consumption and aviation.
- Tokenization: enlarged and costumized tokenizer by Dipper (2016)
- 276 texts with 5,202 paragraphs, 16,046 sentences and 260,114 tokens
- 38,337 word-form types, and 27,902 lemma types.
- 15,622 of these lemma types are inside the databases of trees or monomorphemic words. Coverage of 55.99% with an accuracy of nearly 100% due to the quality of the CELEX and GermaNet data.
- The remaining 44.01% of all lemma types were processed by SMOR and Moremorph with a coverage of 100%.
- sample of 1,006 word forms

	correct analysis (all levels)	no analysis	(partially) flat analysis	wrong analysis
DBs	≈ 55.99%			
DBs + Context + Corpus Look-up	87.77%	7.45%	3.48%	1.29%
+ Recheck	92.44%	2.68%	2.88%	1.99%
+ Recheck + Weighting	93.34%	2.58%	2.78%	1.29%

(5) a. adjective vs. participle: *folgend* 'following', *gewandt* 'turned$_v$, skillful$_{adj}$'
b. constituents not in context: *Metallkäfig* 'metal cage', *Tierärztin* 'vet$_{fem}$'
c. flat analyses: *Roll|vor|Gang* '♯?(to roll|prefix, before|gait), rolling procedure'
d. analysis from GermaNet: ♯?*(Land|Nahme)* '(land|"take"), settlement'
e. frequent homograph:
♯*(Parlament|(*arisch* ar|isch))* '(parliament|(*Aryan* Ar|ian), parliamentary'
f. correction by word-length weighting:
*rollen|(*Vorgang* (*vorgehen* vor|gehen))* 'to roll|(*procedure* (*to proceed* pro|ceed))'

5,696 new entries for monomorphemic lexemes; 8,448 for the new splits.

	correct analysis (all levels)	no analysis	(partially) flat analysis	wrong analysis
DBs	≈ 55.99%			
DBs + Context + Corpus Look-up	87.77%	7.45%	3.48%	1.29%
+ Recheck	92.44%	2.68%	2.88%	1.99%
+ Recheck + Weighting	93.34%	2.58%	2.78%	1.29%

(5) a. adjective vs. participle: *folgend* 'following', *gewandt* 'turned$_v$, skillful$_{adj}$'
b. constituents not in context: *Metallkäfig* 'metal cage', *Tierärztin* 'vet$_{fem}$'
c. flat analyses: *Roll|vor|Gang* '♯?(to roll|prefix, before|gait), rolling procedure'
d. analysis from GermaNet: ♯?*(Land|Nahme)* '(land|"take"), settlement'
e. frequent homograph:
♯*(Parlament|(*arisch* ar|isch))* '(parliament|(*Aryan* Ar|ian), parliamentary'
f. correction by word-length weighting:
*rollen|(*Vorgang* (*vorgehen* vor|gehen))* 'to roll|(*procedure* (*to proceed* pro|ceed))'

5,696 new entries for monomorphemic lexemes; 8,448 for the new splits.

	correct analysis (all levels)	no analysis	(partially) flat analysis	wrong analysis
DBs	≈ 55.99%			
DBs + Context + Corpus Look-up	87.77%	7.45%	3.48%	1.29%
+ Recheck	92.44%	2.68%	2.88%	1.99%
+ Recheck + Weighting	93.34%	2.58%	2.78%	1.29%

(5) a. adjective vs. participle: *folgend* 'following', *gewandt* 'turned$_v$, skillful$_{adj}$'
b. constituents not in context: *Metallkäfig* 'metal cage', *Tierärztin* 'vet$_{fem}$'
c. flat analyses: *Roll|vor|Gang* '♯?(to roll|prefix, before|gait), rolling procedure'
d. analysis from GermaNet: ♯?*(Land|Nahme)* '(land|"take"), settlement'
e. frequent homograph:
♯*(Parlament|(*arisch* ar|isch))* '(parliament|(*Aryan* Ar|ian), parliamentary'
f. correction by word-length weighting:
*rollen|(*Vorgang* (*vorgehen* vor|gehen))* 'to roll|(*procedure* (*to proceed* pro|ceed))'

5,696 new entries for monomorphemic lexemes; 8,448 for the new splits.

	correct analysis (all levels)	no analysis	(partially) flat analysis	wrong analysis
DBs	≈ 55.99%			
DBs + Context + Corpus Look-up	87.77%	7.45%	3.48%	1.29%
+ Recheck	92.44%	2.68%	2.88%	1.99%
+ Recheck + Weighting	93.34%	2.58%	2.78%	1.29%

(5) a. adjective vs. participle: *folgend* 'following', *gewandt* 'turned$_v$, skillful$_{adj}$'
b. constituents not in context: *Metallkäfig* 'metal cage', *Tierärztin* 'vet$_{fem}$'
c. flat analyses: *Roll|vor|Gang* '♯?(to roll|prefix, before|gait), rolling procedure'
d. analysis from GermaNet: ♯?*(Land|Nahme)* '(land|"take"), settlement'
e. frequent homograph:
♯*(Parlament|(*arisch* ar|isch))* '(parliament|(*Aryan* Ar|ian), parliamentary'
f. correction by word-length weighting:
*rollen|(*Vorgang* (*vorgehen* vor|gehen))* 'to roll|(*procedure* (*to proceed* pro|ceed))'

5,696 new entries for monomorphemic lexemes; 8,448 for the new splits.

	correct analysis (all levels)	no analysis	(partially) flat analysis	wrong analysis
DBs	≈ 55.99%			
DBs + Context + Corpus Look-up	87.77%	7.45%	3.48%	1.29%
+ Recheck	92.44%	2.68%	2.88%	1.99%
+ Recheck + Weighting	93.34%	2.58%	2.78%	1.29%

(5) a. adjective vs. participle: *folgend* 'following', *gewandt* 'turned$_v$, skillful$_{adj}$'
b. constituents not in context: *Metallkäfig* 'metal cage', *Tierärztin* 'vet$_{fem}$'
c. flat analyses: *Roll|vor|Gang* '♯?(to roll|prefix, before|gait), rolling procedure'
d. analysis from GermaNet: ♯?*(Land|Nahme)* '(land|"take"), settlement'
e. frequent homograph:
♯*(Parlament|(*arisch* ar|isch))* '(parliament|(*Aryan* Ar|ian), parliamentary'
f. correction by word-length weighting:
*rollen|(*Vorgang* (*vorgehen* vor|gehen))* 'to roll|(*procedure* (*to proceed* pro|ceed))'

5,696 new entries for monomorphemic lexemes; 8,448 for the new splits.

	correct analysis (all levels)	no analysis	(partially) flat analysis	wrong analysis
DBs	≈ 55.99%			
DBs + Context + Corpus Look-up	87.77%	7.45%	3.48%	1.29%
+ Recheck	92.44%	2.68%	2.88%	1.99%
+ Recheck + Weighting	93.34%	2.58%	2.78%	1.29%

(5) a. adjective vs. participle: *folgend* 'following', *gewandt* 'turned$_v$, skillful$_{adj}$'
b. constituents not in context: *Metallkäfig* 'metal cage', *Tierärztin* 'vet$_{fem}$'
c. flat analyses: *Roll|vor|Gang* '♯?(to roll|prefix, before|gait), rolling procedure'
d. analysis from GermaNet: ♯?*(Land|Nahme)* '(land|"take"), settlement'
e. frequent homograph:
♯*(Parlament|(*arisch* ar|isch))* '(parliament|(*Aryan* Ar|ian), parliamentary'
f. correction by word-length weighting:
*rollen|(*Vorgang* (*vorgehen* vor|gehen))* 'to roll|(*procedure* (*to proceed* pro|ceed))'

5,696 new entries for monomorphemic lexemes; 8,448 for the new splits.

	correct analysis (all levels)	no analysis	(partially) flat analysis	wrong analysis
DBs	≈ 55.99%			
DBs + Context + Corpus Look-up	87.77%	7.45%	3.48%	1.29%
+ Recheck	92.44%	2.68%	2.88%	1.99%
+ Recheck + Weighting	93.34%	2.58%	2.78%	1.29%

(5) a. adjective vs. participle: *folgend* 'following', *gewandt* 'turned$_v$, skillful$_{adj}$'
b. constituents not in context: *Metallkäfig* 'metal cage', *Tierärztin* 'vet$_{fem}$'
c. flat analyses: *Roll|vor|Gang* '♯?(to roll|prefix, before|gait), rolling procedure'
d. analysis from GermaNet: ♯?*(Land|Nahme)* '(land|"take"), settlement'
e. frequent homograph:
♯*(Parlament|(*arisch* ar|isch))* '(parliament|(*Aryan* Ar|ian), parliamentary'
f. correction by word-length weighting:
*rollen|(*Vorgang* (*vorgehen* vor|gehen))* 'to roll|(*procedure* (*to proceed* pro|ceed))'

5,696 new entries for monomorphemic lexemes; 8,448 for the new splits.

Outline

An Hybrid Approach for Deep-Level Morphological Analysis

- Starting points: a. morphological trees database b. flat structures from a morphological segmentation tool.
- All plausible combinations of the immediate constituents were evaluated by look-ups in textual environments of a large corpus or inside the set of all types as a back-off strategy.
- Biases towards small constituents with high frequencies on the one side and unsplit words on the other were tackled by insights from investigations in quantitative linguistics.
- The combination of the methods lead to an accuracy of 93% for complex structures and 98.7% for acceptable output.

An Hybrid Approach for Deep-Level Morphological Analysis

- Starting points: a. morphological trees database b. flat structures from a morphological segmentation tool.
- All plausible combinations of the immediate constituents were evaluated by look-ups in textual environments of a large corpus or inside the set of all types as a back-off strategy.
- Biases towards small constituents with high frequencies on the one side and unsplit words on the other were tackled by insights from investigations in quantitative linguistics.
- The combination of the methods lead to an accuracy of 93% for complex structures and 98.7% for acceptable output.

An Hybrid Approach for Deep-Level Morphological Analysis

- Starting points: a. morphological trees database b. flat structures from a morphological segmentation tool.
- All plausible combinations of the immediate constituents were evaluated by look-ups in textual environments of a large corpus or inside the set of all types as a back-off strategy.
- Biases towards small constituents with high frequencies on the one side and unsplit words on the other were tackled by insights from investigations in quantitative linguistics.
- The combination of the methods lead to an accuracy of 93% for complex structures and 98.7% for acceptable output.

An Hybrid Approach for Deep-Level Morphological Analysis

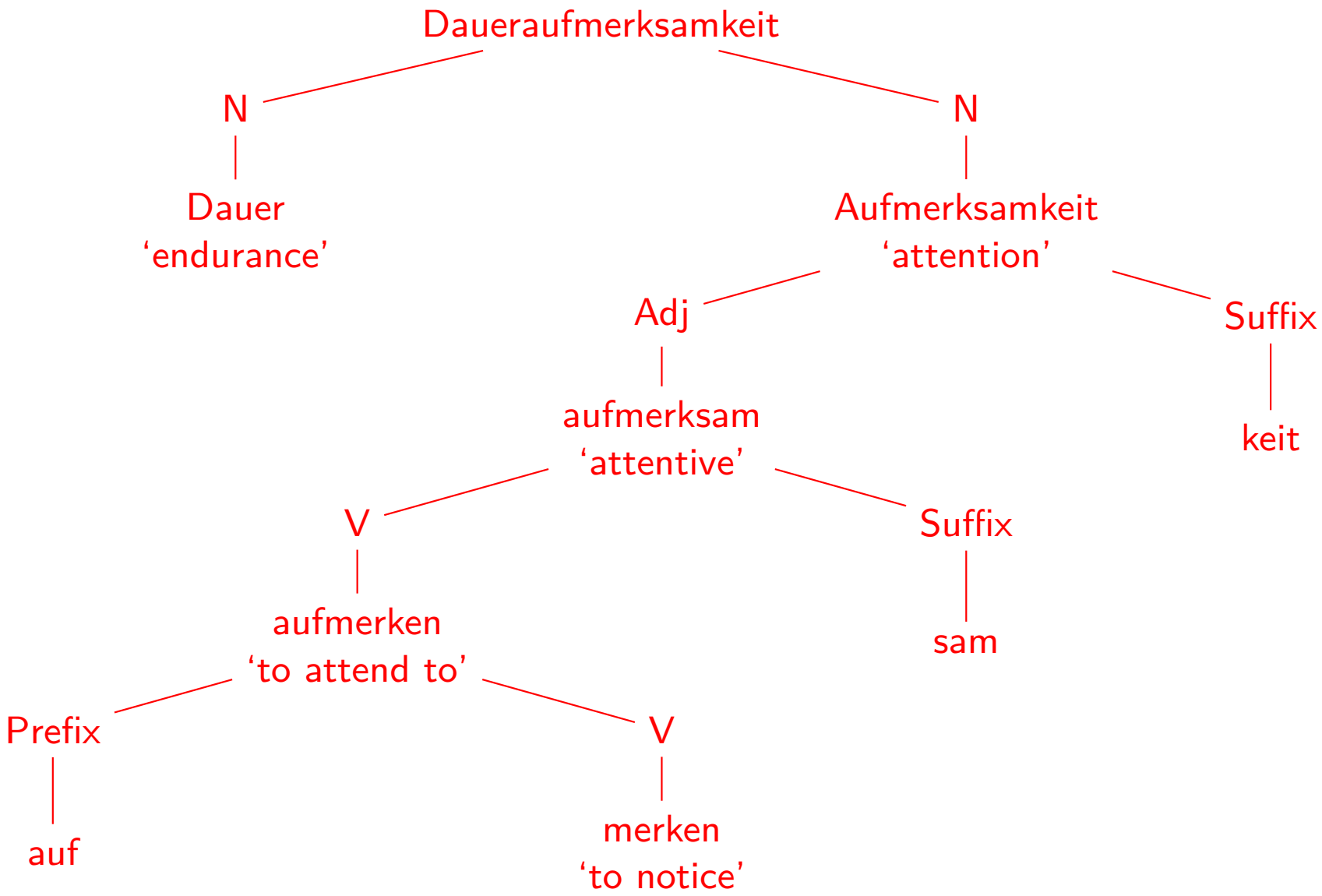
- Starting points: a. morphological trees database b. flat structures from a morphological segmentation tool.
- All plausible combinations of the immediate constituents were evaluated by look-ups in textual environments of a large corpus or inside the set of all types as a back-off strategy.
- Biases towards small constituents with high frequencies on the one side and unsplit words on the other were tackled by insights from investigations in quantitative linguistics.
- The combination of the methods lead to an accuracy of 93% for complex structures and 98.7% for acceptable output.

Future Work

For improvement, there are two directions:

- using larger corpora, to possibly obtain a better fit of the wordlength-frequency relationship.
- On the other hand, inhomogeneous data can blur models. Therefore, analyzing words text by text could help to achieve larger contextual dependency and to find morphological structures fitting to the direct environment. This would result in different structures for orthographical words according to their contexts.

