Indirect Supervision for the Determination and Structural Analysis of Nominal Compounds

Patrick Ziering

Patrick.Ziering@ims.uni-stuttgart.de
www.ziering.de

PhD Defense
21st December 2017
waiting room
French teacher

↓

compounds or phrases?

↓

composed constituents?
waiting room
French teacher
timetable
database

↓ ↓
comounds or phrases? composed constituents?

noun compound analysis

[noun compound] analysis noun [compound analysis]
Outline of the Talk

1 Introduction
Outline of the Talk

1. Introduction

Main Contributions

2. New Insights on the Notion of Compoundhood

3. Cross-lingual Evidence

4. Morphological Regularities
Outline of the Talk

1. Introduction

Main Contributions

2. New Insights on the Notion of Compoundhood

3. Cross-lingual Evidence

4. Morphological Regularities

5. Summary and Conclusions
Outline of the Talk

1 Introduction

Main Contributions

2 New Insights on the Notion of Compoundhood

3 Cross-lingual Evidence

4 Morphological Regularities

5 Summary and Conclusions
Compounds - Characteristics and Motivation

Definition/existence controversial

- **The formation of a new lexeme by adjoining two or more lexemes** (Bauer, 2003)
- **No compounding word formation** (Marchand, 1967)
Compounds - Characteristics and Motivation

Definition/existence controversial

👍 the formation of a new lexeme by adjoining two or more lexemes (Bauer, 2003)

👎 no compounding word formation (Marchand, 1967)

Compounds are located between words and phrases

French  French fries  French car
Compounds - Characteristics and Motivation

Definition/existence controversial

- The formation of a new lexeme by adjoining two or more lexemes (Bauer, 2003)
- No compounding word formation (Marchand, 1967)

Compounds are located between words and phrases

French fries French car French

High productivity but only few individual instances

⇒ problematic for statistical techniques
Nominal Compounds across Languages

Nominal Compounds are a multilingual phenomenon

⇒ Closed compounds for airport in various language families
   spanish: aeropuerto, german: Flughafen, polish: lotnisko, turkish: havaalani
   greek: αεροδρόμιο, hungarian: letištő, russian: аэропорт, finnish: lentokenttä

⇒ Large variation in surface form (e.g., for human rights abuse)

→ German: Mißachtung der Menschenrechte
→ Italian: violazioni dei diritti umani
Nominal Compounds across Languages

Nominal Compounds are a multilingual phenomenon

⇒ Closed compounds for airport in various language families

- aeropuerto (Spanish)
- Flughafen (German)
- lotnisko (Polish)
- havaalani (Turkish)
- αεροδρόμιο (Greek)
- летище (Bulgarian)
- аэропорт (Russian)
- lentokenttä (Finnish)

⇒ Large variation in surface form (e.g., for human rights abuse)

- Mißachtung der Menschenrechte (German)
- violazioni dei diritti umani (Italian)

Compound Analysis Approaches in this Thesis

⇒ Cross-lingual perspective

⇒ Multilingual design
**Compound Analysis - Determination of Compoundhood**

- **Atomic word**: friendship
- **Compound**: French toast
- **Phrase**: French car

Determination of compoundhood
**Compound Analysis - Structural Analysis**

- **Splitting**
  - Ei/dotter 'egg yolk'

- **Structural analysis**

- **Parsing**
  - [egg yolk] yellow

- **Determination of compoundhood**

- **Atomic word**
  - friendship

- **Compound**
  - French toast

- **Phrase**
  - French car
**Compound Analysis - Semantic Analysis**

- **Compositionality**
  - *honey bee vs. honey moon*

- **Splitting**
  - *Ei/dotter ‘egg yolk’*

- **Parsing**
  - *[egg yolk] yellow*

- **Relation**
  - *substance plastic fork*

- **Determinant of compoundhood**

- **Semantic analysis**

- **Structural analysis**

- **Atomic word**
  - *friendship*

- **Compound**
  - *French toast*

- **Phrase**
  - *French car*
**Compound Analysis - in this thesis**

**Compositionality**
- *honey bee* vs. *honey moon*

**Splitting**
- $E_i/dotter$ ‘egg yolk’

**Semantic analysis**

**Determination of compoundhood**

**Structural analysis**

**Parsing**
- [egg yolk] yellow

**Relation**
- *substance* plastic fork

**Atomic word**
- *friendship*

**Compound**
- *French toast*

**Phrase**
- *French car*
1. PRINCIPLE - INDIRECT SUPERVISION

✗ Avoid **supervised** approaches with **task-dependent labels**
   (e.g., annotated compound parse trees)
   ✗ expensive
   ✗ domain/language dependent, ...

