ALeSKo – An annotated learner corpus
Heike Zinsmeister1, 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:

Pre-processing: Both L2 and L1 texts were tokenized, lemmatized and part-of-speech tagged with the TreeTagger (Schiller 1994) using the German STTS tagset (Schiller et al. 1996).

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
  [+specific] – target hypothesis: sg. ein, pl. null
  [-specific] – target hypothesis: sg. ein, pl. null

• 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)

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 questions:
  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, 2 min/markable)

Results

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

Conclusions

• 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

Conclusion and Future Work

• The ALeSKo corpus provides a small but richly annotated resource for the investigation of L2 vs. L1 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)