Thank you for your ‘permanent|attention’



References I

Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. The CELEX lexical database (CD-ROM).

Stefanie Dipper. 2016. Tokenizer for German. https://www.linguistics.rub.de/~dipper/resources/tokenizer.html.

Rainer Gerlach. 1982. Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. In W. Lehfeldt and U. Strauss, editors, *Glottometrika 4*, Brockmeyer, Quantitative Linguistics 14, pages 95–102.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. pages 9–15. http://www.aclweb.org/anthology/W97-0802.

Verena Henrich and Erhard Hinrichs. 2011. Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, 2011*. Association for Computational Linguistics, pages 420–426. http://www.aclweb.org/anthology/R11-1058.

Reinhard Köhler. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Quantitative Linguistics 31. Studienverlag Dr. N. Brockmeyer, Bochum.

Andrea Krott. 1996. Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics* 3(1):29–37. https://doi.org/10.1080/09296179608590061.

References II

Andrea Krott. 2004. Ein funktionalanalytisches Modell der Wortbildung [A functional analytical model of word formation]. In Reinhard Köhler, editor, *Korpuslinguistische Untersuchungen zur Quantitativen und Systemtheoretischen Linguistik [Corpus-linguistic Investigations of Quantitative and System-theoretical Linguistics]*, Elektronische Hochschulschriften an der Universität Trier, Trier, pages 75–126. http://ubt.opus.hbz-nrw.de/volltexte/2004/279/pdf/04_krott.pdf.

Eliza Margaretha and Harald Lüngen. 2014. Building linguistic corpora from wikipedia articles and discussions. *Journal of Language Technology and Computational Linguistics. Special issue on building and annotating corpora of computer-mediated communication. Issues and challenges at the interface between computational and corpus linguistics* 29(2):59 – 82. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-33306, http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf.

Paul Menzerath. 1954. *Die Architektonik des deutschen Wortschatzes*. Phonetische Studien. Dümmler, Bonn ; Hannover ; Stuttgart.

Helmut Schmid. 1999. Improvements in Part-of-Speech Tagging with an Application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, Springer Netherlands, Dordrecht, pages 13–25. https://doi.org/10.1007/978-94-017-2390-9_2.

References III

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association (ELRA). http://www.aclweb.org/anthology/L04-1275.

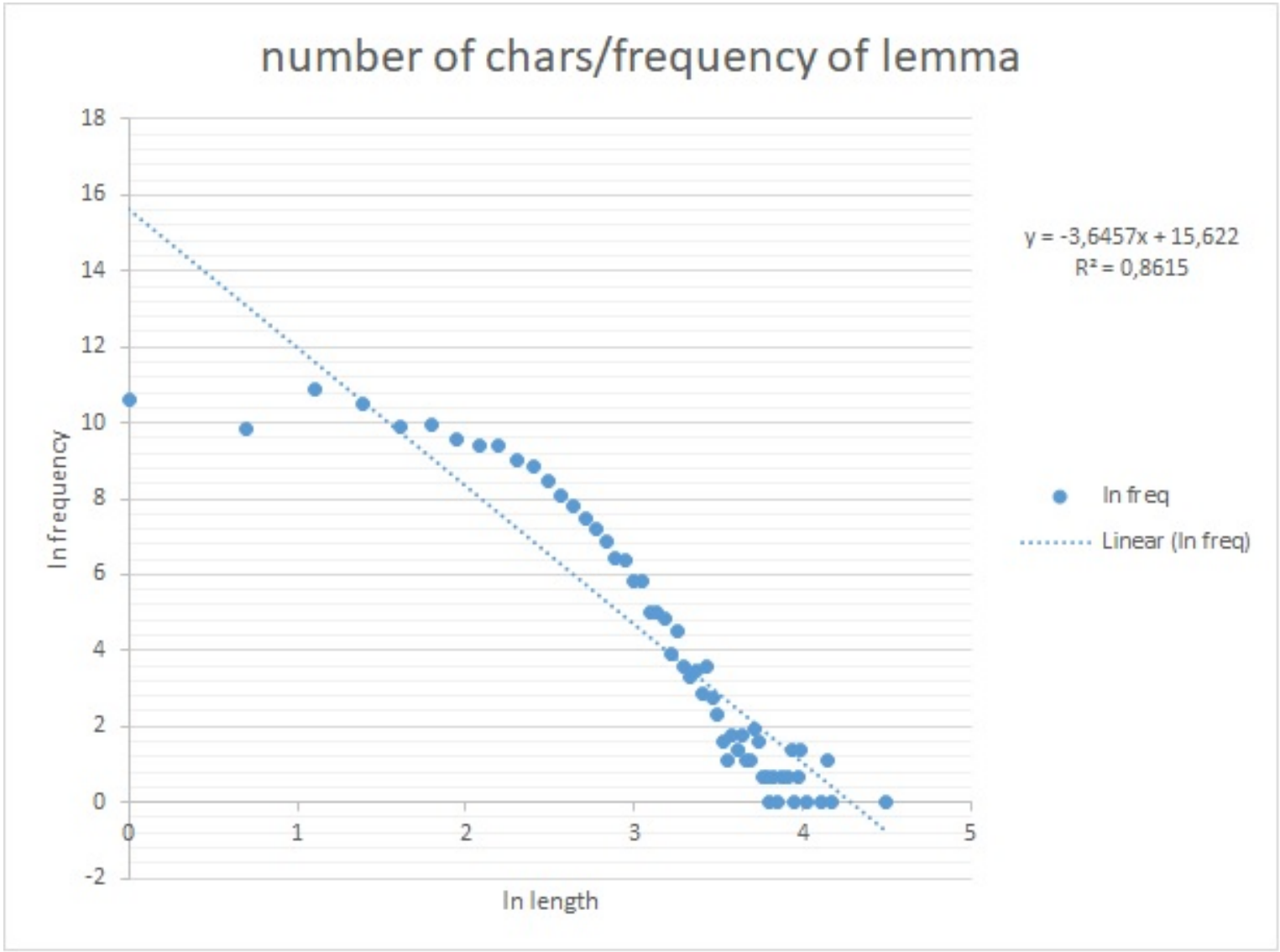
Petra Steiner. 2017. Merging the Trees - Building a Morphological Treebank for German from Two Resources. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, January 23-24, 2018, Prague, Czech Republic*. pages 146–160. https://aclweb.org/anthology/W17-7619.

Petra Steiner and Reinhard Rapp. in press. Building and Exploiting Lexical Databases for Morphological Parsing. In *Proceedings of The International Conference on Contemporary Issues in Data Science, March 5-8, 2019, Zanjan, Iran*. Springer, Lecture Notes in Computer Science.

Petra Steiner and Josef Ruppenhofer. 2018. Building a Morphological Treebank for German from a Linguistic Database. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1613.

More slides ....

Figure 4: length/frequeny of lemmas from MLD corpus

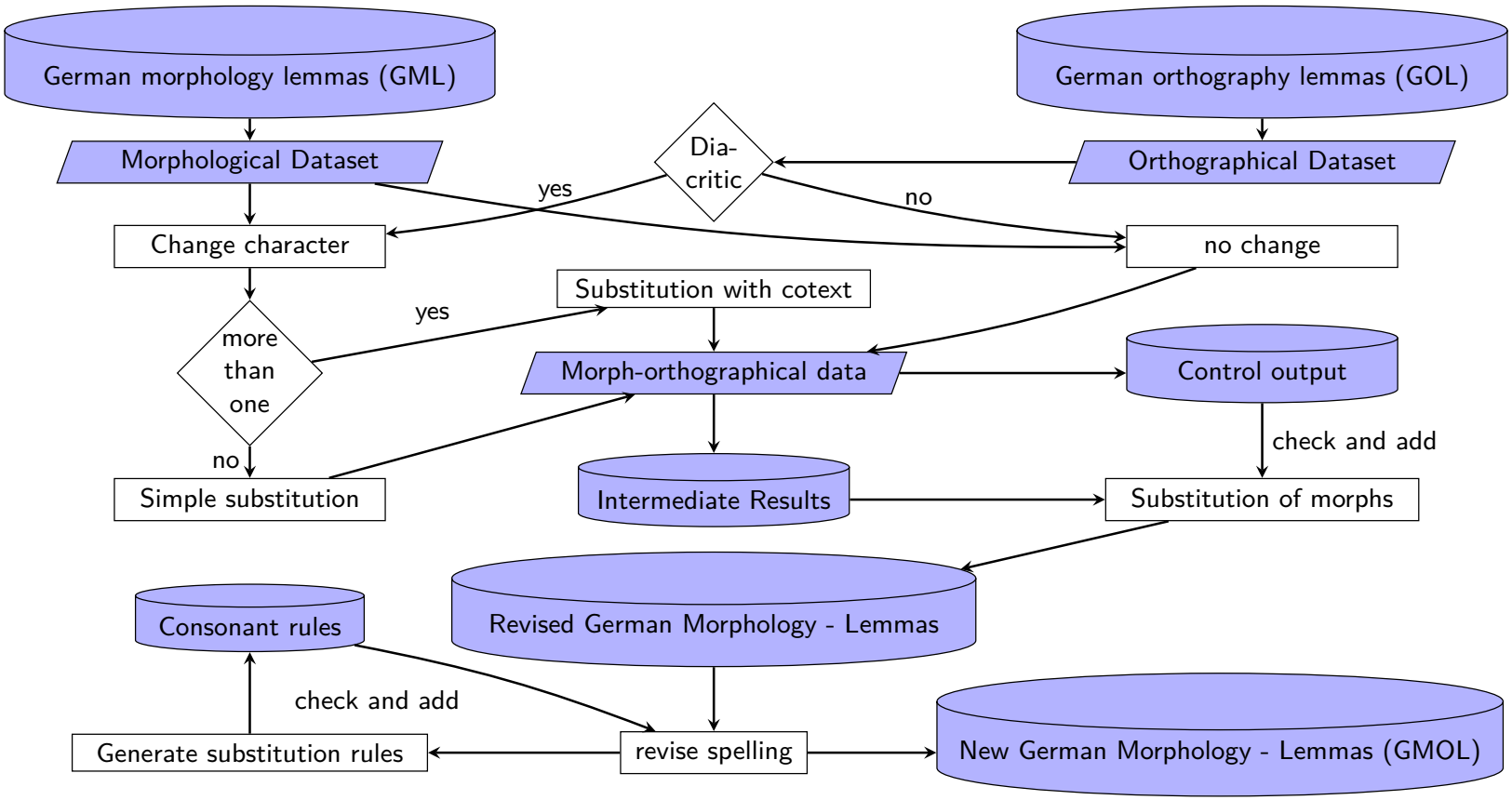


Reliable Morphological Data

- CELEX: standard resource for German lexical data
  - 51,728 entries
  - 38,650 derivatives or compounds, 2,402 conversions
  - core vocabulary
  - outdated format
  - outdated spelling
- GermaNet:
  - rich vocabulary, complex lexemes
  - segmentation restricted to nominal compounds
  - approx. 68,000 entries of compounds

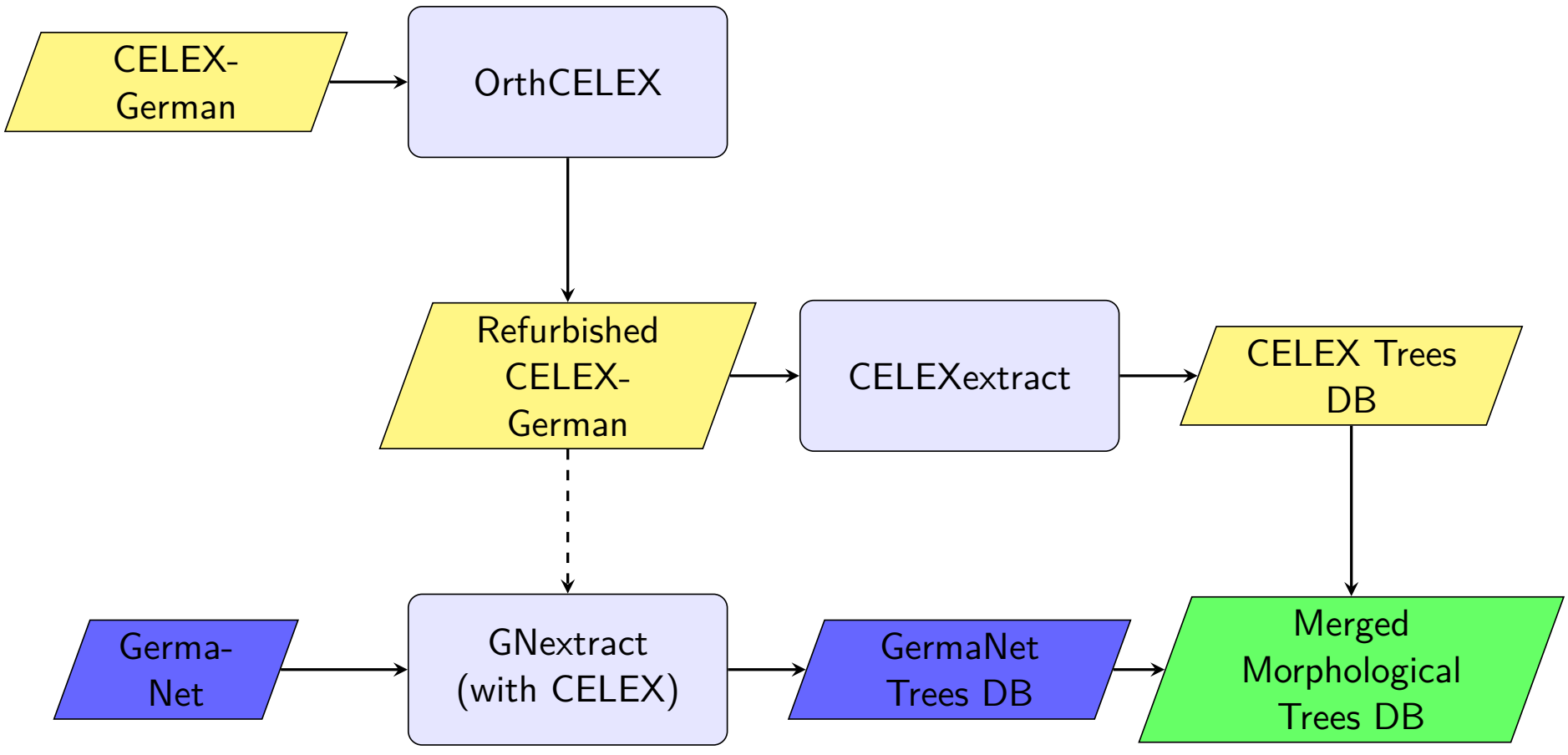
Transfer of CELEX to the Modern Standard

- outdated format, e.g. *Abschlu$* (orthographical dataset) *Abschluss* (morphological dataset) $\longmapsto$ *Abschluß* 'conclusion'
- outdated spelling, e.g. *Abschluß* 'conclusion' $\longmapsto$ *Abschluss*

