L1 transfer versus fixed chunks: A learner corpus-based study on L2 German

Margit Breckle¹, Maike Müller², Melanie Seiss¹ & Heike Zinsmeister²

Bildungswissenschaftliche Universität Litauens¹ & Universität Konstanz²

Motivation

How do Chinese L2 learners of German start a declarative sentence?

- Previous studies revealed a significant overuse of information-structural functions (Zinsmeister & Breckle 2010)
- Explanations of the overuse:
  1. L2 learners transfer information-structural concepts from their L1.
  2. L2 learners extensively use ‘chunks’ (formulaic language).
- Research question:
  What influence does chunking have on L2 texts
  (i) in general?
  (ii) on sentence beginnings?

The ALeSKo corpus

- Multi-level annotated learner corpus
- Argumentative essays of Chinese L2 learners (-62) and German L1 speakers
- L2 sub-corpus (4 texts, 11,716 tokens) consists of
  1. L2 macro (4 texts, 5,914 tokens)
  2. L2 micro (18 texts, 5,802 tokens)

Transfer of information structure

- Background:
  - Chinese is a topic-prominent language
  - German is a verb-second language
- Results:
  - No significant differences in the relative use of information-structural categories in the prefactual
  - Slight preference for topic in the L2 data
- Conclusion:
  - chunks may contain information-structural concepts themselves
  - focus on chunk analysis

Automatic detection of copied material from the title/task

- Only all -chunks were evaluated
- Function words ignored (i.e., aus dem, 'of the life', sie einer, 'the they are')
- Example:

  **Title**
  Ist Urlaub die vergebliche Flucht aus dem Alltag?
  Are holidays an unsuccessful escape from everyday life?

  **String**
  dass Urlaub nicht die vergebliche Flucht ist
  'dass Urlaub nicht die vergebliche Flucht ist'

  **3-gram candidates**
  - vergebliche Flucht
  - vergebliche Flucht

  **Chunk**
  - die vergebliche Flucht

Manual classification of chunk types

- Most common sequences of words (2-, 3-, 4-grams) sorted by frequency
- No context provided for annotators
- Two/three independent annotators
- Chunk categories:
  - all – the whole n-gram is part of the title or the task description
    - e.g. für die Gesellschaft ('to the society')
  - w – the n-gram contains one or more words of the title or the task description
    - e.g. die Gesellschaft/und die ('the society')
  - t – only for the L2 texts: n-gram contains a (part of a) chunk that was taught in the L2 class prior to the essay writing
    - e.g. Verarmung der (immaterialer interpersonal relations)
  - p – chunking (idioms, collocations, patterns, and sentence frames that are memorised, rather stable, and possibly target language-like)
    - e.g. meiner Meinung nach ('in my opinion')
  - na – not applicable (i.e., sequences that are not formulaic)

Results

- Fewer copies from the title/task description (all, w, mixed) in the L1 n-grams
- Less other chunks and unrelated n-grams (na) in the L2 n-grams
- Significant less copied material in L2 prefields that in L1 prefields
- Substantial annotator agreement of Cohen’s κ = .8
  - in L2 texts κ = .86
  - in L1 texts κ = .72

Automatic classification of chunk types

- Type and coverage of the top-ranked chunks (manual annotation)

  - More copied material in L2 texts in general
  - Significant less copied material in L2 prefields that in L1 prefields
  - Ratio of copied material in the prefactual higher in L1 corpora
  - L2 learners use chunks (copied material) more often in general but do not oversaturate in the prefactual
  - L1 authors use chunks more often to start a sentence than L2 learners

Summary

- Explanation "L2 learners extensively use chunks" for overuse of sentences starting with an information-structural function disconfirmed
- A more efficient way to identify chunks of the class ‘na’ needed
- Automatic determination of collocations in a reference corpus of argumentative writing

References

- Fewer copies from the title/task description (all, w, mixed) in the L1 n-grams
- Less other chunks and unrelated n-grams (na) in the L2 n-grams
- Significant less copied material in L2 prefields that in L1 prefields
- Substantial annotator agreement of Cohen’s κ = .8
  - in L2 texts κ = .86
  - in L1 texts κ = .72

