ALeSKo – An annotated learner corpus

Heike Zinsmeister[1], University of Konstanz, Germany
Margit Breckle, Vilnius Pedagogical University, Lithuania

Introduction

Learner corpora collect texts of second language (L2) learners in order to provide an empirical basis for studies in second language acquisition.
Examples of research questions which can be answered by corpus-based studies:

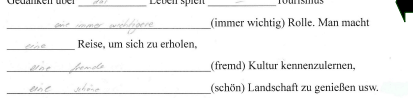
- Does the article use in L2 German correspond to the article use of L1 German? If not, how do they differ?
- Does the L2 learners' use of Vorfeld constituents differ from the use in L1 German?

Properties of learner corpora:

- provide quantitative, empirical findings
- reusable analyses
- make studies reproducible
- multi-functional

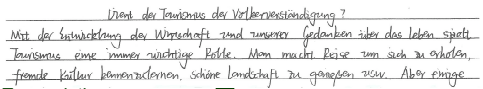
Existing freely available learner corpus of L2 German: FALKO corpus

Fill-in-the-blanks texts – article use:



From data collection to empirical studies:

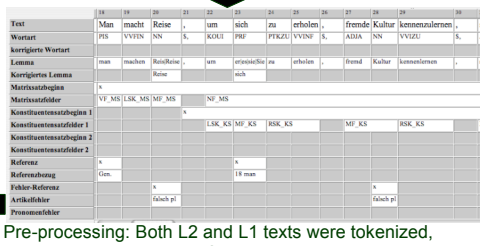
Hand-written text:



Transcription:

Dient der Tourismus der Völkerverständigung?
Mit der Entwicklung der Wirtschaft und unserer Gedanken über das Leben spielt Tourismus eine immer wichtige Rolle. Man macht Reise, um sich zu erholen, fremde Kultur kennenzulernen, schöne Landschaft zu genießen usw. Aber einige [...]

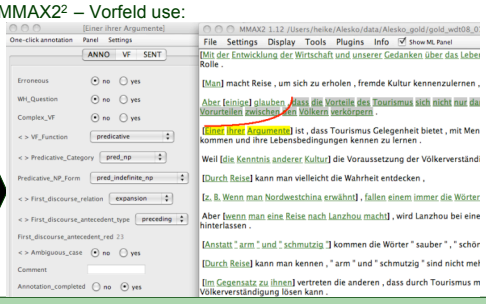
EXMARaLDA:



Pre-processing: Both L2 and L1 texts were tokenized, lemmatized and part-of-speech tagged with the TreeTagger (Schmid 1994) using the German STTS tagset (Schiller et al. 1999).

The ALeSKo corpus consists of L2 essays (Chinese L2 learners of German), L1 essays, metadata as well as detailed annotation guidelines:
wdt07: 25 L2 texts – topic: *Are holidays an unsuccessful escape from everyday life?* (6,902 tokens)
wdt08: 18 L2 texts – topic: *Does tourism support understanding among nations?* (6,685 tokens)
Falko Essays L1 0.5: 39 essays – different topics (34,155 tokens)
The ALeSKo corpus is influenced by the FALKO corpus, e. g. annotation with EXMARaLDA (Schmidt 2004), topological field annotation (Doolittle 2008).

MMAX2[2] – Vorfeld use:



Study I : Article Use

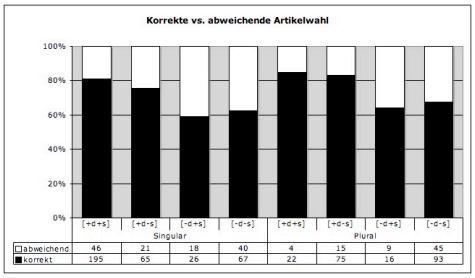
- Research questions:
  - ➢ Do Chinese L2 learners of German use the article system in the same way as German L1 speakers do – given that Chinese does not have articles?
  - ➢ Do they use it in the same way to express the semantic dimensions of specificity and definiteness or is one of these dimensions easier to grasp?
- Hypotheses (cf. Ionin 2003):

	[+definite] – target hypothesis: sg. *d-*, pl. *d-*	[–definite] – target hypothesis: sg. *ein-*, pl. null
[+specific]	correct use of sg./pl. *d-*	wrong use of sg./pl. *d-* in addition to target language-like sg. *ein-* and pl. null
[–specific]	wrong use of sg. *ein-*, pl. null in addition to target language-like sg./pl. *d-*	correct use of sg. *ein-*, pl. null

Method

- Experiment related to Ionin et al. (2007)
- Offline experiment on the basis of sub-corpus wdt08
- The texts were reformatted so that all articles were deleted and replaced by a gap / line preceding the nouns.
- The fill-in-the-blanks texts were filled in by L1 Germans (see screenshot above)
  ➡ comparison of L1 and L2 article use in *identical* contexts

Results



- Significant divergence in L1 and L2 article use
- Significant difference between L2 article use in definite vs. indefinite contexts
- Specificity does not have a significant influence on article use.