✗ Avoid **unsupervised** approaches
   ✗ worse performance than supervised (Vadas and Curran, 2007)
1. Principle - Indirect Supervision

✗ Avoid **supervised** approaches with **task-dependent labels**
   (e.g., annotated compound parse trees)
   - expensive
   - domain/language dependent, ...

✗ Avoid **unsupervised** approaches
   - worse performance than supervised ([Vadas and Curran, 2007](#))

---

**Indirect Supervision**

✔ Exploit **task-independent** information
  → beneficial in an indirect way
  → e.g., cross-lingual equivalents in parallel data

✔ **High quality training data**
  → close to gold labels

- Knowledge resources
  - bilingual dictionaries
  - lexical resources (e.g., WordNet)
- Extensive human support
  - transformation rules
  - list of linking elements

- Knowledge resources
  - bilingual dictionaries
  - lexical resources (e.g., WordNet)
- Extensive human support
  - transformation rules
  - list of linking elements

× Knowledge resources
  × bilingual dictionaries
  × lexical resources (e.g., WordNet)

× Extensive human support
  × transformation rules
  × list of linking elements

✓ (Parallel) corpora
✓ Linguistically informed strategies
  → language-independent assumptions

- Knowledge resources
  - bilingual dictionaries
  - lexical resources (e.g., WordNet)
- Extensive human support
  - transformation rules
  - list of linking elements

(Parallel) corpora

Linguistically informed strategies
→ language-independent assumptions

Multilingually applicable
Independent of specific domains
Main contributions

1. **New Insights on the Notion of Compoundhood**
   - Various *linguistic criteria* for compoundhood
   - *Cross-lingual* study on nominal compounding

2. **Novel Compound Analysis Approaches**
   - Indirectly supervised and *manual-resource-lean*

   - **Cross-lingual Evidence on Association Strength**
     - *Compound identification*
     - *Compound parsing*

   - **Morphological Regularities:**
     - *Compound splitting* (and constituent normalization)
Outline of the Talk

1. Introduction
2. Main Contributions
   - New Insights on the Notion of Compoundhood
3. Cross-lingual Evidence
4. Morphological Regularities
5. Summary and Conclusions
Some Linguistic Criteria in Literature

Compound
Some Linguistic Criteria in Literature

Spelling

blackbird

Compound
Some Linguistic Criteria in Literature

- Spelling: blackbird
- Inseparability: black ugly bird

Compound
Some Linguistic Criteria in Literature

- Spelling: *blackbird*
- Inseparability: *black ugly bird*
- Inability to modify the modifier: *very blackbird*
Some Linguistic Criteria in Literature

- **Spelling**
  - *blackbird*

- **Inseparability**
  - *black ugly bird*

- **Inability to modify the modifier**
  - *very blackbird*

- **Head replacement by ‘one’**
  - *blackbird ≠ black one*
Some Linguistic Criteria in Literature

- **Spelling**
  - *blackbird*

- **Inseparability**
  - *black ugly bird*

- **Inability to modify the modifier**
  - *very blackbird*

- **Modifier-specific inflection**
  - *shoe seller frontiersman*

- **Head replacement by ‘one’**
  - *blackbird ≠ black one*
Some Linguistic Criteria in Literature

- Spelling: blackbird
- Inseparability: black ugly bird
- Inability to modify the modifier: very blackbird
- Prosody: blackbird
- Modifier-specific inflection: shoe seller, frontiersman
- Head replacement by ‘one’: blackbird ≠ black one
Linguistic Criterion Inspection

Which linguistic criteria are most relevant for the compoundhood determination?
Linguistic Criterion Inspection

Which linguistic criteria are most relevant for the compoundhood determination?

Nominal Compound Resource

Two experienced judges

→ annotated and rated nominal compound candidates
  → degree of compoundhood
  → validity of linguistic criteria

⇒ 824 NCs → Gold standard for identification
Linguistic Criterion Inspection

Which linguistic criteria are most relevant for the compoundhood determination?

