Interfixes as boundary markers in compounds

Harald Baayen\textsuperscript{1}, Andrea Krott\textsuperscript{2}, Wolfgang Dressler\textsuperscript{3}, Robert Schreuder\textsuperscript{1}, Gonia Jarema\textsuperscript{4}, and Gary Libben\textsuperscript{5}

Nijmegen\textsuperscript{1}, Birmingham\textsuperscript{2}, Vienna\textsuperscript{3}, Montreal\textsuperscript{4}, Edmonton\textsuperscript{5}. 
Triconstituent compounds

- Dutch and German have many tri-constituent compounds.
Dutch and German have many tri-constituent compounds.

Consider the following examples from Dutch:

- *aardolieproduct* → ‘petrochemical product’
- *gevarendriehoek* → ‘emergency triangle’
- *huisartsenpraktijk* → ‘family doctor’s office’
Dutch and German have many tri-constituent compounds.

Consider the following examples from Dutch:

- *aardolieproduct* ‘petrochemical product’
- *gevarenendrieck* ‘emergency triangle’
- *huisartsenpraktijk* ‘family doctor’s office’
Dutch and German have many tri-constituent compounds.

Consider the following examples from Dutch:

- aardolieproduct: ‘petrochemical product’
- gevaren(drie)hoek: ‘emergency triangle’
- huisartsenpraktijk: ‘family doctor’s office’

The question we address is whether interfixes such as en and s are useful for parsing such triconstituent compounds into their immediate constituents.
Immediate Constituent Structure

- Many triconstituent compounds are left branching:
Immediate Constituent Structure

- Many triconstituent compounds are left branching:
- ‘petrochemical product’:  
  \(aard+olie+product\) (‘earth oil product’)

Interfixes as boundary markers in compounds – p.3/18
Immediate Constituent Structure

- Many triconstituent compounds are left branching:
- ‘petrochemical product’: 
  \[aard+olie+product\] (‘earth oil product’)
- But triconstituent compounds can also be right branching:
Immediate Constituent Structure

- Many triconstituent compounds are left branching:
- ‘petrochemical product’: 
  \( aard + olie + product \) (‘earth oil product’)
- But triconstituent compounds can also be right branching:
- ‘old car dump’: 
  \( auto + kerk + hof \) (‘car church yard’)

Interfixes as boundary markers in compounds – p.3/18
Grimm, and recently Fuhrhop (1998) argued that interfixes in left-branching compounds signal the major constituent boundary:
Grimm, and recently Fuhrhop (1998) argued that interfixes in left-branching compounds signal the major constituent boundary:

‘family doctor’s office’

C1  C2  C3

huis  arts  praktijk

‘house’  ‘doctor’  ‘practice’
Grimm, and recently Fuhrhop (1998) argued that interfixes in left-branching compounds signal the major constituent boundary:

‘family doctor’s office’

\[
\begin{array}{ccc}
C1 & C2 & C3 \\
\text{huis} & \text{arts} & \text{en} & \text{praktijk} \\
\text{‘house’} & \text{‘doctor’} & \text{BOUNDARY} & \text{‘practice’}
\end{array}
\]
Grimm, and recently Fuhrhop (1998) argued that interfixes in left-branching compounds signal the major constituent boundary:

‘family doctor’s office’

\[ \begin{array}{ccc}
C1 & C2 & C3 \\
\text{huis} & \text{arts} & \text{en} & \text{praktijk} \\
\text{‘house’} & \text{‘doctor’} & \text{BOUNDARY} & \text{‘practice’} \\
\end{array} \]

We report lexical statistics and some first experimental results addressing this separation hypothesis.
Types

Interfixes as boundary markers in compounds – p.5/18
Exponential decrease in type counts as the number of simple constituents increases.
Interfixation Rates

The diagram illustrates the proportion of interfixes in German and Dutch compounds with different numbers of constituents. The x-axis represents the number of constituents (2, 3, 4), while the y-axis shows the proportion with interfixes. The graph shows that the number of constituents affects the use of interfixes, with higher proportions in German compared to Dutch compounds.
There is a higher interfixation rate for tri-constituent compounds compared to bi-constituent compounds.
Overview of Asymmetries

- The higher interfixation rate in tri-constituent compounds turns out to be significant, even when we take into account that there are more slots available for interfixation in tri-constituent compounds.
The higher interfixation rate in tri-constituent compounds turns out to be significant, even when we take into account that there are more slots available for interfixation in tri-constituent compounds.

In fact, there are surprisingly few tri-constituent compounds with both possible interfixation slots filled.
The higher interfixation rate in tri-constituent compounds turns out to be significant, even when we take into account that there are more slots available for interfixation in tri-constituent compounds.

In fact, there are surprisingly few tri-constituent compounds with both possible interfixation slots filled.

\[
\begin{align*}
C1 & \quad SLOT1 & \quad C2 & \quad SLOT2 & \quad C3 \\
\text{besluit} & \quad \text{vorming} & \quad s & \quad \text{fase}
\end{align*}
\]
Overview of Asymmetries

- The higher interfixation rate in tri-constituent compounds turns out to be significant, even when we take into account that there are more slots available for interfixation in tri-constituent compounds.