Poster based on:


Figure 1: Proportion of information structure-related prefactuals (dark grey) and unrelated ones (light grey)

Table 1: Ten most frequent 3-grams in L1 and L2 learner

<table>
<thead>
<tr>
<th>3-gram</th>
<th>freq</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>dass sich die</td>
<td>6</td>
<td>na</td>
</tr>
<tr>
<td>ob sich Kriminalität</td>
<td>6</td>
<td>w</td>
</tr>
<tr>
<td>Kriminalität nicht ausgezahlt</td>
<td>5</td>
<td>all</td>
</tr>
<tr>
<td>Kriminalität hat der Frauen mehr geschadet als genutzt</td>
<td>5</td>
<td>all</td>
</tr>
<tr>
<td>auf lange Sicht</td>
<td>4</td>
<td>o</td>
</tr>
<tr>
<td>in Verbrechen begangen</td>
<td>4</td>
<td>o</td>
</tr>
<tr>
<td>euch für ihn ausgezahlt</td>
<td>4</td>
<td>na</td>
</tr>
<tr>
<td>für ihn ausgezahlt</td>
<td>4</td>
<td>w</td>
</tr>
<tr>
<td>für ihn ausgezahlt</td>
<td>4</td>
<td>w</td>
</tr>
</tbody>
</table>

Table 2: Frequencies and coverage of data presented in Fig. 2

<table>
<thead>
<tr>
<th>Type</th>
<th>Coverage</th>
<th>L1 (2,112 tokens)</th>
<th>L2 (5,912 tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-grams</td>
<td>7.87%</td>
<td>3.82%</td>
<td>5.59%</td>
</tr>
<tr>
<td>3-grams</td>
<td>2.02%</td>
<td>1.54%</td>
<td>1.75%</td>
</tr>
<tr>
<td>4-grams</td>
<td>0.89%</td>
<td>0.69%</td>
<td>0.77%</td>
</tr>
<tr>
<td>5-grams</td>
<td>0.78%</td>
<td>0.50%</td>
<td>0.63%</td>
</tr>
</tbody>
</table>

Figure 2: Results of the automatic analysis of copied material in general

Figure 3: Results of the automatic analysis of copied material in the prefactual

Figure 4: Chunks types of rank 1-10 (normalised)

Table 1: Ten most frequent 3-grams in L1 and L2 learner

<table>
<thead>
<tr>
<th>3-gram</th>
<th>freq</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>dass sich die</td>
<td>5</td>
<td>all</td>
</tr>
<tr>
<td>ob sich Kriminalität</td>
<td>5</td>
<td>o</td>
</tr>
<tr>
<td>Kriminalität hat der Frauen mehr geschadet als genutzt</td>
<td>4</td>
<td>o</td>
</tr>
<tr>
<td>auf lange Sicht</td>
<td>4</td>
<td>o</td>
</tr>
<tr>
<td>in Verbrechen begangen</td>
<td>4</td>
<td>o</td>
</tr>
<tr>
<td>euch für ihn ausgezahlt</td>
<td>4</td>
<td>o</td>
</tr>
<tr>
<td>für ihn ausgezahlt</td>
<td>4</td>
<td>o</td>
</tr>
</tbody>
</table>

Table 2: Frequencies and coverage of data presented in Fig. 2

<table>
<thead>
<tr>
<th>Type</th>
<th>Coverage</th>
<th>L1 (2,112 tokens)</th>
<th>L2 (5,912 tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-grams</td>
<td>7.87%</td>
<td>3.82%</td>
<td>5.59%</td>
</tr>
<tr>
<td>3-grams</td>
<td>2.02%</td>
<td>1.54%</td>
<td>1.75%</td>
</tr>
<tr>
<td>4-grams</td>
<td>0.89%</td>
<td>0.69%</td>
<td>0.77%</td>
</tr>
<tr>
<td>5-grams</td>
<td>0.78%</td>
<td>0.50%</td>
<td>0.63%</td>
</tr>
</tbody>
</table>