Nominal Compound Resource

Two experienced judges

→ annotated and rated nominal compound candidates
  → degree of compoundhood
  → validity of linguistic criteria

⇒ 824 NCs → Gold standard for identification

Goal

Correlation between compoundhood and linguistic criteria

→ Classification experiments
Classification Experiment

- Target class: compoundhood rating (1, 2, 3)
- 10-fold cross-validation, decision tree classifier
- Each linguistic criterion as single feature
**Classification Experiment**

- Target class: compoundhood rating (1, 2, 3)
- 10-fold cross-validation, decision tree classifier
- Each linguistic criterion as single feature

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spelling</strong></td>
<td>45.1%</td>
</tr>
<tr>
<td><strong>Inseparability</strong></td>
<td>50.7%</td>
</tr>
<tr>
<td><strong>Inability to Modify the Modifier</strong></td>
<td>50.1%</td>
</tr>
<tr>
<td><strong>Replacement by ‘one’</strong></td>
<td>48.2%</td>
</tr>
<tr>
<td><strong>Modifier-specific inflection</strong></td>
<td>48.2%</td>
</tr>
<tr>
<td><strong>Prosody</strong></td>
<td>48.9%</td>
</tr>
<tr>
<td><strong>Majority class baseline</strong></td>
<td>38.7%</td>
</tr>
</tbody>
</table>

→ Results are in line with additional experiments
## Classification Experiment

- Target class: compoundhood rating (1, 2, 3)
- 10-fold cross-validation, decision tree classifier
- Each linguistic criterion as single feature

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling</td>
<td>45.1%</td>
</tr>
<tr>
<td>Inseparability</td>
<td>50.7%</td>
</tr>
<tr>
<td>Inability to Modify the Modifier</td>
<td>50.1%</td>
</tr>
<tr>
<td>Replacement by ‘one’</td>
<td>48.2%</td>
</tr>
<tr>
<td>Modifier-specific inflection</td>
<td>48.2%</td>
</tr>
<tr>
<td>Prosody</td>
<td>48.9%</td>
</tr>
<tr>
<td>Majority class baseline</td>
<td>38.7%</td>
</tr>
</tbody>
</table>

→ Results are in line with additional experiments
How much do people agree in the notion of compoundhood?
Linguistic Criterion Inspection

How much do people agree in the notion of compoundhood?

<table>
<thead>
<tr>
<th># extractions</th>
<th>Jacc</th>
<th>P</th>
<th>R</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anno1 ∩ Anno2</td>
<td>119</td>
<td>66</td>
<td>100</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.660</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.555</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.603</td>
</tr>
</tbody>
</table>

- Moderate Inter-Annotator Agreement
Linguistic Criterion Inspection

How much do people agree in the notion of compoundhood?

<table>
<thead>
<tr>
<th># extractions</th>
<th>Jacc</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anno1 ∩ Anno2</td>
<td>119</td>
<td>66</td>
<td>100</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td>0.660</td>
<td>0.555</td>
<td>0.603</td>
<td></td>
</tr>
</tbody>
</table>

- **Moderate** Inter-Annotator Agreement
  - In line with Vincze et al. (2011) who reported a Jaccard of 0.552
Linguistic Criterion Inspection

How much do people agree in the notion of compoundhood?

<table>
<thead>
<tr>
<th># extractions</th>
<th>Jacc</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anno1 ∩ Anno2</td>
<td>119</td>
<td>66</td>
<td>100</td>
<td>0.431</td>
</tr>
</tbody>
</table>

- Moderate Inter-Annotator Agreement
  - In line with Vincze et al. (2011) who reported a Jaccard of 0.552
- Serves as upper bound for compound identification
What do the surface patterns of cross-lingual equivalents look like?
What do the surface patterns of cross-lingual equivalents look like?

→ Noun-noun compound gold standard (Ó Séaghdha, 2007)
   → Matches with Europarl: 468 types, 11,793 tokens

→ Word alignment to 9 European languages:
What do the surface patterns of cross-lingual equivalents look like?

→ Noun-noun compound gold standard (Ó Séaghdha, 2007)
  → Matches with Europarl: 468 types, 11,793 tokens

→ Word alignment to 9 European languages:

Results

~ 30% of all equivalents are closed compounds
~ 70% of Germanic equivalents are closed compounds
~ 80% of Romance equivalents are phrases
Conclusion: The Most Important Criteria

Inseparability
CONCLUSION: THE MOST IMPORTANT CRITERIA

- Inseparability
- Inability to modify the modifier
Conclusion: The Most Important Criteria

- Inseparability
- Inability to modify the modifier
- Prosody
Conclusion: The Most Important Criteria

- Inability to modify the modifier
- Prosody
- Inseparability
- Cross-lingual surface forms
CONCLUSION: THE MOST IMPORTANT CRITERIA