Study II: Vorfeld Use

- In German, there is a preference hierarchy for the Vorfeld (Speyer 2007):
  i. frame-setting elements / brand-new elements
  ii. elements that belong to a salient set of elements (poset)
  iii. certain previously mentioned entities (backward-looking center aka 'topic')
- Research question:
  - ➢ Which Vorfeld constituents do Chinese L2 learners of German use in comparison to L1 Germans?
- Hypotheses:
  Chinese L2 learners of German transfer the presentation of information structure from their L1 into German:
  - ➢ **H1**: overuse of backward-looking centers (unclear: frame-setting)
  - ➢ **H2:** underuse of brand-new elements and poset

Method

- Vorfeld annotation extracted from EXMARaLDA annotation
- Annotation performed as one-click-annotation in MMAX2 (see screenshot of annotation tool above)
- Annotation levels – only relevant annotation levels appear according to previous choice, e. g. Vorfeld function, information status, discourse coreference, ...
- Annotation procedure: first version of annotation > comparison of annotations (if two annotations available) > final version based on expert decision
- Annotation time (1st): about 1.5 min/markable, 30 min/L2 text, 60 min/L1 text

Results

- Study based on subcorpus: 36 L2 texts and 21 L1 texts
- Exclusion of erroneous (non-VF) markables, wh-questions, and complex VF

Ex.: **Und [Reise]** macht Menschen auch müde (wdt07_03)
and travel makes people also tired

- Comparable use of categories and function
- L2 learners do not use expletive Vorfeld *es*

Ad H2: For brand-new elements and poset elements L2 and L1 do **not** differ significantly (tendencies as predicted)
Ad H1: L2 use significantly more often a backward-looking center in Vorfeld than L1 (in general and with respect to NPs or coreferential elements)

Ex.: Durch Reisen können **sie** auch andere Kultur und Lebensstile kennenlernen .
by travelling can they also other culture and lifestyles get_to_know
[Sie] können auch ihre Kenntnisse durch Reisen erweitern. (wdt07_22)
they can also their knowledge by travelling broaden

Conclusion and Future Work

- The ALeSKo corpus provides a small but richly annotated resource for the investigation of L2 acquisition of German.
- Not yet investigated:
  - ➢ Complex Vorfeld use (e. g. Vor-Vorfeld, coordination, parentheses)
  - ➢ Discourse relations (e. g. contrast or contingency)
- Not yet annotated: use of non-Vorfeld elements
- Evaluation of texts:
  - ➢ Development of criteria for classifying readability and coherence of the texts
- Aspects of sustainability:
  - ➢ Conversion into PAULA format (Dipper 2005)
  - ➢ Make data accessible by e. g. ANNIS (Zeldes et al. 2009)
  - ➢ Integration of data in FALKO repository (Lüdeling et al. 2008)

Dipper, S. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation Schema. In Proceedings of Berliner XML-Tage 2005, 39-50.
Doolittle, S. 2008. Entwicklung und Evaluierung eines auf dem Stellungsfeldermodell basierenden syntaktischen Annotationsverfahrens für Lernerkorpora innerhalb einer Mehrebenen-Architektur mit Schwerpunkt auf schriftlichen Texten fortgeschrittener Deutschlerner. Magister's thesis. Humboldt Universität Berlin.
Ionin, T. 2003. Article semantics in Second Language Acquisition. PhD thesis. Massachusetts Institute of Technology.
Ko, H., T. Ionin & K. Wexler. 2007. The role of semantic features in the acquisition of English articles by Russian and Korean speakers. In J. M. Liceras, H. Zobl and H. Goodluck (eds.), The role of formal features in second language acquisition. Lawrence Erlbaum Associates.
Lüdeling, A., S. Doolittle, H. Hirschmann, K. Schmidt & M. Walter. 2008. Das Lernerkorpus Falko. Deutsch als Fremdsprache 2 (2008), 67-73.
Müller, C. & M. Strube. 2006. Multi-Level Annotation of Linguistic Data with MMAX2. In S. Braun, K. Kohn, J. Mukherjee (eds.) Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, 197-214.
Schiller, A., S. Teufel, C. Stöckert & C. Thielen 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht. Institut für maschinelle Sprachverarbeitung, Stuttgart.
Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the International Conference on New Methods in Language Processing.
Schmidt, T. 2004. EXMARaLDA – ein Modellierungs- und Visualisierungsverfahren für die computergestützte Transkription gesprochener Sprache. In: Buchberger, E. (ed.). Proceedings of Konvens 2004.
Speyer, A. 2007. Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. Zeitschrift für Sprachwissenschaft, 26, 83-115.
Zeldes, A., J. Ritz, A. Lüdeling & C. Chiarcos. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. In Proceedings of Corpus Linguistics 2009.

1) H.Z.'s research was partly financed by Europäischer Sozialfonds in Baden-Württemberg.
2) MMAX2: see Müller & Strube (2006).