- In fact, there are surprisingly few tri-constituent compounds with both possible interfixation slots filled.

```
C1     SLOT1    C2     SLOT2    C3
besluit  vorming    s    fase
```
The higher interfixation rate in tri-constituent compounds turns out to be significant, even when we take into account that there are more slots available for interfixation in tri-constituent compounds.

In fact, there are surprisingly few tri-constituent compounds with both possible interfixation slots filled.

- **C1**
  - **SLOT1**: besluit
  - **C2**: vorming
  - **SLOT2**: s
  - **C3**: fase
- **vrede**: s
- **vraag**: stuk
Overview of Asymmetries

- The higher interfixation rate in tri-constituent compounds turns out to be significant, even when we take into account that there are more slots available for interfixation in tri-constituent compounds.

- In fact, there are surprisingly few tri-constituent compounds with both possible interfixation slots filled.

<table>
<thead>
<tr>
<th>C1</th>
<th>SLOT1</th>
<th>C2</th>
<th>SLOT2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>besluit</td>
<td>vorming</td>
<td>s</td>
<td>fase</td>
<td>stuk</td>
</tr>
<tr>
<td>vrede</td>
<td>s</td>
<td>vraag</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Interfixes as boundary markers in compounds – p.7/18
Overview of Asymmetries

- The higher interfixation rate in tri-constituent compounds turns out to be significant, even when we take into account that there are more slots available for interfixation in tri-constituent compounds.

- In fact, there are surprisingly few tri-constituent compounds with both possible interfixation slots filled.

<table>
<thead>
<tr>
<th>C1</th>
<th>SLOT1</th>
<th>C2</th>
<th>SLOT2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>besluit</td>
<td>vorming</td>
<td>s</td>
<td>fase</td>
<td></td>
</tr>
<tr>
<td>vrede</td>
<td>s</td>
<td>vraag</td>
<td>s</td>
<td>stuk</td>
</tr>
<tr>
<td>krijg</td>
<td>s</td>
<td>man</td>
<td>s</td>
<td>eer</td>
</tr>
</tbody>
</table>
Overview of Asymmetries

- The higher interfixation rate in tri-constituent compounds turns out to be significant, even when we take into account that there are more slots available for interfixation in tri-constituent compounds.

- In fact, there are surprisingly few tri-constituent compounds with both possible interfixation slots filled.

<table>
<thead>
<tr>
<th>C1</th>
<th>SLOT1</th>
<th>C2</th>
<th>SLOT2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>besluit</td>
<td>vorming</td>
<td>s</td>
<td>fase</td>
<td></td>
</tr>
<tr>
<td>vrede</td>
<td>s</td>
<td>vraag</td>
<td>stuk</td>
<td></td>
</tr>
<tr>
<td>krijg</td>
<td>s</td>
<td>man</td>
<td>s</td>
<td>eer</td>
</tr>
</tbody>
</table>
Many triconstituent compounds contain a bi-constituent compound that itself is a well-known word of the language.
Many triconstituent compounds contain a bi-constituent compound that itself is a well-known word of the language.

\[
\begin{array}{ccc}
C1 & SLOT1 & C2 & SLOT2 & C3 \\
\text{besluit} & \text{vorming} & s & \text{fase} \\
\end{array}
\]
Existing Embedded Compounds

Many triconstituent compounds contain a bi-constituent compound that itself is a well-known word of the language.

\[ \begin{array}{cccc}
C1 & SLOT1 & C2 & SLOT2 & C3 \\
besluit & vorming & s & fase \\
vrede & s & vraag & stuk \\
\end{array} \]
Many triconstituent compounds contain a bi-constituent compound that itself is a well-known word of the language.

<table>
<thead>
<tr>
<th>C1</th>
<th>SLOT1</th>
<th>C2</th>
<th>SLOT2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>besluit</td>
<td>vorming</td>
<td>s</td>
<td>fase</td>
<td></td>
</tr>
<tr>
<td>vrede</td>
<td>s</td>
<td>vraag</td>
<td>stuk</td>
<td></td>
</tr>
<tr>
<td>krijg</td>
<td>s</td>
<td>man</td>
<td>s</td>
<td>eer</td>
</tr>
</tbody>
</table>
Many triconstituent compounds contain a bi-constituent compound that itself is a well-known word of the language.

To what extent might the structure of tri-constituent compounds be given simply by the bi-constituent compounds that the lexicon makes available?

<table>
<thead>
<tr>
<th>C1</th>
<th>SLOT1</th>
<th>C2</th>
<th>SLOT2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>besluit</td>
<td>vorming</td>
<td>s</td>
<td>fase</td>
<td></td>
</tr>
<tr>
<td>vrede</td>
<td>s</td>
<td>vraag</td>
<td></td>
<td>stuk</td>
</tr>
<tr>
<td>krijg</td>
<td>s</td>
<td>man</td>
<td>s</td>
<td>eer</td>
</tr>
</tbody>
</table>
Embedded Compounds: German

interfixes as boundary markers in compounds – p.9/18
The existing bi-constituent compounds almost always provide the correct left or right branch.
The high interfixation rate for tri-constituent compounds and the scarcity of tri-constituent compounds with two interfixes argues in favor of the separation hypothesis.
The high interfixation rate for tri-constituent compounds and the scarcity of tri-constituent compounds with two interfixes argues in favor of the separation hypothesis.