- Inseparability
- Inability to modify the modifier
- Prosody
- Cross-lingual surface forms

Compounds
Conclusion: The Most Important Criteria

- Inseparability
- Inability to modify the modifier
- Prosody
- Cross-lingual surface forms
Outline of the Talk

1. Introduction

Main Contributions

2. New Insights on the Notion of Compoundhood

3. Cross-lingual Evidence

4. Morphological Regularities

5. Summary and Conclusions
Cross-lingual Evidence on Association Strength

Cross-lingual Supervision

→ Indirect supervision
→ Exploit cross-lingual evidence
→ Based on parallel data
→ Used for splitting (K & K, 2003)
Cross-lingual Evidence on Association Strength

Cross-lingual Supervision

→ Indirect supervision
→ Exploit cross-lingual evidence
→ Based on parallel data
→ Used for splitting (K & K, 2003)

Cross-lingual Supervision with Association Strength

(Ziering and Van der Plas, 2014, 2015a, 2015b)

• Spelling of cross-lingual equivalents
  → compound identification
• Relative sentence position of constituent equivalents
  → compound parsing
## 1. Compound Identification

### The Task
Identification of compounds in context

### Limitations in Previous Approaches
- Lauer (1995) *(two-noun compounds)*
- Ramisch et al. (2010) *(MWETOOLKIT)*
- Vincze et al. (2011) *(rely on Wikipedia links)*
- Ziering et al. (2013) *(base NPs)*

### Publication Related to the PhD Thesis
⇒ Ziering and Van der Plas (COLING, 2014)
**Closed Compounds among Equivalents**

Based on PoS patterns, e.g., ADJ+NN

*last year*
Closed Compounds among Equivalents

based on PoS patterns
e.g., ADJ+NN

parallel corpus

- sidste år
- letztes Jahr
- vorig jaar
- förra året

last year
Closed Compounds among Equivalents

Based on PoS patterns e.g., ADJ+NN

Parallel corpus

closed compounds?

last year

sidste år

letztes Jahr

vorig jaar

förra året

0
Closed Compounds among Equivalents

based on PoS patterns
e.g., ADJ+NN

financial crisis
Closed Compounds among Equivalents

- based on PoS patterns e.g., ADJ+NN

- financial crisis → finanskrise
- Finanzkrise
- financiële crisis
- finansiella krisen
**Closed Compounds among Equivalents**

Based on PoS patterns, e.g., ADJ+NN

Parallel corpus

Closed compounds?

- **financial crisis**
- **finanskrise**
- **Finanzkrise**
- **financiële crisis**
- **finansiella krisen**
Closed Compounds among Equivalents

based on PoS patterns
e.g., ADJ+NN

agricultural policy
Based on PoS patterns, e.g., ADJ+NN, parallel corpus:

- agricultural policy
- landbrugspolitik
- Agrarpolitik
- landbouwbeleid
- jordbrukspolitiken
Based on PoS patterns, e.g., ADJ+NN, closed compounds among equivalents can be identified. For example, the agricultural policy based on PoS patterns is represented by parallel corpus. The closed compounds for agricultural policy include:

- **Danish**: landbrugspolitik
- **German**: Agrarpolitik
- **Dutch**: landbouwbeleid
- **Swedish**: jordbrukspolitiken

This approach allows for the identification of closed compounds among equivalents across different languages.
Closed Compounds among Equivalents

Based on PoS patterns e.g., ADJ+NN

Parallel corpus

Closed compounds?

Agricultural policy

Landbrugspolitik

Agrarpolitik

Landbouwbeleid

Jordbrukspolitiken

Closed-Compound Restriction (CCR)
Closed Compounds among Equivalents

Based on PoS patterns e.g., ADJ+NN

Parallel corpus

Closed compounds?

agricultural policy

landbrugspolitik

Agrarpolitik

landbouwbeleid

jordbrukspolitiken

Closed-Compound Restriction (CCR)
Closed Compounds among Equivalents

Based on PoS patterns e.g., ADJ+NN

Parallel corpus

Closed compounds?