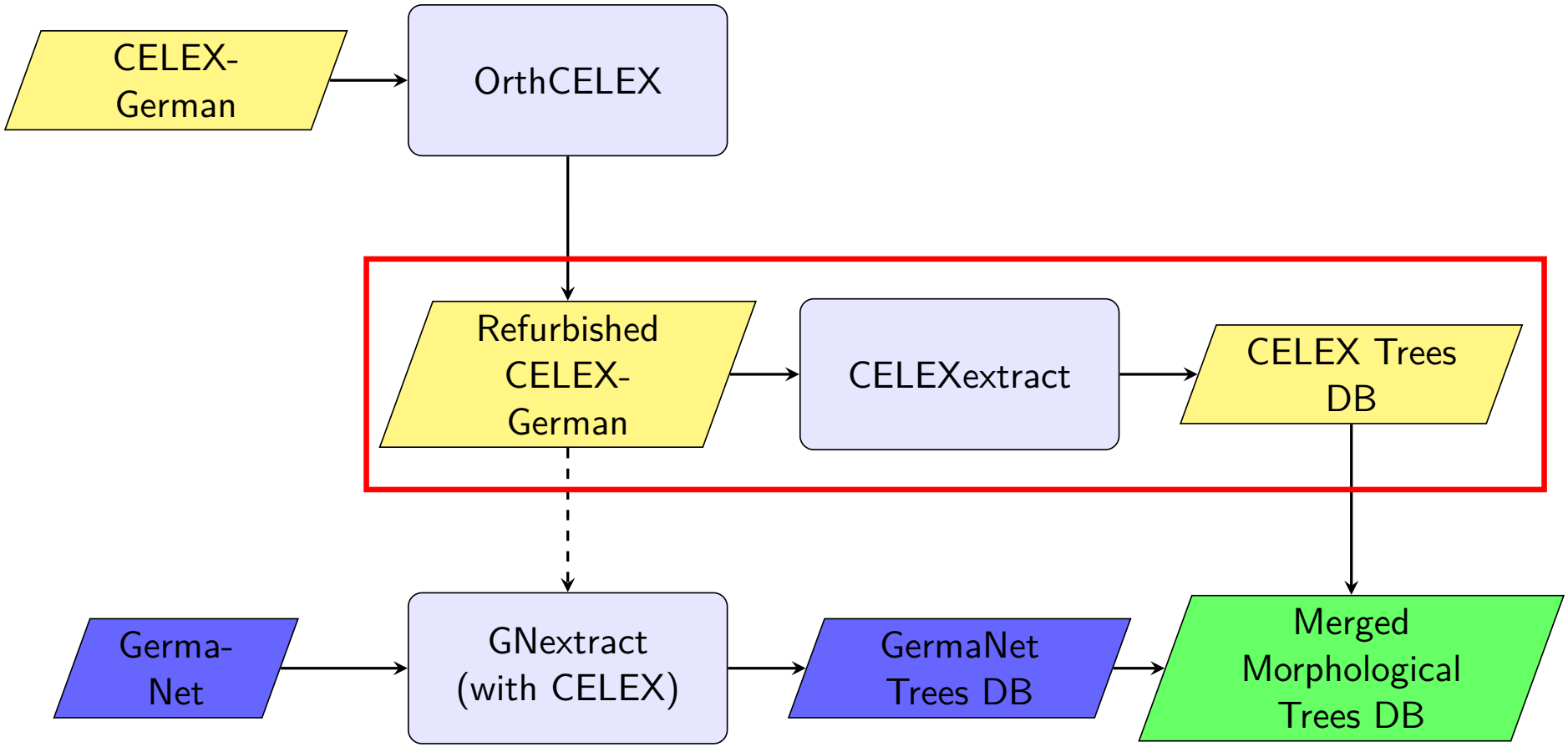


see ?

Overview of the Data Processing



Overview of the Data Processing



CELEX Trees I

Examples

97\Abdrift\0\C\1\Y\Y\Y\ab+drift\xV\N\N\N\
((ab)[N|.V],((treib)[V])[V])[N]\Y\N\N\N\S3/P3\N
'leeway - away|to float'

207\Abgangszeugnis\4\C\1\Y\Y\Y\Abgang+s+Zeugnis\NxN\N\N\N\
((((ab)[V|.V],(geh)[V])[V])[N],(s)[N|N.N],((zeug)[V],(nis)[N|V.])[N]) [...]
'leaving certificate - leave|certificate'

605\Abschlussprüfung\C\1\Y\Y\Y\Abschluss+Prüfung\NN\N\N\N\
((((ab)[V|.V],(schließ)[V])[V])[N],((prüf)[V],(ung)[N|V.])[N]\[...]
'final exam - conclusion|exam'

CELEX Trees II

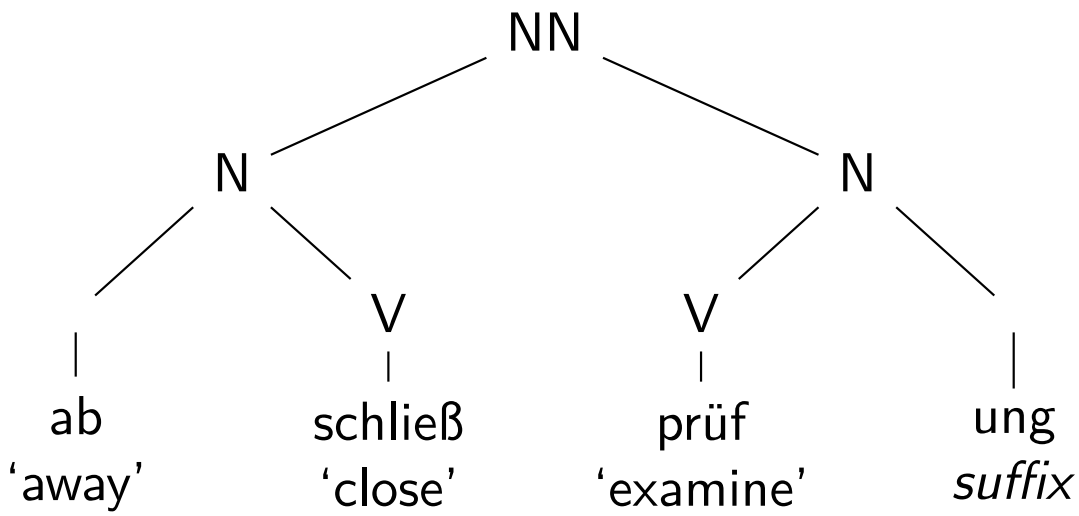


Figure 5: Morphological analysis of *Abschlussprüfung* 'final exam'

CELEX Trees III

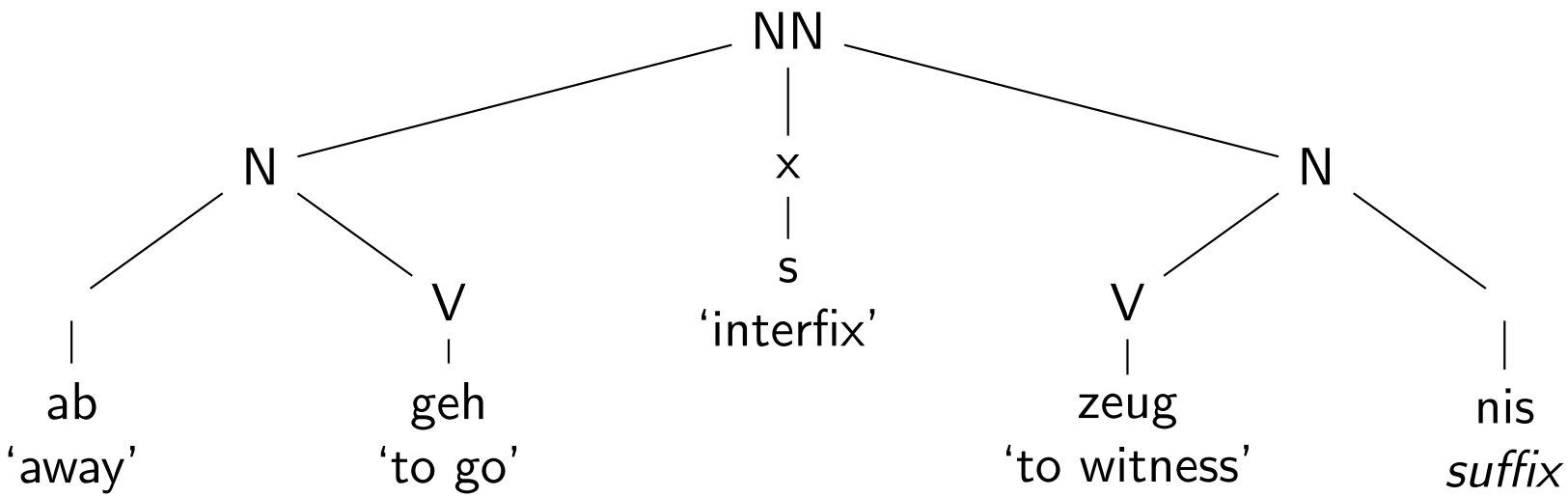
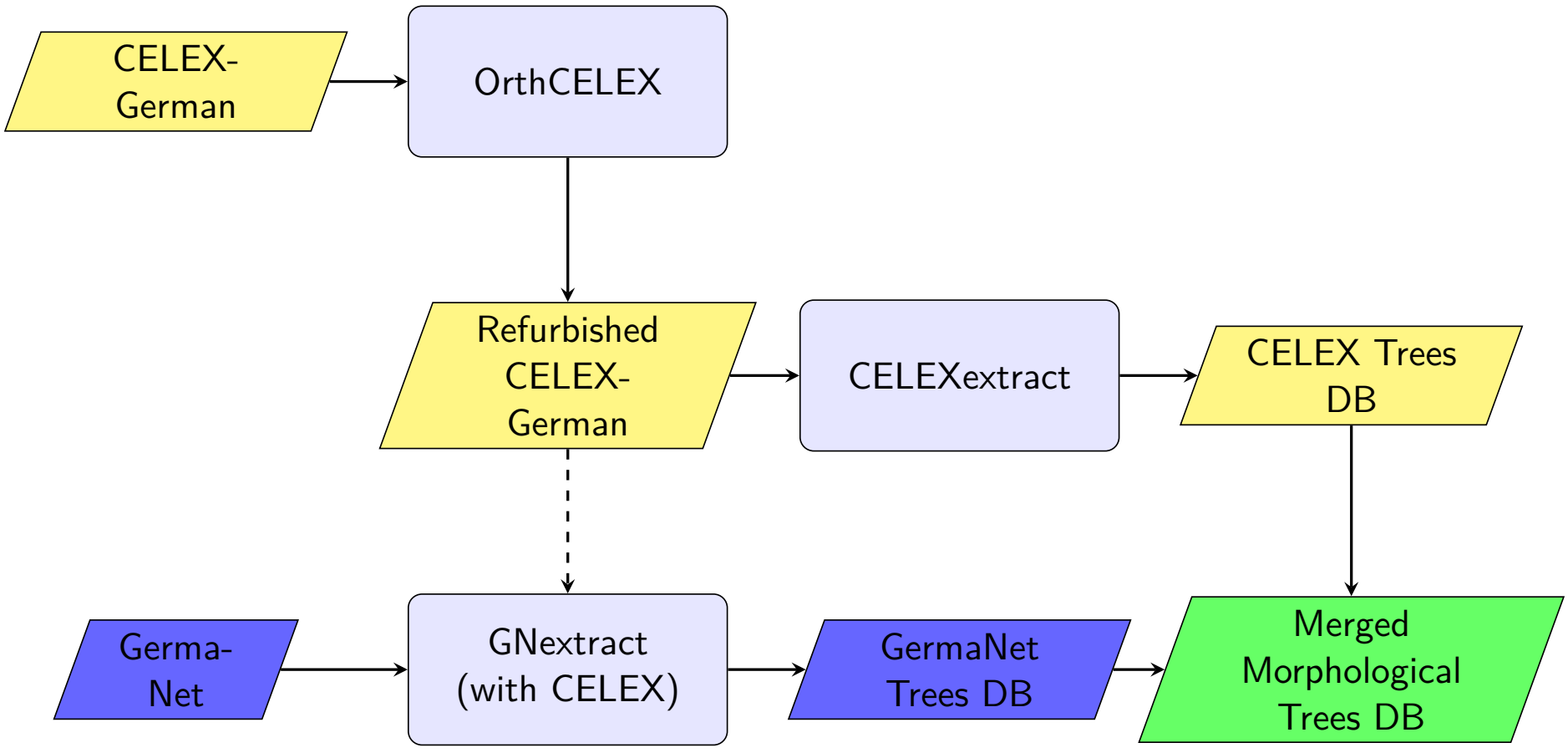
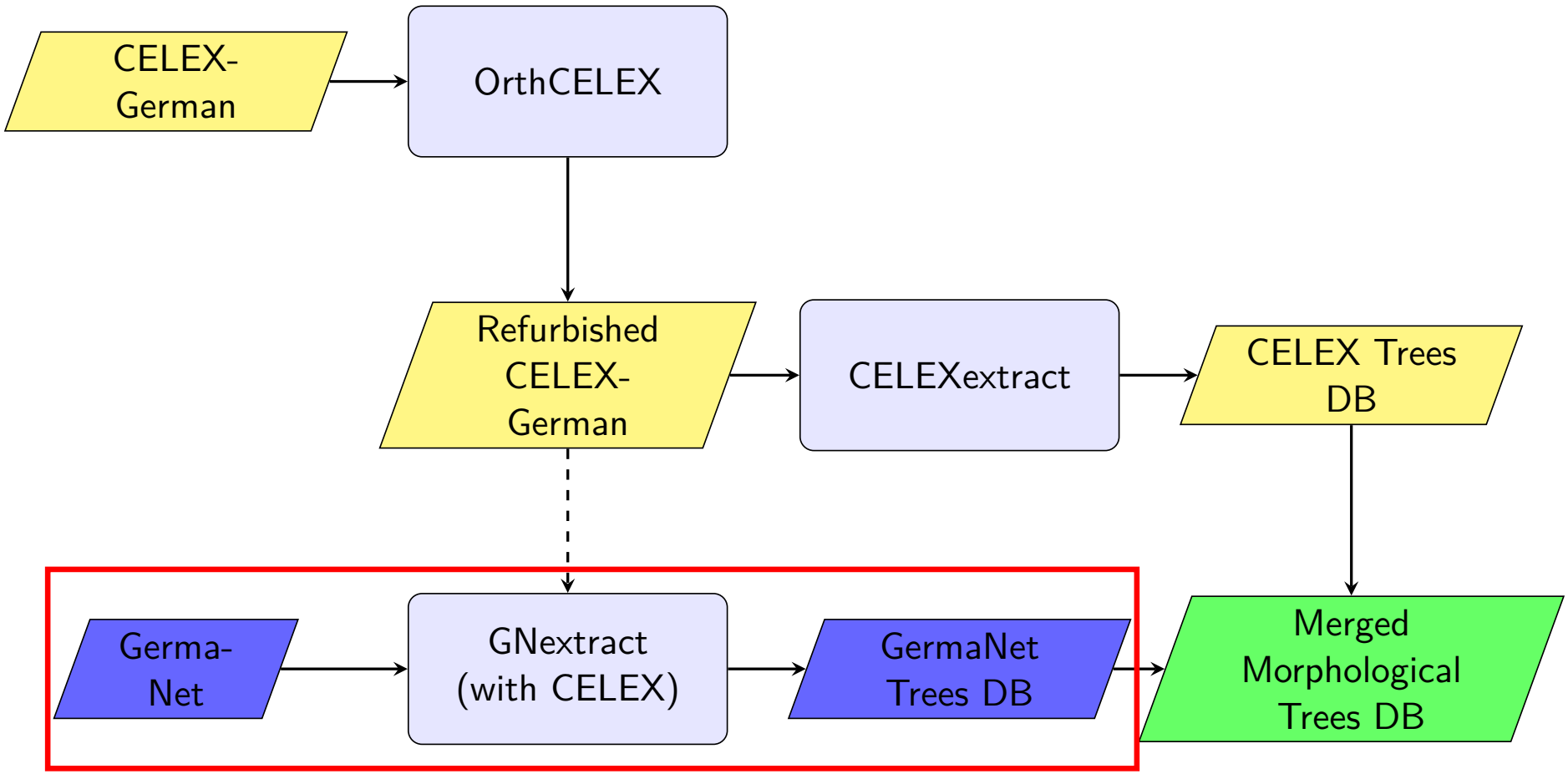