However, embedded bi-constituent compounds almost always provide the immediate constituent structure for free. Bi-constituent compounds straddling the main boundary are rare and infrequent. This argues against the separation hypothesis.
Positional Distribution: German

- Proportion of types
- Left Branching
- Right Branching

Interfixes as boundary markers in compounds – p.11/18
Left and right branches have few internal interfixes. Interfixes typically appear at the major boundary.
Positional Distribution: Dutch

Interfixes as boundary markers in compounds – p.12/18
In Dutch, this pattern is restricted to the -s- interfix.
Key Question

- The distributional properties of affixes argue in favor of the separation hypothesis, not only for left-branching compounds (Grimm, Fuhrhop), but also for right-branching compounds.
The distributional properties of affixes argue in favor of the separation hypothesis, not only for left-branching compounds (Grimm, Fuhrhop), but also for right-branching compounds.

The availability in the lexicon of most left and right branches argues against the separation hypothesis.
Key Question

- The distributional properties of affixes argue in favor of the separation hypothesis, not only for left-branching compounds (Grimm, Fuhrhop), but also for right-branching compounds.

- The availability in the lexicon of most left and right branches argues against the separation hypothesis.

- Moreover, the choice of the interfix is largely determined by the distributions of interfixes in the left and right constituent families.
Key Question

- The distributional properties of affixes argue in favor of the separation hypothesis, not only for left-branching compounds (Grimm, Fuhrhop), but also for right-branching compounds.

- The availability in the lexicon of most left and right branches argues against the separation hypothesis.

- Moreover, the choice of the interfix is largely determined by the distributions of interfixes in the left and right constituent families.

- Might German and Dutch speakers nevertheless be sensitive to the present distributional branching asymmetries?
Materials

- Tri-constituent compounds consisting of three monomorphemic pseudo-stems with distinct coda-onset transitions (e.g., *bluig-froep-taafl*).
Materials

- Tri-constituent compounds consisting of three monomorphemic pseudo-stems with distinct coda-onset transitions (e.g., bluig-froep-taaf).

- Four conditions: no interfix, an interfix in slot 1, and interfix in slot 2, and interfixes in both slots.
Materials

- Tri-constituent compounds consisting of three monomorphemic pseudo-stems with distinct coda-onset transitions (e.g., *bluig-froep-taaf*).
- Four conditions: no interfix, an interfix in slot 1, and interfix in slot 2, and interfixes in both slots.
- The interfix was always an *-s-*. 
Materials

- Tri-constituent compounds consisting of three monomorphemic pseudo-stems with distinct coda-onset transitions (e.g., *bluig-froep-taaf*).
- Four conditions: no interfix, an interfix in slot 1, and interfix in slot 2, and interfixes in both slots.
- The interfix was always an -s-.
- The experiment was run in German and in Dutch.
We asked participants to indicate how they would hyphenate these tri-constituent compounds.
Task

- We asked participants to indicate how they would hyphenate these tri-constituent compounds.
- The compounds were presented on paper, with slight spacing between the letters.
Task

- We asked participants to indicate how they would hyphenate these tri-constituent compounds.
- The compounds were presented on paper, with slight spacing between the letters.
- Participants were asked to indicate by means of a vertical slash where they would hyphenate the compounds.
Task

- We asked participants to indicate how they would hyphenate these tri-constituent compounds.
- The compounds were presented on paper, with slight spacing between the letters.
- Participants were asked to indicate by means of a vertical slash where they would hyphenate the compounds.
- Are their hyphenation responses co-determined by the presence of an interfix?
Experiment: German

Interfixes as boundary markers in compounds – p.16/18
Hyphenation responses follow the interfix. The same pattern was observed for Dutch.
Conclusions

- Interfixes tend to occur at major constituent boundaries.
Conclusions

- Interfixes tend to occur at major constituent boundaries.
- Dutch and German speakers appear to be sensitive to this distributional pattern.
Conclusions

- Interfixes tend to occur at major constituent boundaries.
- Dutch and German speakers appear to be sensitive to this distributional pattern.
- The separation hypothesis is supported, even though the choice of the linker is predominantly determined analogically, and even though the embedded existing bi-constituent compounds provide the constituent structure for free.
Conclusions

- Interfixes tend to occur at major constituent boundaries.
- Dutch and German speakers appear to be sensitive to this distributional pattern.
- The separation hypothesis is supported, even though the choice of the linker is predominantly determined analogically, and even though the embedded existing bi-constituent compounds provide the constituent structure for free.
- These results support the hypothesis that language users make use of whatever probabilistic distributional information is available to optimize language processing.
The End

Thank you to the Linux and LEDE developers.