Agricultural policy

Landbrugspolitik

Agrarpolitik

Landbouwbeleid

Jordbrukspolitiken

Closed-Compound Restriction (CCR)
Compound Identification Experiments

- Precision highest with CCR=4
**Compound Identification Experiments**

- Precision highest with $CCR=4$
- Highest recall for $CCR \geq 0$
- Strong recall drop
**Compound Identification Experiments**

- Precision highest with $CCR=4$
- Highest recall for $CCR \geq 0$
- Strong recall drop
- $F_1$ highest for $CCR \geq 1$
**Compound Identification Experiments**

- $F_1$ highest for $\text{CCR} \geq 1$
- Upper bound is also moderate
**Compound Identification Experiments**

- $F_1$ highest for $CCR \geq 1$
- Upper bound is also moderate

- $\checkmark$ CCR precise
- $\times$ Over-restricted
Nominal Compound Database

Precise identification

⇒
High quality compound collection

✓ Useful resource for NLP

→ Cross-lingual compound parsing
Ziering and Van der Plas (2014/2015a,b)

✓ Provides many parallel examples for linguistics studies

→ Deverbal compounds in English and Romanian
Iordachioaia (2017)

✓ Resource for CCR ≥ 0 to CCR = 4 available
2. **Compound Parsing**

### The Task
Determine the tree structure of a compound

### Limitations in Previous Approaches
- **Barker (1998)** (semi-automatic, NPs)
- **Lauer (1994)** (knowledge-rich, 3NCs)
- **Pitler et al. (2010)** (supervised, NPs)

### Publications related to the PhD thesis
- Ziering and Van der Plas (COLING, 2014)
- Ziering and Van der Plas (IWCS, 2015)
- Ziering and Van der Plas (RANLP, 2015)
2. **Compound Parsing**

**The Task**
Determine the tree structure of a compound

**Limitations in Previous Approaches**
- **Barker (1998)** (semi-automatic, NPs)
- **Lauer (1994)** (knowledge-rich, 3NCs)
- **Pitler et al. (2010)** (supervised, NPs)

**Publications related to the PhD thesis**
- Ziering and Van der Plas (COLING, 2014)
- Ziering and Van der Plas (IWCS, 2015)
- Ziering and Van der Plas (RANLP, 2015)
CROSS-LINGUAL COMPOUND PARSING METHOD

- **Deterministic Bottom-Up Parsing (DBUP)**
- Iteratively merges constituents with smallest Aligned Word Distance (AWD)
**CROSS-LINGUAL COMPOUND PARSING METHOD**

- **Deterministic Bottom-Up Parsing (DBUP)**
- Iteratively merges constituents with smallest Aligned Word Distance (AWD)

**Bottom-up Parser**

**Word Alignment**

**Aligned Word Distance**

- twin
- pipe
- undersea
- gas
- pipeline

- onderzeese
- gaspijpleiding
- met
- dubbele
- pijp

- 0
**Cross-lingual Compound Parsing Method**

- **Deterministic Bottom-Up Parsing (DBUP)**
- Iteratively merges constituents with smallest Aligned Word Distance (AWD)
**CROSS-LINGUAL COMPOUND PARSING METHOD**

- **Deterministic Bottom-Up Parsing (DBUP)**
- Iteratively merges constituents with smallest Aligned Word Distance (AWD)

[Diagram showing word alignment and aligned word distance]
**Cross-lingual Compound Parsing Method**

- **Deterministic Bottom-Up Parsing (DBUP)**
- Iteratively merges constituents with smallest Aligned Word Distance (AWD)
Cross-lingual Compound Parsing Method

- **Deterministic Bottom-Up Parsing (DBUP)**
- Iteratively merges constituents with smallest Aligned Word Distance (AWD)
**Cross-lingual Compound Parsing Method**

- **Deterministic Bottom-Up Parsing (DBUP)**
- Iteratively merges constituents with smallest Aligned Word Distance (AWD)

```
<table>
<thead>
<tr>
<th>twin pipe undersea gas pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>twin pipe</td>
</tr>
<tr>
<td>undersea gas pipeline</td>
</tr>
<tr>
<td>twin</td>
</tr>
<tr>
<td>pipe</td>
</tr>
<tr>
<td>undersea</td>
</tr>
<tr>
<td>gas</td>
</tr>
<tr>
<td>pipeline</td>
</tr>
<tr>
<td>onderzeese</td>
</tr>
<tr>
<td>gaspijpleiding</td>
</tr>
<tr>
<td>met</td>
</tr>
<tr>
<td>dubbele</td>
</tr>
<tr>
<td>pijp</td>
</tr>
</tbody>
</table>
```