Figure 6: Morphological analysis of *Abgangszeugnis* 'leaving certificate'

Overview of the Data Processing



Overview of the Data Processing



GermaNet

- Lexical-semantic database, hierarchically structured in synsets
- same approach as WordNet

(Hamp and Feldweg, 1997)

<synset id="s5552" category="nomen" class="Artefakt"> <lexUnit id="l8355" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no"> <orthForm>Werkstück</orthForm> **<compound> <modifier category="Nomen">Werk</modifier> <modifier category="Verb">**
werken</modifier> <head>Stück</head> </compound> </lexUnit>
</synset>

GermaNet

- Lexical-semantic database, hierarchically structured in synsets
- same approach as WordNet

(Hamp and Feldweg, 1997)

<synset id="s5552" category="nomen" class="Artefakt"> <lexUnit id="l8355" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no"> <orthForm>Werkstück</orthForm> **<compound> <modifier category="Nomen">Werk</modifier> <modifier category="Verb"> werken</modifier> <head>Stück</head> </compound>** </lexUnit> </synset>

GN Trees I: Compounds of GermaNet

- GermaNet compounds (Henrich and Hinrichs, 2011), version 11 with 66,059 compounds of which some have ambiguous structures
- remove proper names, foreign word expressions (*After-Show-Party*, *Bodenseeregion* 'Lake of Constance region')
- remove deficient entries, e.g. with missing parts-of-speech classes or affixoids
- add interfixes (Fugen/filler letters) by heuristics
  Abfahrtszeit 'departure time' GermaNet: Abfahrt|zeit
  $\longmapsto$ Abfahrt|s|zeit

GN Trees II: Compound structures from GermaNet

1. generate flat compound entries
   - *Beitragssatz* : Beitrag|s|Satz ‘contribution rate’
   - *Beitragssatzsicherung* : Beitragssatz|Sicherung ‘contribution rate safeguarding’
   - *Beitragssatzsicherungsgesetz* : Beitragssatzsicherung|s|Gesetz ‘contribution rate safeguarding law’
2. infer GN complex structure by recursive look-up

   Beitragssatz

   Beitragssatz|Sicherung

   Beitragssatzsicherung|s|Gesetz

GN Trees II: Compound structures from GermaNet

1. generate flat compound entries
   - *Beitragssatz* : Beitrag|s|Satz ‘contribution rate’
   - *Beitragssatzsicherung* : Beitragssatz|Sicherung ‘contribution rate safeguarding’
   - *Beitragssatzsicherungsgesetz* : Beitragssatzsicherung|s|Gesetz ‘contribution rate safeguarding law’
2. infer GN complex structure by recursive look-up

   Beitragssatz — Beitrag|s|Satz

   Beitragssatz|Sicherung

   Beitragssatzsicherung|s|Gesetz

GN Trees II: Compound structures from GermaNet

1. generate flat compound entries
   - *Beitragssatz* : Beitrag|s|Satz ‘contribution rate’
   - *Beitragssatzsicherung* : Beitragssatz|Sicherung ‘contribution rate safeguarding’
   - *Beitragssatzsicherungsgesetz* : Beitragssatzsicherung|s|Gesetz ‘contribution rate safeguarding law’
2. infer GN complex structure by recursive look-up

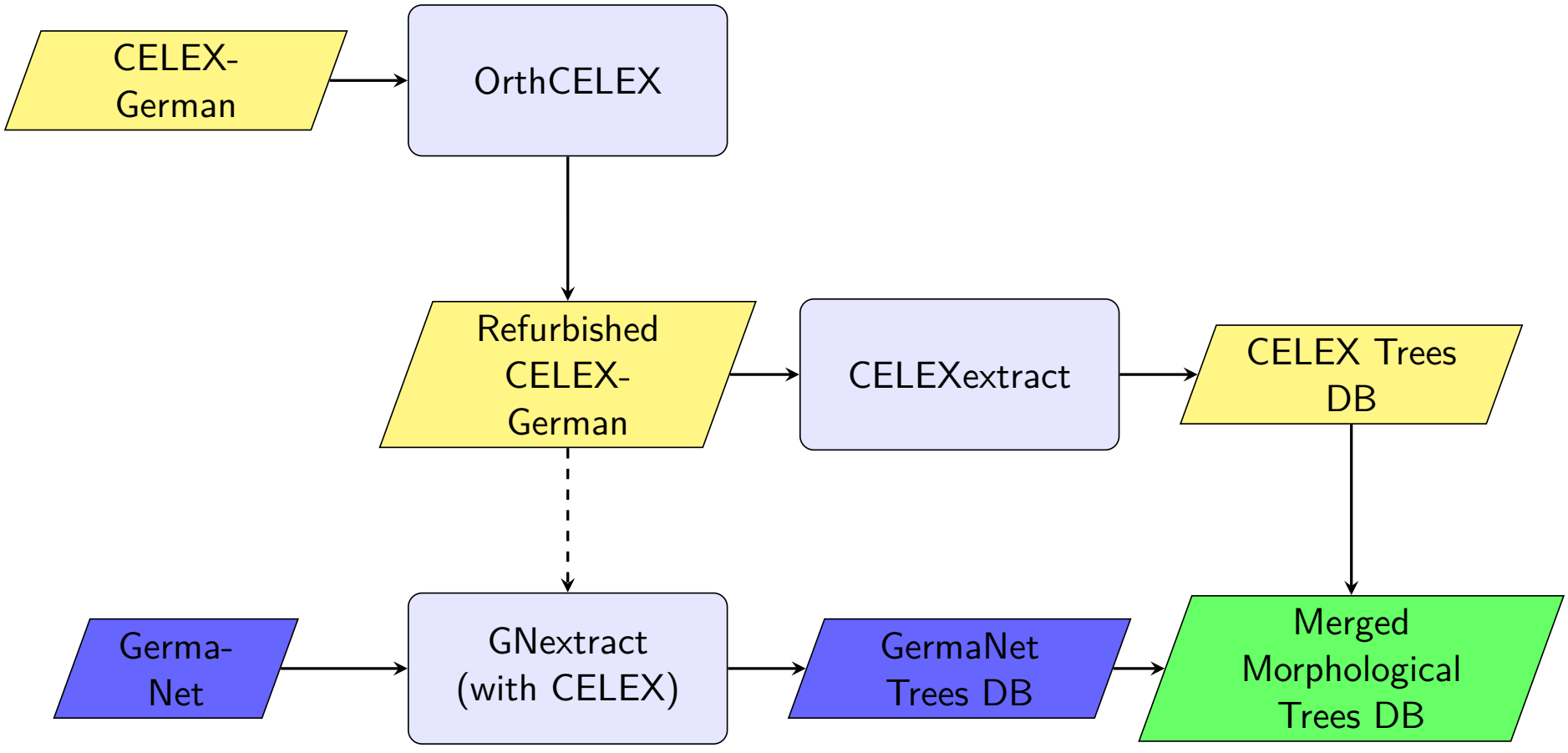
Beitragssatz	*insert*	Beitrag|s|Satz
Beitragssatz|Sicherung		(Beitrag|s|Satz)|Sicherung
Beitragssatzsicherung|s|Gesetz		

GN Trees II: Compound structures from GermaNet

1. generate flat compound entries
   - *Beitragssatz* : Beitrag|s|Satz ‘contribution rate’
   - *Beitragssatzsicherung* : Beitragssatz|Sicherung
     ‘contribution rate safeguarding’
   - *Beitragssatzsicherungsgesetz* : Beitragssatzsicherung|s|Gesetz
     ‘contribution rate safeguarding law’
2. infer GN complex structure by recursive look-up

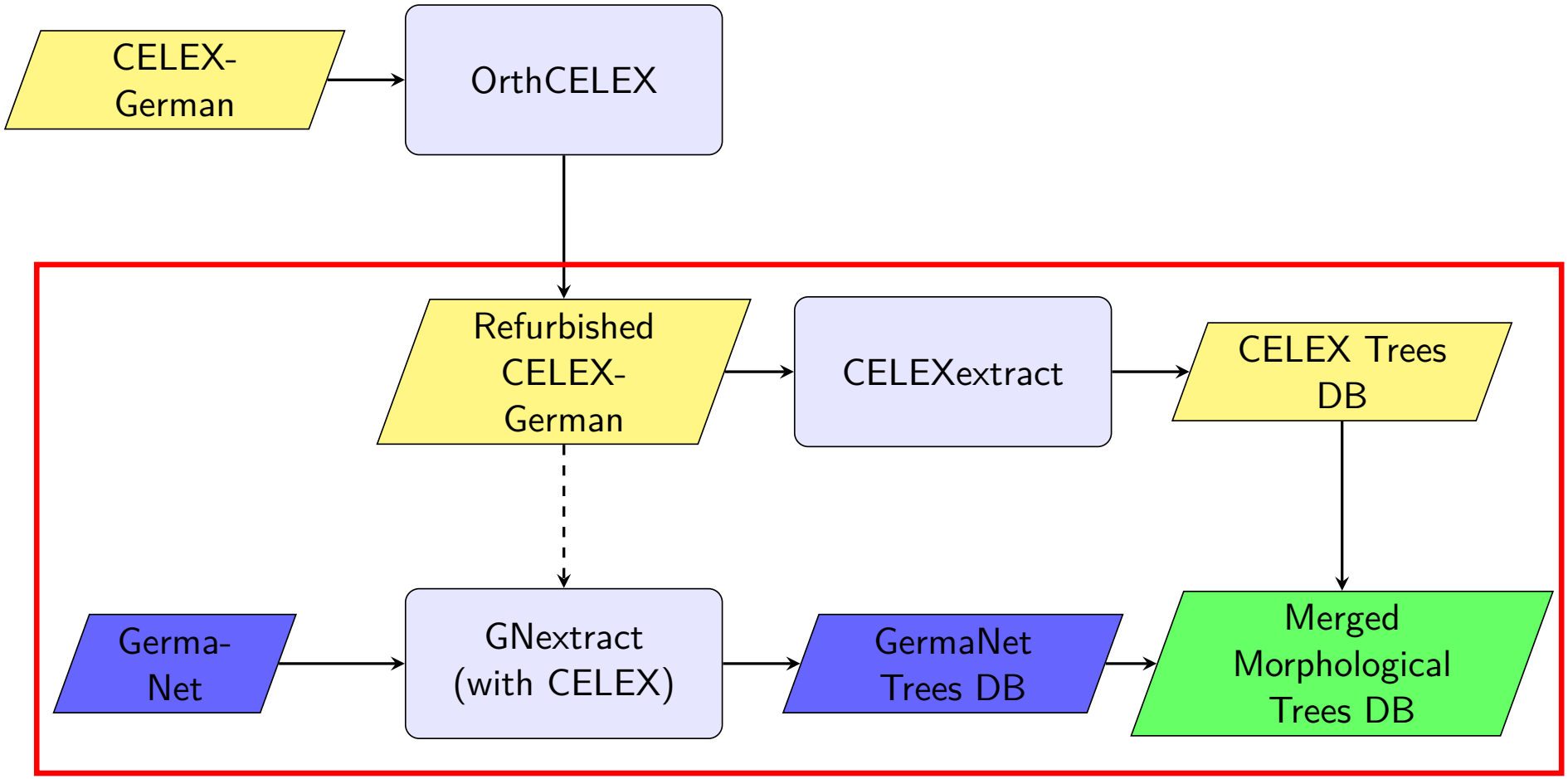
Beitragssatz	*insert*	Beitrag|s|Satz
Beitragssatz|Sicherung	*insert*	(Beitrag|s|Satz)|Sicherung
Beitragssatzsicherung|s|Gesetz		((Beitrag|s|Satz)|Sicherung)|s|Gesetz

Overview of the Data Processing



(Steiner, 2017)

Overview of the Data Processing



(Steiner, 2017)

Combining GN Trees and CELEX Trees

- GermaNet compound structures
- immediate constituents and other information from CELEX

Beitrag 'contribution'	(*beitragen$_V$* bei$_x$|tragen$_V$)
Sicherung 'safeguarding'	(*sichern$_V$* sicher$_A$|n$_x$)|ung$_x$
Gesetz 'law'	ge$_x$|setzen$_V$

- infer complex (derivative) structures by recursive look-up

Beitragssatz	Beitrag|s|Satz
Beitragssatzsicherung	(Beitrag|s|Satz)|(sichern|ung)
Beitragssatzsicherungsgesetz	((Beitrag|s|Satz)|(sichern|ung))|s|(ge|setzen)

Some mistakes/questionable or missing analyses:

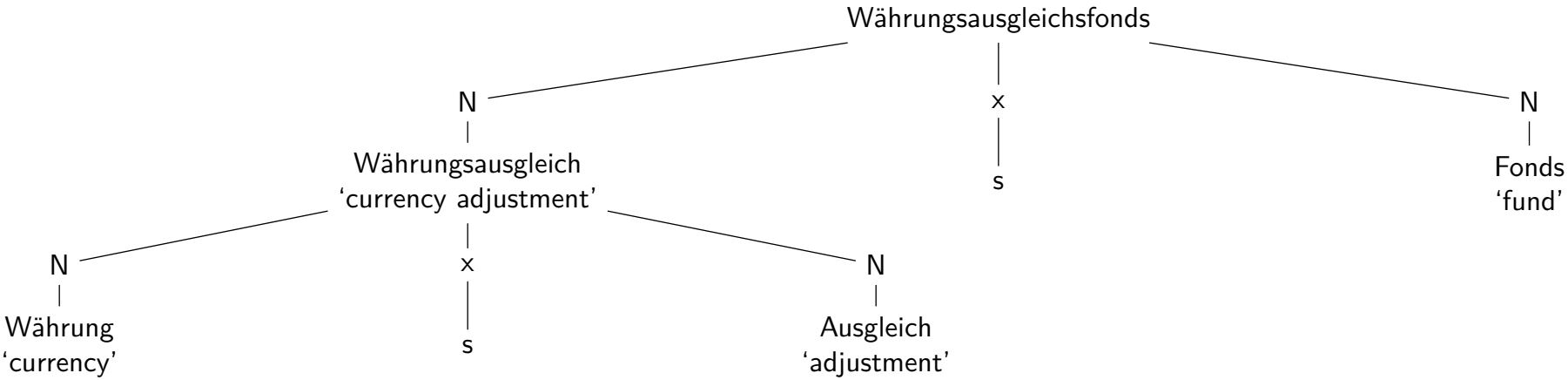
- approx. 2000 missing segmentations
- *Restrukturierungsmaßnahmen*: Restrukturierung|s|(Maß|Nahme)

Formats of Output

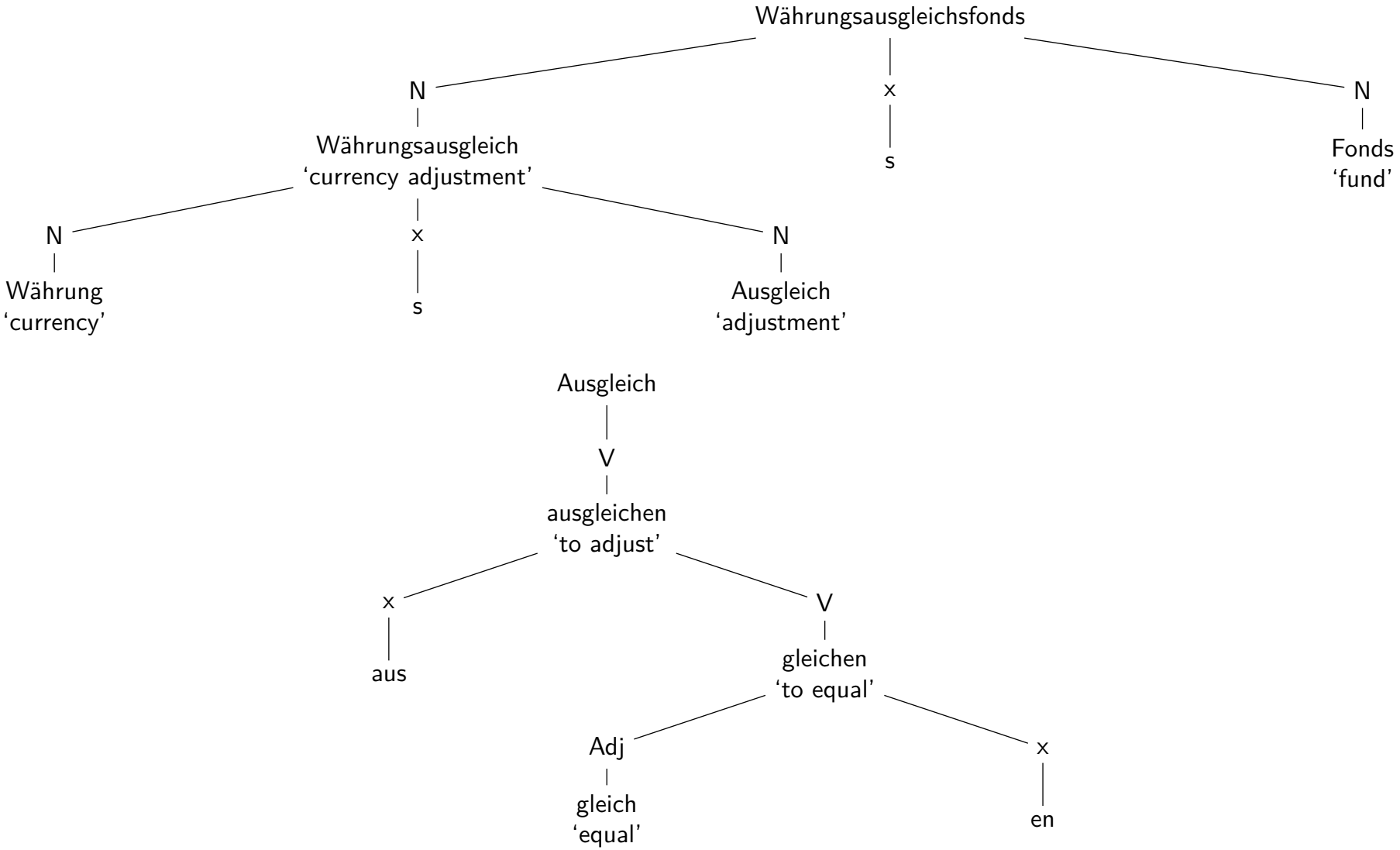
Optional parameters:

- Depth of analysis for compounds
- Parts of speech for the constructs and/or the smallest constituents
- Choice of the output format (parentheses or a notation with | for the splits on the same level)
- Addition of filler letters for GN
- Transfering the GN annotation scheme to CELEX scheme
- Removing compounds with proper names and/or foreign words as constituents for GN
- Analysis of conversions for CELEX
- Depth of analysis for conversions for CELEX
- Dissimilarity measure for CELEX diachronic analyses

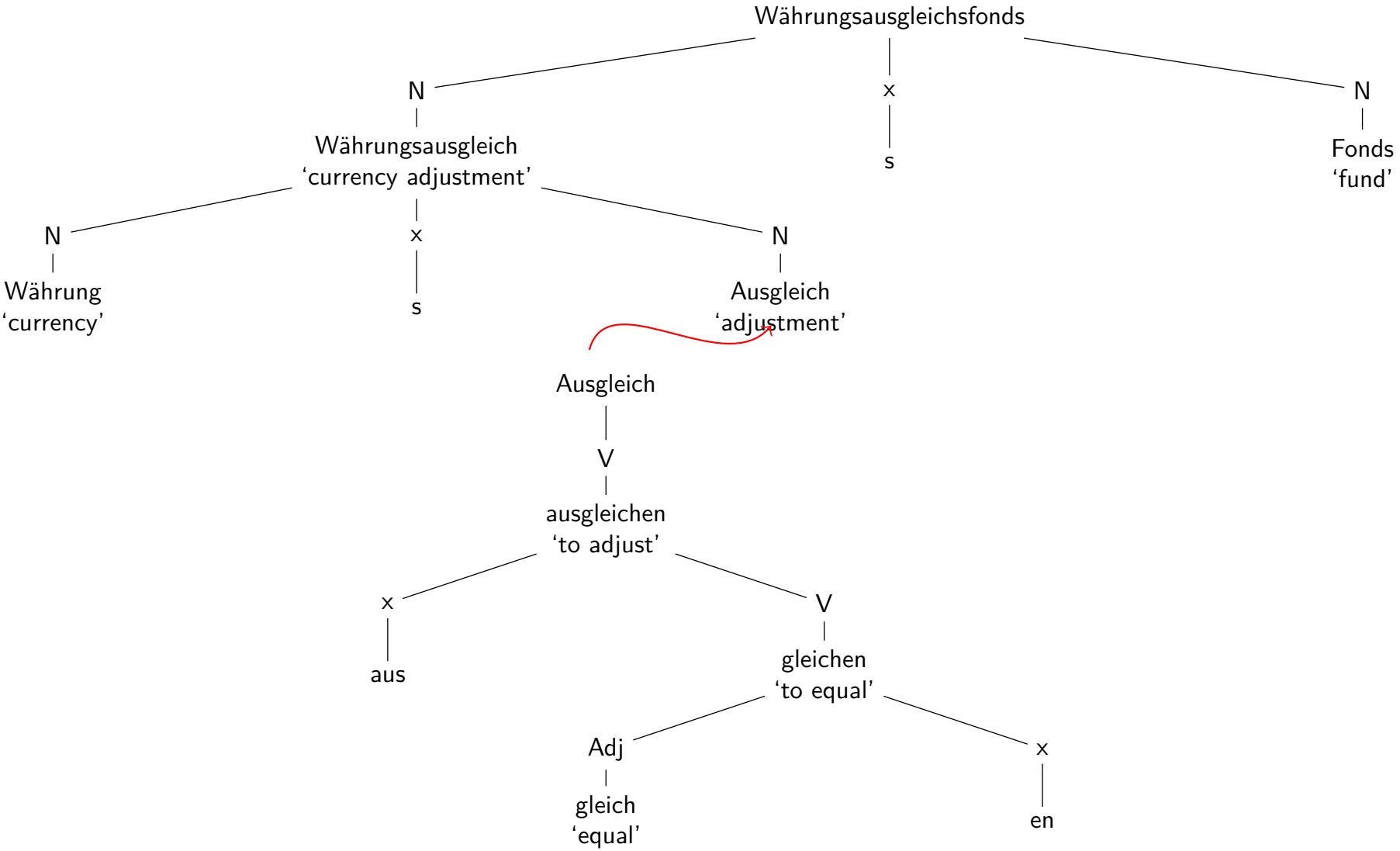
Example


Währungsausgleichsfonds
N
Währungsausgleich
‘currency adjustment’
x
N
Währung
‘currency’
x
s
N
Ausgleich
‘adjustment’
x
s
N
Fonds
‘fund’

Example


Währungsausgleichsfonds
N
x
N
Währungsausgleich
‘currency adjustment’
s
Fonds
‘fund’
N
x
N
Währung
‘currency’
s
Ausgleich
‘adjustment’
Ausgleich
V
ausgleichen
‘to adjust’
x
V
aus
gleichen
‘to equal’
Adj
x
gleich
‘equal’
en

Example


Währungsausgleichsfonds
N
x
N
Währungsausgleich
‘currency adjustment’
s
Fonds
‘fund’
N
x
N
Währung
‘currency’
s
Ausgleich
‘adjustment’
Ausgleich
V
ausgleichen
‘to adjust’
x
V
aus
gleichen
‘to equal’
Adj
x
gleich
‘equal’
en

Example

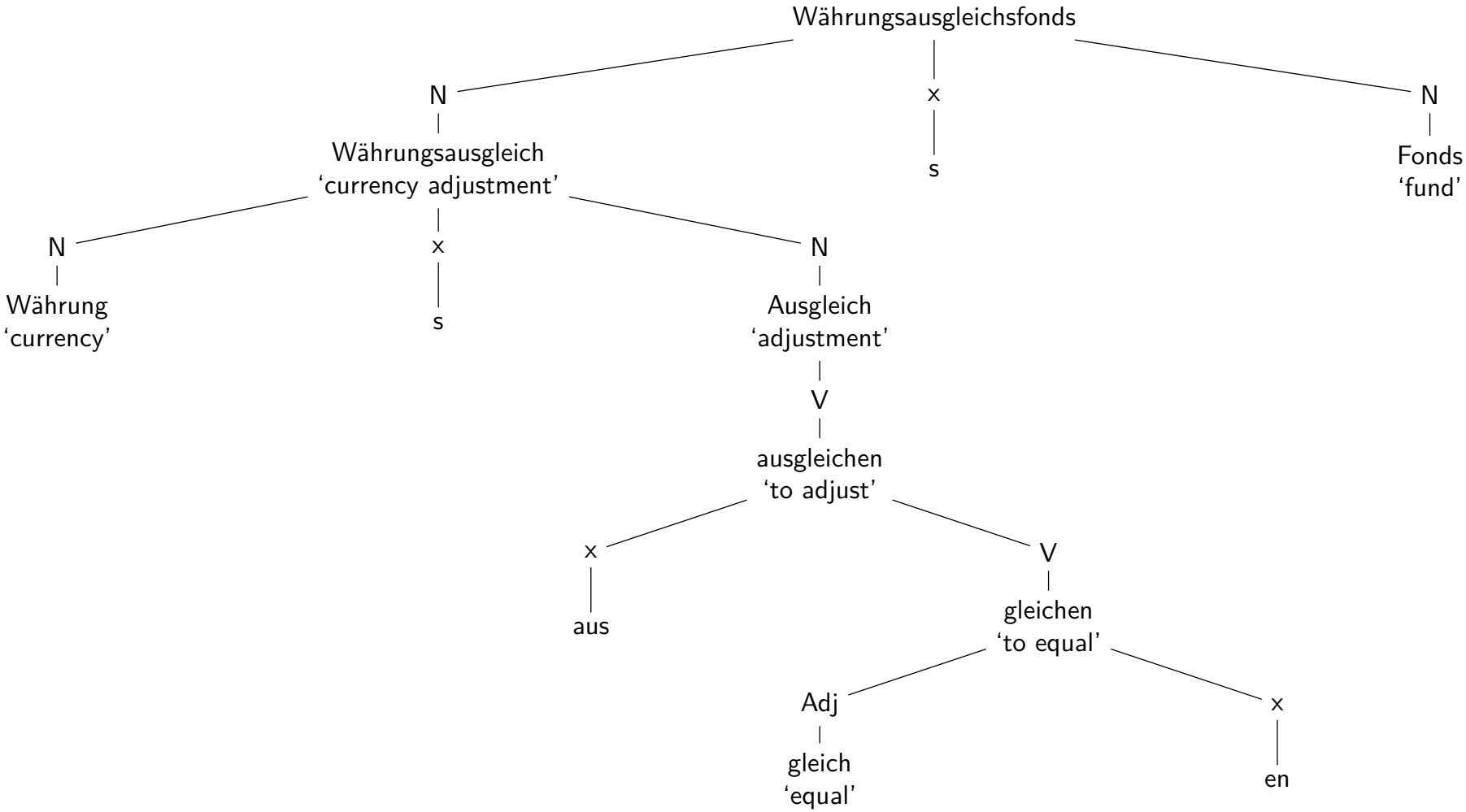


Figure 8: Merged morphological analysis of *Währungsausgleichsfonds* ‘currency adjustment fund’

Small Examples of List Representations

```
a. Abschlussprüfung (*Abschluss_N* (*abschließen_V*
ab_x|schließen_V))|(*Prüfung_N* prüfen_V|ung_x)
b. Abschlussprüfung (*Abschluss_N* (*abschließen_V* (ab_x)
(schließen_V)))(*Prüfung_N* (prüfen_V)(ung_x))
c. Abschlussprüfung Abschluss_N|Prüfung_N

a. Abdrift ab_x|(driften_V)
b. Abdrift (ab_x)(*driften_V* treiben_V)
c. Abdrift ab_x|driften_V

a. Abgangszeugnis (*Abgang_N* (*abgehen_V* ab_x|gehen_V))
|s_x|*Zeugnis_N* (zeugen_V|nis_x)
b. Abgangszeugnis (*Abgang_N* (*abgehen_V* (ab_x)(gehen_V)))
(s_x)(*Zeugnis_N* (zeugen_V)(nis_x))
c. Abgangszeugnis Abgang_N|s_x|Zeugnis_N
```

a: | notation, threshold 0.5; b: parenthesis notation and no restrictions on diachronic conversions; c: flat representation of the immediate constituent.