**Bottom-up Parser**

**Word Alignment**

**Aligned Word Distance**
Compound Parsing Experiments

- Test set: 278 three-noun compound tokens from Europarl → IAA rate of 90.3% (on first half)
- $\chi^2$: unsupervised approach by Nakov and Hearst (2005)
- Cross-lingual evidence is sparse: (smaller coverage)
  ⇒ Backoff model: DBUP → $\chi^2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBUP → $\chi^2$</td>
<td>93.5%</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>87.4%</td>
</tr>
<tr>
<td>Majority class baseline</td>
<td>80.9%</td>
</tr>
</tbody>
</table>
Compound Parsing Experiments

- Test set: 278 three-noun compound tokens from Europarl → IAA rate of 90.3% (on first half)
- $\chi^2$: unsupervised approach by Nakov and Hearst (2005)
- Cross-lingual evidence is sparse: (smaller coverage)
  ⇒ Backoff model: DBUP $\rightarrow \chi^2$

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBUP $\rightarrow \chi^2$</td>
<td>93.5%</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>87.4%</td>
</tr>
<tr>
<td>Majority class baseline</td>
<td>80.9%</td>
</tr>
</tbody>
</table>

→ DBUP $\rightarrow \chi^2$ performs best
→ comparable to human performance
Outline of the Talk

1. **Introduction**

Main Contributions

2. **New Insights on the Notion of Compoundhood**

3. **Cross-lingual Evidence**

4. **Morphological Regularities**

5. **Summary and Conclusions**
**Morphological Regularities**

**Monolingual Approach to Indirect Supervision**

→ Avoid knowledge about compound morphology (e.g., linking elements)
→ Exploit morphological regularities
→ Based on monolingual lemmatized data

⇓

**Compound splitting**
Constituent Inflection

‘Compound-specific inflection’ (Lieber and Štekauer, 2009)

constituent inflection (vs. word inflection)
Constituent Inflection

‘Compound-specific inflection’ (Lieber and Štekauer, 2009)
constituent inflection (vs. word inflection)

Huhn ‘chicken’
\[u \rightarrow ü \downarrow \ominus \text{er}\]
\[Hühner\text{suppe}\]
‘Compound-specific inflection’ (Lieber and Štekauer, 2009)

**Constituent inflection** (vs. **word inflection**)

- **Huhn** ‘chicken’
  - u → ü
  - Hühnersuppe

- **aalbes** ‘currant’
  - ↓ ⊕ sen
  - aalbessensap ‘currant juice’
Constituent Inflection

‘Compound-specific inflection’ (Lieber and Štekauer, 2009)

Constituent inflection (vs. word inflection)

Huhn ‘chicken’

\[\text{u} \rightarrow \text{ü} \Downarrow \otimes \text{er} \]

Hühnersuppe

aalbes ‘currant’

\[\Downarrow \oplus \text{sen} \]

aalbessen sap ‘currant juice’

frontier

\[\Downarrow \oplus \text{s} \]

frontiersman
Constituent Inflection

Linguistic Theories

German and Dutch linking elements stem from case and number endings (Neef, 2009, Booij & v. Santen, 1998)

- German: Huhn → Hühner_{plural}
- Dutch: aalbes → aalbessen_{plural}

Indirect Supervision for Constituent Inflection

- Usage of number and case endings
  ⇒ Constituent inflection ≈ word inflection

  - manual-resource-lean
  - cover non-paradigmatic cases
    - Liebesbrief ‘love letter’
**Morphological Operation Patterns (MOPs)**

<table>
<thead>
<tr>
<th>MOPs</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>u → ü, +er</td>
<td>&lt;Huhn, Hühner&gt; ‘chicken’</td>
</tr>
<tr>
<td>+en</td>
<td>&lt;arts, artsen&gt; ‘doctor’</td>
</tr>
</tbody>
</table>

→ MOPs derived from word inflection: **Word MOPs**
### Morphological Operation Patterns (MOPs)

<table>
<thead>
<tr>
<th>MOPs</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>u → ü, ⊕er</td>
<td>&lt;Huhn, Hühner&gt; ‘chicken’</td>
</tr>
<tr>
<td>⊕en</td>
<td>&lt;arts, artsen&gt; ‘doctor’</td>
</tr>
</tbody>
</table>

→ MOPs derived from word inflection: **Word MOPs**

### Functional Application

- MOPs can be applied to lemmas (e.g., for generation)
  \[ u \rightarrow ü, \oplus er \] \( Huhn \) = \( Hühner \)

- Inverse application to forms (e.g., for **compound splitting**)
  \[ ü \rightarrow u, \ominus er \] \( Hühner \) = \( Huhn \)
**The Task**