Merged GermanTrees

The parameters for the deep-level analyses are 6 for the levels of complex words and 2 for conversions. The Levenshtein dissimilarity threshold was set to 0.5. Double entries were removed.

Structures	**GN entries**	**CELEX entries**	**German Trees**
flat	67,452	40,097	100,095
deep-level	68,163	40,097	104,424
merged with CELEX	68,171	n/a	100,986
merged with CELEX plus simplex words	68,171	n/a	112,086

Table 1: Databases of German word trees

One Complex Database, Two Segmenters

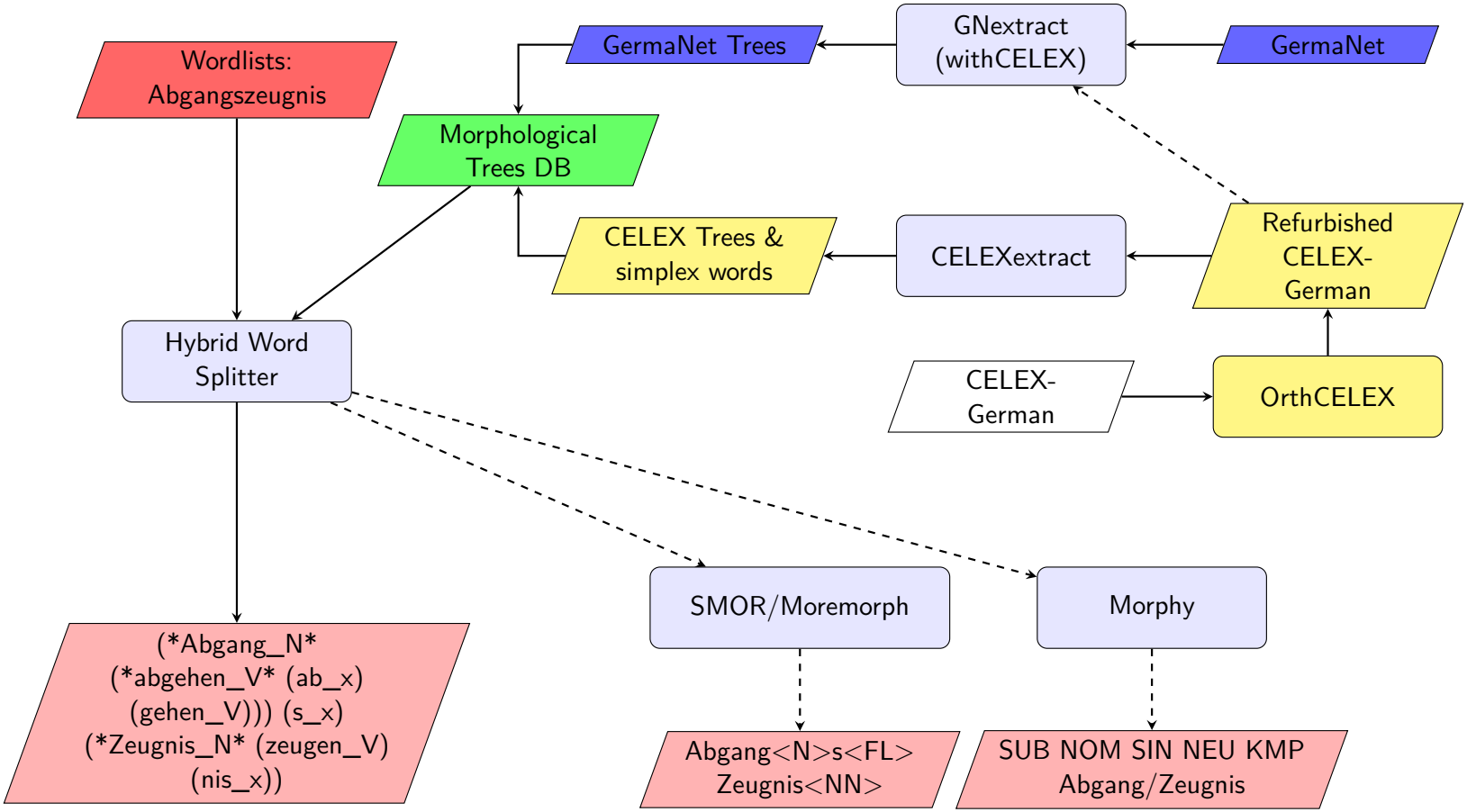


Figure 9: Morphological trees database and two different word segmenters as alternative methods for word splitting

Coverage of the Lemma Forms

- Corpus: *Korpus Magazin Lufthansa Bordbuch (MLD)*, part of the DeReKo-2016-I (Institut für Deutsche Sprache 2016) corpus (see Kupietz et al. 2010), an in-flight magazine with articles on traveling, consumption and aviation.
- Tokenization: enlarged and costumized tokenizer by Dipper (2016)
- 276 texts with 5,202 paragraphs, 16,046 sentences and 260,115 tokens

	lemma types	**recall**	**lemmas in text**	**recall**
corpus size	29,313		260,014	
MergedDB + simplex	14,446	49.29%	157,535	60.59%

Coverage of the Lemma Forms

- Corpus: *Korpus Magazin Lufthansa Bordbuch (MLD)*, part of the DeReKo-2016-I (Institut für Deutsche Sprache 2016) corpus (see Kupietz et al. 2010), an in-flight magazine with articles on traveling, consumption and aviation.
- Tokenization: enlarged and costumized tokenizer by Dipper (2016)
- 276 texts with 5,202 paragraphs, 16,046 sentences and 260,115 tokens

	lemma types	**recall**	**lemmas in text**	**recall**
corpus size	29,313		260,014	
MergedDB + simplex	14,446	49.29%	157,535	60.59%
+ Morphy	21,953	74.89%	241,117	92.73%

Coverage of the Lemma Forms

- Corpus: *Korpus Magazin Lufthansa Bordbuch (MLD)*, part of the DeReKo-2016-I (Institut für Deutsche Sprache 2016) corpus (see Kupietz et al. 2010), an in-flight magazine with articles on traveling, consumption and aviation.
- Tokenization: enlarged and costumized tokenizer by Dipper (2016)
- 276 texts with 5,202 paragraphs, 16,046 sentences and 260,115 tokens

	lemma types	**recall**	**lemmas in text**	**recall**
corpus size	29,313		260,014	
MergedDB + simplex	14,446	49.29%	157,535	60.59%
+ Morphy	21,953	74.89%	241,117	92.73%
+ Moremorphs	27,907	95.20%	256,903	98.80%

Table 2: Recall of Tree DBs (Steiner and Rapp, in press)

Conclusion

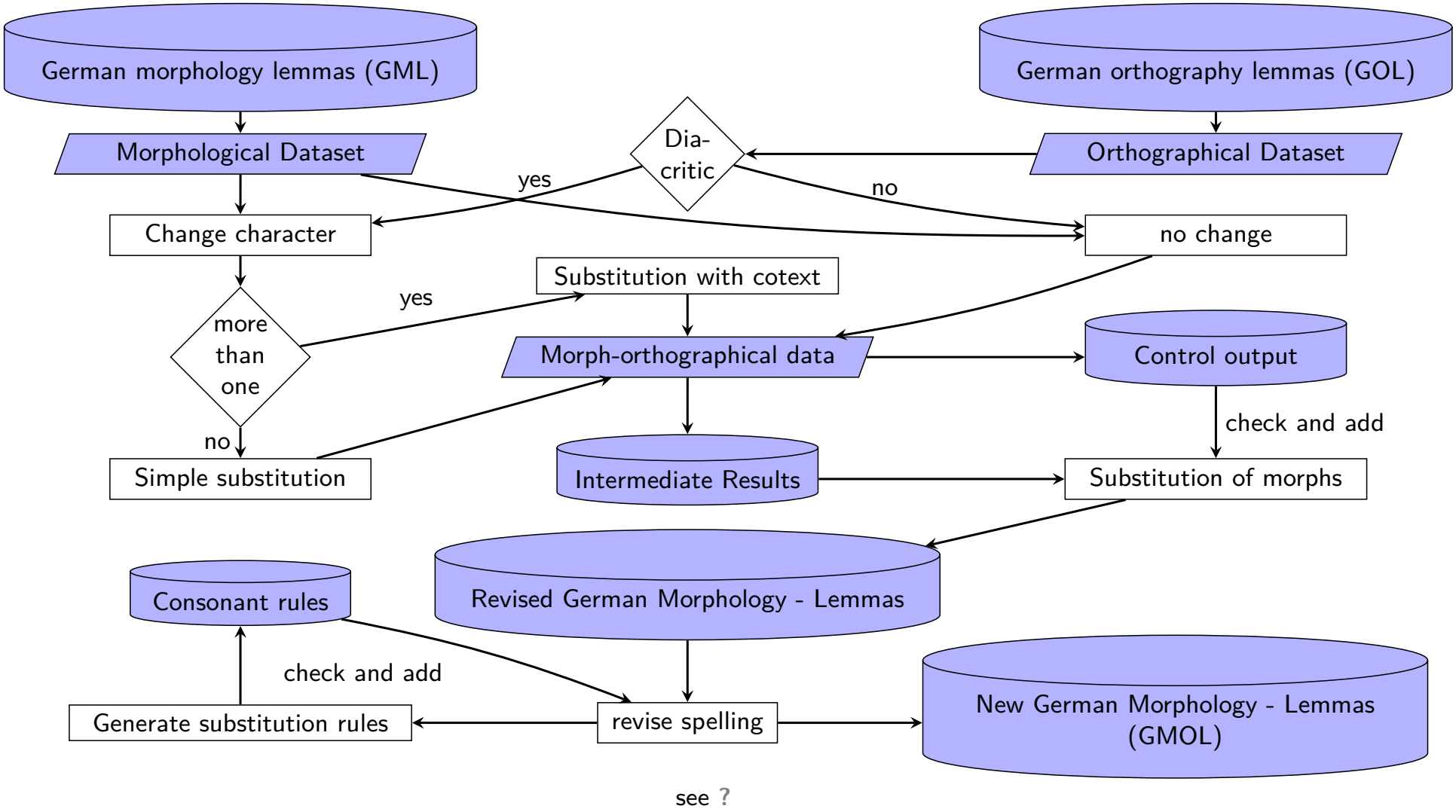
- 100,986 merged trees
- Currently the biggest available data resource of its kind
- Text coverage of 60.59%
- Combined with Morphy: 92.73%
- Combined with SMOR: 98.80%
- Downloads without data:
  `https://github.com/petrasteiner/morphology`

The authors were partially supported by the German Research Foundation (DFG) under grant RU 1873/2-1 and by a Marie Curie Career Integration Grant within the 7th European Community Framework Programme.

The Lexical Database CELEX

- Dutch, English, and German lexical information
- combined with information on word-formation types and frequencies
- manually annotated multi-tiered word structures (Baayen et al., 1995)
- outdated format, e.g.
  Abschlu$ (orthographical dataset)
  Abschluss (morphological dataset)
  $\longmapsto$ *Abschluß* (modern format) 'conclusion'
- outdated spelling, e.g.
  Abschluß 'conclusion' $\longmapsto$ *Abschluss* (modern spelling).

Transfer of CELEX to the Modern Standard



see ?

CELEX Revision - Facts and Figures

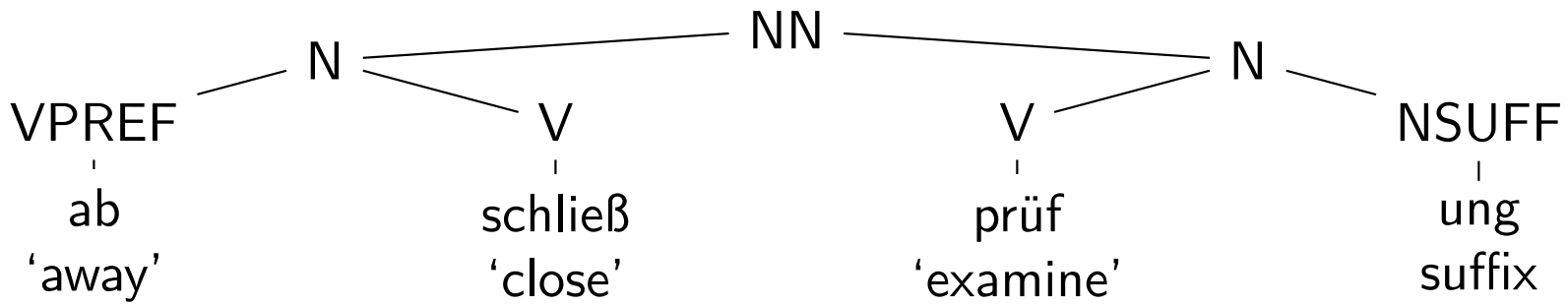
- 51,728 entries
- 10,106 entries with diacritics
- 576 entries with updated spelling
- 38,683 complex entries (morphological deep-level analyses)

The Lexical Database CELEX: Morphological Structures

- manually annotated multi-tiered word structures (Baayen et al., 1995)
- needs only a few repairs of missing constituents or wrong analyses

Example (Morphological analysis of *Abschlußprüfung* 'final exam')

Abschlusspruefung ('final exam') Abschluss+Pruefung
((((ab)[V|.V],(schliess)[V])[V])[N], ((pruef)[V],(ung)[N|V.])[N])



CELEX Trees IV: Restriction of Diachronic Information

- Cut two forms f_1, f_2 with length l_1 and l_2 to the strings s_1, s_2 of the smaller length ($min(l_1, l_2)$) and calculate the Levensthein distance (LD) of these. Special characters such as *ä* or *ß* are transformed to *a* and *ss*, uppercase characters to lowercase. Then the quotient of both values is compared to a threshold t as in (3):

$$\frac{LD(s_1, s_2)}{min(l_1, l_2)} < t \tag{3}$$

  Example: the stem of the derived form *treib* and its component *driften* are reduced to the smaller size (5): *drift* and *treib*. (4) shows that the analysis will stop for a threshold at 0.8 or below.

$$\frac{LD(drift, treib)}{min(l_1, l_2)} = \frac{4}{5} \tag{4}$$

- Plus a small list of exceptions. Steiner and Ruppenhofer (2018)

Algorithm 1: Building a merged morphological treebank

```
Input: CELEX-German revised, GN flat compounds
Output: A DB of Morphological Trees
initialization of parameters: depth of analysis, linguistic
  information, levenshtein threshold, parts of speech, style of
  output;
add CELEX data to the knowledge base
  forall entries of GN flat compounds do
    if entry is a compound then
      foreach constituent of entry do
        if depth of analysis reached then
          retrieve linguistic information/PoS as
            required;
          return linguistic information and
            constituent
        end
        else if constituent not found in GN data
          then
          depth of analysis++;
          analysedeepercelex part with
            parameters and depth;
          return result of analysedeepercelex
        end
        else
          foreach part of constituent do
            depth of analysis++;
            analysedeeper part with
              parameters and depth;
            return result of analysedeeper
          end
        end
      end
    end
end
```

```
sub analysedeeper part (parameters and level)
  if part is simplex or depth of analysis reached
  then
    retrieve linguistic information/PoS as required;
    return linguistic information and part
end
else if constituent not found in GN data then
    depth of analysis++;
    analysedeepercelex part with parameters and depth;
    return result of analysedeepercelex
end
else
    depth of analysis++;
    foreach subpart of part do
      analysedeeper subpart
        return result of analysedeeper subpart
    end
end
sub analysedeepercelex part (parameters and level)
  if part is simplex or depth of analysis reached
  then
    retrieve linguistic information/PoS as required;
    return linguistic information and part
end
else
    foreach subpart of part do
      analysedeepercelex subpart
        if levenshtein threshold and
        analysedeepercelex subpart is dissimilar then
          skip deeper analysis;
          return subpart
        end
        else
          return result of analysedeepercelex
            subpart
        end
    end
end
```

Conversion or ambiguity?

Examples (GermaNet vs. CELEX)

Werkstück	Werk|Stück	'work(noun)|piece'
Werkstück	werken|Stück	'to work|piece'
Glaswerkstück	Glas|(Werk|Stück)	'glass|work(noun)|piece'
Glaswerkstück	Glas|(werken|Stück)	'glass|to work|piece'

Werkstück	(*werken_V* (Werk_N)(en_x))(Stück_N)	
	'(*to work_V* (work_N)(en(suffix))(piece_N)'	

Compounding or Prefixation/Conversion?

Examples (GermaNet vs. CELEX)

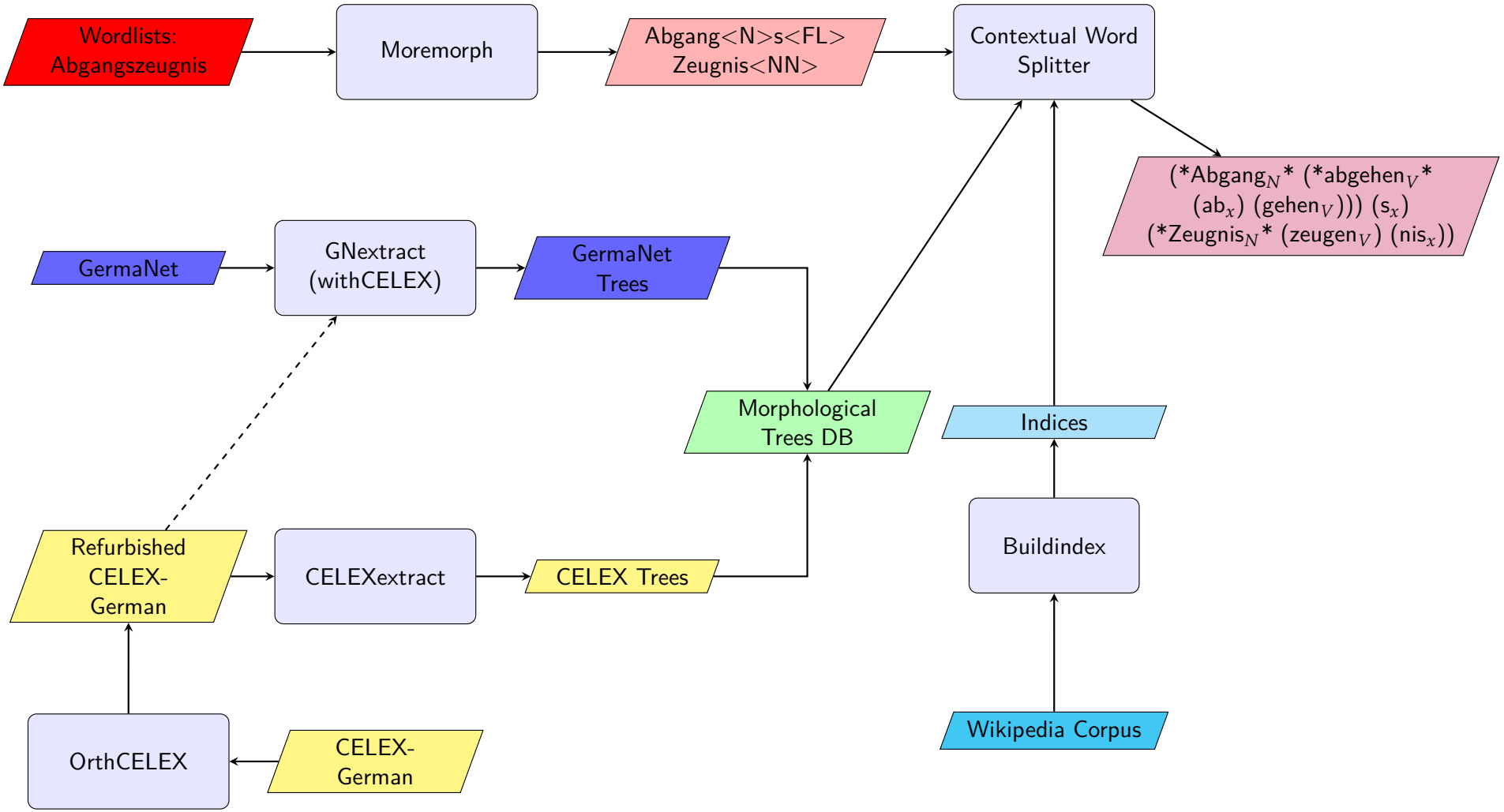
Abwasser	(ab_P)(Wasser_N)	'(away_P)(water_N) waste water'
	(ab_x)(Wasser_N)	'(away_x)(water_N) waste water'
afroasiatisch	(afro_R)(Asiatisch_N)	'(afro_R)(Asian_N)'
afroamerikanisch	(afro_x)(amerikanisch_A)	'(afro_x)(American_A)'
Maßnahme	(Maß_N)(Nahme_N)	'(measure_n)(taking_N) measure'
	maßnehmen_V	'(to measure_take_V) measure'

Mapping the Morphological Tagsets

Part of Speech/morph type	GN	CELEX	GN Trees
noun	nomen, Nomen	N	N
adjective	Adjektiv	A	A
adverb	Adverb	B	B
preposition	Präposition	P	P
verb	Verb, verben	V	V
article	Artikel	D	D
interjection	Interjektion	I	I
pronoun	Pronomen	O	O
abbreviation	Abkürzung	X	X
word group	Wortgruppe	n	n
root/confix	Konfix	R	R
filler letters, affixes	-	x	x

Table 3: Mapping of two morphological tagsets

Outlook


Wordlists: Abgangszeugnis
Moremorph
Abgang<N>s<FL> Zeugnis<NN>
Contextual Word Splitter
(*Abgang$_N$* (*abgehen$_V$* (ab$_x$) (gehen$_V$))) (s$_x$) (*Zeugnis$_N$* (zeugen$_V$) (nis$_x$))
GermaNet
GNextract (withCELEX)
GermaNet Trees
Morphological Trees DB
Indices
Refurbished CELEX-German
CELEXextract
CELEX Trees
Buildindex
OrthCELEX
CELEX-German
Wikipedia Corpus

Morphs and immediate constituents I

Three datasets with frequency information were extracted from CELEX:

- all morphs with their frequencies within the lemmas (13,419 entries)

ung	3588	ein	755
er	3066	ge	750
ig	2531	auf	681
s	2327	über	630
e	2120	um	557
ver	1694	vor	517
n	1581	bar	485
lich	1273	heit	475
be	1236	ent	475
ier	1215	los	455
un	1141	en	423
aus	983	ation	394
keit	974	t	382
ab	896	unter	381
an	845	zu	378
isch	836	in	374

- all immediate constituents with their frequencies within the lemmas (21,406 entries)
- all immediate constituents within the lemmas with their frequencies as found in the *Mannheim Corpus*.

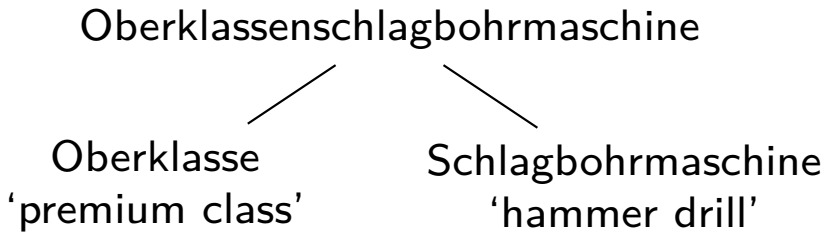
Morphs and immediate constituents II

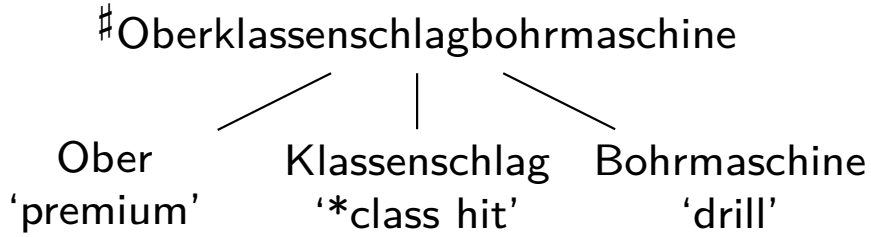
Preliminary results: smallest parts of GN trees.

11905	s	771	al
8961	ung	730	Land
5939	n	728	ation
5339	e	721	ion
5324	er	715	Zeit
2198	ge	634	stellen
1564	be	631	fahren
1537	ver	622	Bau
1452	en	620	heit
1197	schaft	577	bauen
913	es	569	Arbeit

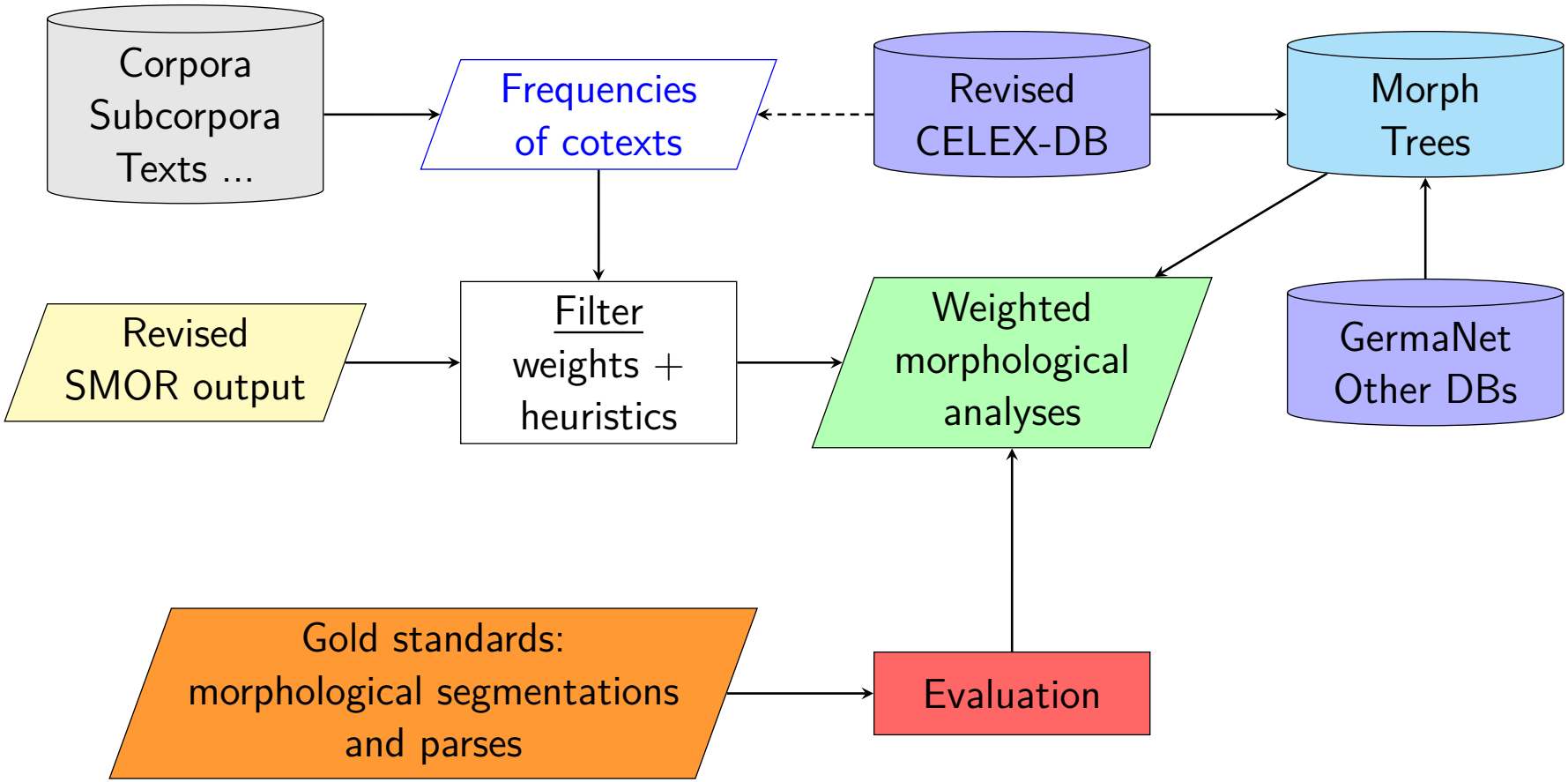
Characteristics of German Word-Formation Version 2

- language with complex processes of word formation
- most common are compounding and derivation
  Oberklasse-
  Kompaktschlagbohrmaschine
  'Premium class compact hammer drill (machine)'
- many combinatorially possible analyses





Overview - New Version



SMOR

- Stuttgarter Morphologisches Analysewerkzeug
- Morphological analyzer based on two-level morphology, implemented as a set of finite-state transducers (Schmid et al., 2004)
- Main lexicon with 41,944 entries, proper name lexicons with 15,188 entries and different datasets with other morphological information

(6)
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Nom><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Nom><Sg>

SMOR

- Stuttgarter Morphologisches Analysewerkzeug
- Morphological analyzer based on two-level morphology, implemented as a set of finite-state transducers (Schmid et al., 2004)
- Main lexicon with 41,944 entries, proper name lexicons with 15,188 entries and different datasets with other morphological information

(6)
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Nom><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Nom><Sg>

SMOR

- Stuttgarter Morphologisches Analysewerkzeug
- Morphological analyzer based on two-level morphology, implemented as a set of finite-state transducers (Schmid et al., 2004)
- Main lexicon with 41,944 entries, proper name lexicons with 15,188 entries and different datasets with other morphological information

(6)

ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Nom><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Nom><Sg>