Identify constituents in closed compounds

**SP:** Hühner | suppe  
**Norm:** Huhn + Suppe

**Limitations in Previous Approaches**

- Fritzinger and Fraser (2010) (SMOR)
- Weller and Heid (2012) (transformation rules)

**Publications related to the PhD thesis**

⇒ Ziering and Van der Plas (NAACL, 2016)
- Ziering and Müller and van der Plas (MWE, 2016)
- Jagfeld and Ziering and van der Plas (ACL, 2017)
**The Task**

Identify constituents in closed compounds

**SP:** Hühner | suppe  
**Norm:** Huhn + Suppe

**Limitations in Previous Approaches**

- Fritzinger and Fraser (2010) (SMOR)
- Weller and Heid (2012) (transformation rules)

**Publications related to the PhD thesis**

⇒ Ziering and Van der Plas (NAACL, 2016)

- Ziering and Müller and van der Plas (MWE, 2016)
- Jagfeld and Ziering and van der Plas (ACL, 2017)
Compound Splitting with Word MOPs

Hühnersuppenrezept
‘chicken soup recipe’
**Compound Splitting with Word MOPs**

- **Hühnersuppenrezept**
  - ‘chicken soup recipe’

  **binary splitting**

  - **Hü** | **hnersuppenrezept**
  - **Hühner** | **suppenrezept**
  - **Hühnersuppen** | **rezept**
  - **Hühnersuppenreze** | **pt**

  **Normalization using Word MOPs**

  - *Hühnersuppe* + *Rezept*
  - highest scored analysis

  **Recursive process**

  - 97 / 111
**Compound Splitting with Word MOPs**

*Hühnersuppenrezept*

‘*chicken soup recipe*’

**binary splitting**

\[
\begin{align*}
Hü & | \quad hnersuppenrezept \\
Hühner & | \quad suppenrezept \\
Hühnersuppen & | \quad rezept \\
Hühnersuppenreze & | \quad pt
\end{align*}
\]

\[
\begin{align*}
Huhn & + \quad Suppenrezept \\
Hühnersuppe & + \quad Rezept
\end{align*}
\]

normalization using Word MOPs
**Compound Splitting with Word MOPs**

*Hühnersuppenrezept*  
‘chicken soup recipe’  

**Binary Splitting**

1. **Hü | hnersuppenrezept**  
2. **Hühner | suppenrezept**  
3. **Hühnersuppen | rezept**  
4. **Hühnersuppenreze | pt**

**Normalization using Word MOPs**  

- **Hühnersuppe ✓ + Rezept ✓**  
- Highest scored analysis
**Compound Splitting with Word MOPs**

- **Hühnersuppenrezept**
  - ‘chicken soup recipe’

**Binary Splitting**

- **Huhn** | **Suppenrezept**
- **Hühner** | **Suppenrezept**
- **Hühnersupper** | **Rezept**
- **Hühnersuppen** | **Rezept**

**Normalization using Word MOPs**

- **Hühnersuppe** ✓ + **Rezept** ✓
- highest scored analysis
- recursive process

**Example**:

- **Hühnersuppenrezept**
- **Hühnersuppe** + **Rezept**
### Compound Splitting Experiments (Deutsch)

<table>
<thead>
<tr>
<th>System</th>
<th>SPX</th>
<th>NormX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P$</td>
<td>$R$</td>
</tr>
<tr>
<td>ZvdP2016</td>
<td>98.1</td>
<td>96.9</td>
</tr>
<tr>
<td>FF2010</td>
<td>98.5</td>
<td>92.3</td>
</tr>
<tr>
<td>WH2012</td>
<td>98.1</td>
<td>96.8</td>
</tr>
</tbody>
</table>

- Splitting gold standard ([Henrich and Hinrichs, 2011](#))
- Split point detection (**SPX**) and normalization (**NormX**)

---
### Compound Splitting Experiments

<table>
<thead>
<tr>
<th>System</th>
<th>SPX</th>
<th></th>
<th>NormX</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P$</td>
<td>$R$</td>
<td>$Acc$</td>
<td>$F_1$</td>
</tr>
<tr>
<td>ZvdP2016</td>
<td>98.1</td>
<td>96.9</td>
<td>97.5</td>
<td></td>
</tr>
<tr>
<td>FF2010</td>
<td>98.5</td>
<td>92.3</td>
<td>95.3</td>
<td></td>
</tr>
<tr>
<td>WH2012</td>
<td>98.1</td>
<td>96.8</td>
<td>97.4</td>
<td></td>
</tr>
</tbody>
</table>