SMOR

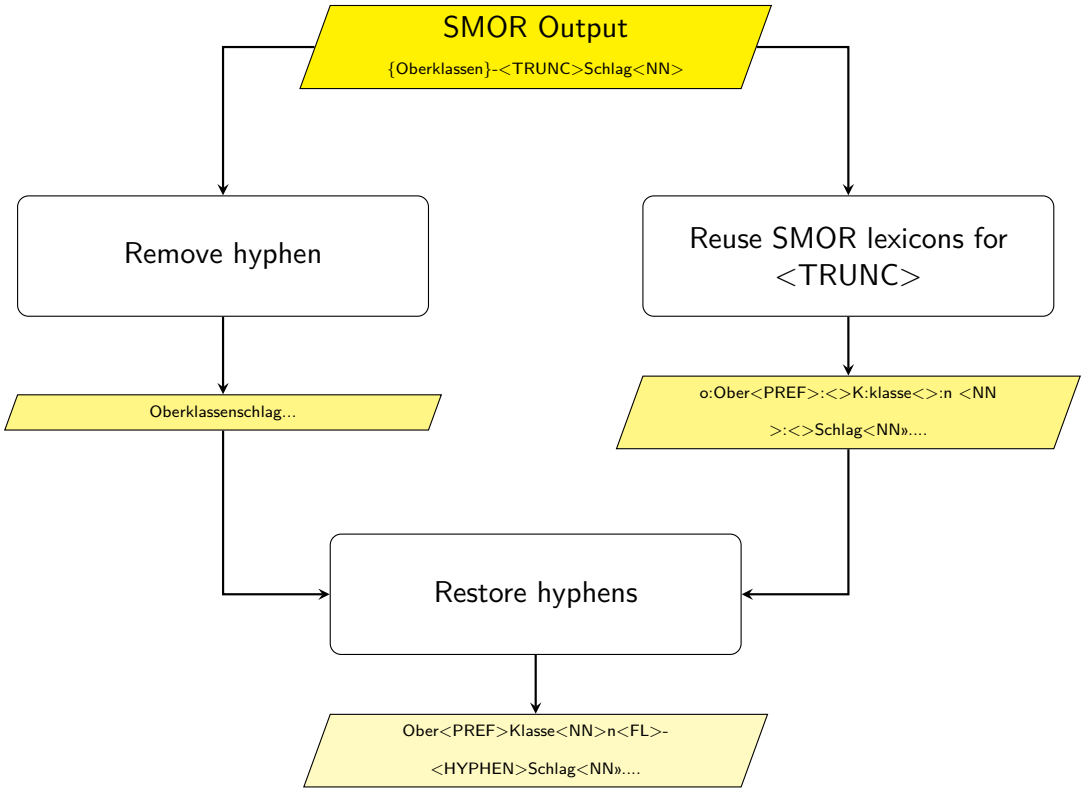
- Stuttgarter Morphologisches Analysewerkzeug
- Morphological analyzer based on two-level morphology, implemented as a set of finite-state transducers (Schmid et al., 2004)
- Main lexicon with 41,944 entries, proper name lexicons with 15,188 entries and different datasets with other morphological information

(6)

ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN> Schlag <NN>bohren<V>Maschine<+NN><Fem><Nom><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Acc><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Dat><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Gen><Sg>
ober<PREF>Klasse<NN> schlagen <V><NN><SUFF>bohren<V>Maschine<+NN><Fem><Nom><Sg>

SMOR tag <TRUNC>

(7) {Oberklassen}-<TRUNC>Schlag<NN>bohren <V>Maschine<+NN><Fem><Acc><Sg>



(8) Ober<PREF>Klasse<NN>n<FL>-<HYPHEN>Schlag<NN>bohren<V>Maschine<+NN>

Coverage / Form of Output

Table 4 summarizes the changes for 1,101 items from our gold standard data.

Method	*t*	*n*	*a*	*r*
(a) SMOR baseline	105	3	0	0.00
(b) remove hyphens	48	2	58	0.54
(c) reanalyze TRUNC	39	3	66	0.61
(d) combine (b) and (c)	2	2	104	0.96

Table 4: Analyzed hyphenated forms; *t*: analyses containing TRUNC; *n*: hyphenated forms without analyses; *a*: correctly pre-analyzed hyphenated word form; *r*: relative frequency of *a*

Geometric Mean Score

We use the geometric mean as in (5)

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \quad for \quad x_i ... x_n, \tag{5}$$

Anbaumenge , $x_1 = 845$ for *an* , $x_2 = 168$ for *bau* , $x_3 = 8$ for *Menge*
$gm(An|bau|Menge) = 104.33$
Frequencies of morphs, of constituents, of corpus frequencies

Geometric Mean Score

We use the geometric mean as in (5)

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \quad for \quad x_i ... x_n, \tag{5}$$

Anbaumenge , $x_1 = 845$ for *an* , $x_2 = 168$ for *bau* , $x_3 = 8$ for *Menge*
$gm(An|bau|Menge) = 104.33$
Frequencies of morphs, of constituents, of corpus frequencies

Geometric Mean Score

We use the geometric mean as in (5)

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \quad for \quad x_i ... x_n, \tag{5}$$

Anbaumenge , $x_1 = 845$ for *an* , $x_2 = 168$ for *bau* , $x_3 = 8$ for *Menge*
$gm(An|bau|Menge) = 104.33$
Frequencies of morphs, of constituents, of corpus frequencies

Combinatorial structure of morphological analyses

There is isomorphy to the permuted integer partitions of *n*

(9) *Drahtseilakt* 'High-wire act'

a. [[*Draht*],[*seil*],[*akt*]]

b. [[*Draht*],[seilakt]]

c. [[Drahtseil],[*akt*]]

d. [[*Drahtseilakt*]]

Corresponding integer compositions

(10) a. 1-1-1

b. 1- 2

c. 2 -1

d. 3

The algorithm for processing the combinatorially possible analyses makes use of this analogy. $c(n) = 2^{n-1}; n >= 1$

Combinatorial structure of morphological analyses

There is isomorphy to the permuted integer partitions of *n*

(9) *Drahtseilakt* 'High-wire act'

a. [[*Draht*],[*seil*],[*akt*]]

b. [[*Draht*],[seilakt]]

c. [[Drahtseil],[*akt*]]

d. [[*Drahtseilakt*]]

Corresponding integer compositions

(10) a. 1-1-1

b. 1- 2

c. 2 -1

d. 3

The algorithm for processing the combinatorially possible analyses makes use of this analogy. $c(n) = 2^{n-1}; n >= 1$

Combinatorial structure of morphological analyses

There is isomorphy to the permuted integer partitions of *n*

(9) *Drahtseilakt* 'High-wire act'

a. [[*Draht*],[*seil*],[*akt*]]

b. [[*Draht*],[seilakt]]

c. [[Drahtseil],[*akt*]]

d. [[*Drahtseilakt*]]

Corresponding integer compositions

(10) a. 1-1-1

b. 1- 2

c. 2 -1

d. 3

The algorithm for processing the combinatorially possible analyses makes use of this analogy. $c(n) = 2^{n-1}; n >= 1$

Path pruning

(11) Compositions of *abwechslungsreich*

a. $[[ab],[wechsl],[ung,s],[reich]]$
b. $[[ab],[wechsl],[ung,s,reich]]$
c. $[[ab],[wechsl,ung,s],[reich]]$
d. $[[ab],[wechsl,ung,s,reich]]$
e. $[[ab,wechsl],[ung,s],[reich]]$
f. $[[ab,wechsl],[ung,s,reich]]$
g. $[[ab,wechsl,ung,s],[reich]]$
h. $[[ab,wechsl,ung,s,reich]]$

(12)

Be	nutz	er	unter	stütz	ung
VPREF	V	NNSUFF	VPART	V	NNSUFF
be.pref	use	er.suff	below	support	ung.suff

The Lexical Database CELEX: Example

- orthographical data, e.g.
  605\Abschlu$pr”ufung\14\Ab-schlu$-prü-fung\N\
  Abschlu$prüfung\Ab-schlu$-prü-fung\N
- morphological data, e.g.
  605\Abschlusspruefung\14\C\1\Y\Y\Y\Abschluss+Prue-
  fung\NN\N\
  N\N\((((ab)[V|.V],(schliess)[V])[V])[N],((pruef)[V],(ung)[N|V.])[N])[N]\
  Y\N\N\N\S3/P3\N
- phonological data, e.g.
  605\Abschlusspruefung\14\’&p-SIUs-pry-fUN\[ap][SIUs][pry:][fUN]
  \’&p-SIUs-pry-fUN\[ap][SIUs][pry:][fUN]\[VC][CCVC][CCVV][CVC]\
  [VC][CCVC][CCVV][CVC]\ap#Sli:s#pry:f+UN\ap#Sli:s#pry:f+UN
- syntactic data, e.g.
  605\Abschlusspruefung\14\1\2\\N\N\\\\\\\\\\\\\\\\
- frequency data

CELEX Trees IV: Some repairs

- missing constituents and missing parts of speech information within the morphological trees
- missing constituents within the field of immediate constituency information
- inconsistent morphological analyses e.g. for phrasal compounds.

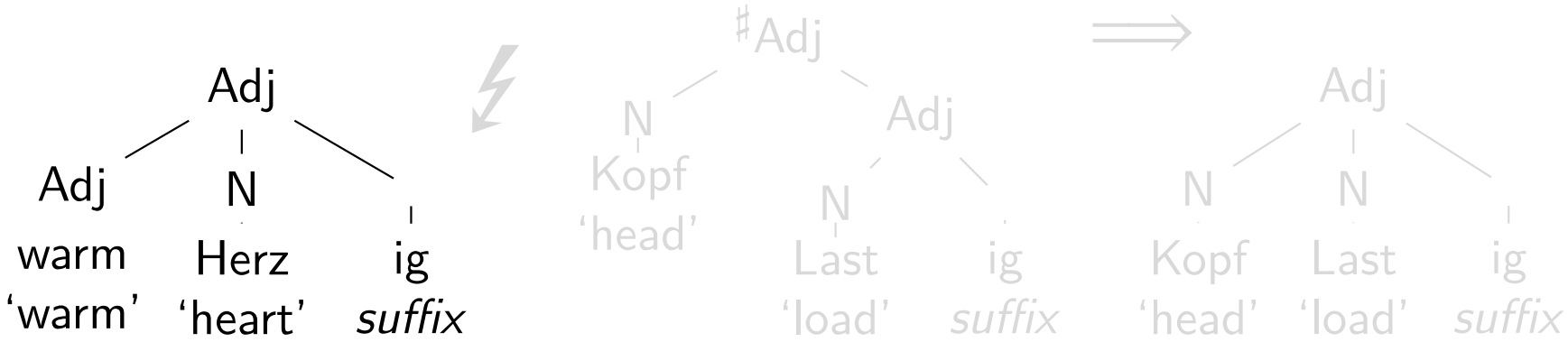


Figure 10: *warmherzig* ‘warm-heartedly’ and *kopflastig* ‘top heavy’

CELEX Trees IV: Some repairs

- missing constituents and missing parts of speech information within the morphological trees
- missing constituents within the field of immediate constituency information
- inconsistent morphological analyses e.g. for phrasal compounds.

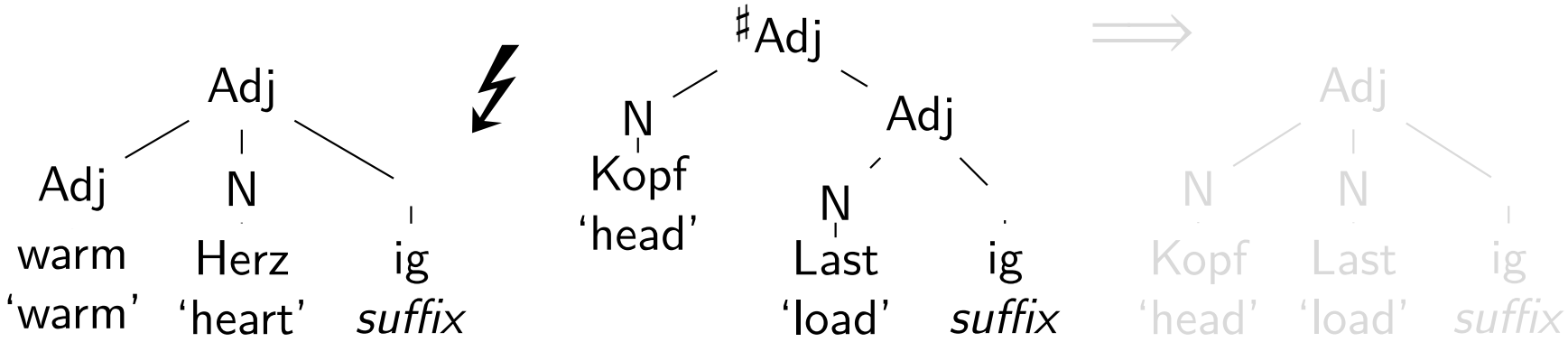


Figure 10: *warmherzig* ‘warm-heartedly’ and *kopflastig* ‘top heavy’

CELEX Trees IV: Some repairs

- missing constituents and missing parts of speech information within the morphological trees
- missing constituents within the field of immediate constituency information
- inconsistent morphological analyses e.g. for phrasal compounds.

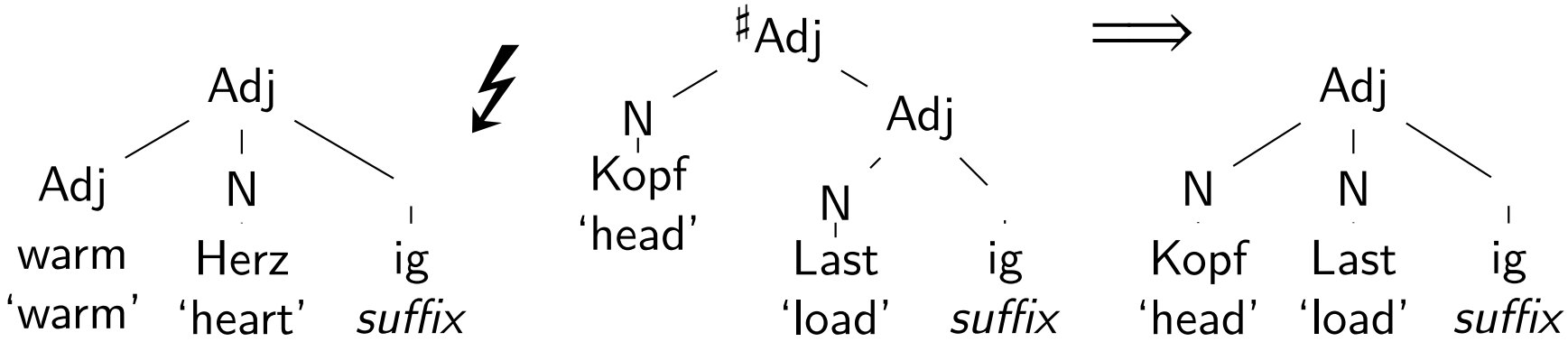


Figure 10: *warmherzig* ‘warm-heartedly’ and *kopflastig* ‘top heavy’