- Splitting gold standard ([Henrich and Hinrichs, 2011](#))
- Split point detection (SPX) and normalization (NormX)
- Manual-resource-rich state-of-the-art approaches
### Compound Splitting Experiments

<table>
<thead>
<tr>
<th>System</th>
<th>SPX</th>
<th>NormX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>ZvdP2016</td>
<td>98.1</td>
<td>96.9</td>
</tr>
<tr>
<td>FF2010</td>
<td>98.5</td>
<td>92.3</td>
</tr>
<tr>
<td>WH2012</td>
<td>98.1</td>
<td>96.8</td>
</tr>
</tbody>
</table>

- Splitting gold standard ([Henrich and Hinrichs, 2011](#))
- Split point detection (SPX) and normalization (NormX)
- Manual-resource-rich state-of-the-art approaches
- ZvdP2016 competes (even sometimes outperforms) in SPX
## Compound Splitting Experiments

<table>
<thead>
<tr>
<th>System</th>
<th>SPX</th>
<th>NormX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>ZvdP2016</td>
<td>98.1</td>
<td>96.9</td>
</tr>
<tr>
<td>FF2010</td>
<td>98.5</td>
<td>92.3</td>
</tr>
<tr>
<td>WH2012</td>
<td>98.1</td>
<td>96.8</td>
</tr>
</tbody>
</table>

- Splitting gold standard ([Henrich and Hinrichs, 2011](#))
- Split point detection (SPX) and normalization (NormX)
- Manual-resource-rich state-of-the-art approaches
  - ZvdP2016 competes (even sometimes outperforms) in SPX
  - Manual-resource-rich outperforms in normalization
**COMPONENT SPLITTING EXPERIMENTS (🇳🇱 / 🇿🇦)**

→ Multilingual applicability: Dutch and Afrikaans

<table>
<thead>
<tr>
<th>System</th>
<th>Language</th>
<th>SP Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZvdP2016</td>
<td>🇳🇱</td>
<td>94.6%</td>
</tr>
<tr>
<td>Verh2014</td>
<td>🇿🇦</td>
<td>82.9%</td>
</tr>
</tbody>
</table>

- Only SP Acc (published numbers of supervised approach)
- Splitting gold standard *(Verhoeven et al. 2014)*
Multilingual applicability: Dutch and Afrikaans

<table>
<thead>
<tr>
<th>System</th>
<th>Language</th>
<th>SP Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZvdP2016</td>
<td>🇳🇱</td>
<td>94.6%</td>
</tr>
<tr>
<td>Verh2014</td>
<td>🇿🇦</td>
<td>82.9%</td>
</tr>
<tr>
<td></td>
<td>🇳🇱</td>
<td>91.5%</td>
</tr>
<tr>
<td></td>
<td>🇿🇦</td>
<td>88.3%</td>
</tr>
</tbody>
</table>

- Only $SP_{Acc}$ (published numbers of supervised approach)
- Splitting gold standard (*Verhoeven et al.* 2014)
- Better for Dutch
- Worse for Afrikaans

→ Afrikaans training corpus is too small
Outline of the Talk

1. Introduction

Main Contributions

2. New Insights on the Notion of Compoundhood

3. Cross-lingual Evidence

4. Morphological Regularities

5. Summary and Conclusions
Most linguistic criteria have only limited predictive power
⇒ promising feature: cross-lingual spelling
Summary and Conclusions

- Most linguistic criteria have only limited predictive power ⇒ promising feature: cross-lingual spelling

- Only moderate agreement in the notion of compoundhood ⇒ in line with previous work
Summary and Conclusions

- Most linguistic criteria have only limited predictive power ⇒ promising feature: cross-lingual spelling

- Only moderate agreement in the notion of compoundhood ⇒ in line with previous work

- Cross-lingual supervision
  - ✓ compound parsing
  - ✗ compound identification (precise but too restrictive)
Summary and Conclusions

Most linguistic criteria have only limited predictive power ⇒ promising feature: cross-lingual spelling

Only moderate agreement in the notion of compoundhood ⇒ in line with previous work

Cross-lingual supervision
  ✓ compound parsing
  ✗ compound identification (precise but too restrictive)

Morphological regularities
  ✓ Competitive to manual-resource-rich state-of-the-art compound splitters
Thank you for your attention!