First publ. in: Semantic approaches in natural language processing : proceedings of the Conference on Natural Language Processing 2010 / Manfred Pinkal, Ines Rehbein, Sabine Schulte im Walde, Angelika Storrer (ed.). Saarbrücken: Universaar, 2010, pp. 181-185

Starting a sentence in L2 German – Discourse annotation of a learner corpus

Heike Zinsmeister
Department of Linguistics
University of Konstanz
Konstanz, Germany
heike.zinsmeister@uni-konstanz.de

Margit Breckle
Department of German Philology and Didactics
Vilnius Pedagogical University
Vilnius, Lithuania
margit.breckle@gmx.de

Abstract

Learner corpora consist of texts produced by second language (L2) learners.[1] We present ALeSKo, a learner corpus of Chinese L2 learners of German and discuss the multi-layer annotation of the left sentence periphery – notably the *Vorfeld*.

1 Introduction

Learner corpora consist of texts produced by foreign language (L2) learners. Normally, they are designed as *comparable corpora* which consist of pairs of monolingual corpora selected according to the same set of criteria. In the case of learner corpora, they comprise similar texts in one target language produced by speakers with different L1 backgrounds or with different L2 levels. Furthermore, for reasons of comparison the corpus can contain similar texts by L1 speakers of the target language.

There are two main approaches for investigating the data in a learner corpus (cf. Granger (2008): 267–268): (i) *contrastive interlanguage analysis* (CIA), which assumes that L2 learners use an interim language that differs from the target language in a way that can be observed quantitatively, and (ii) *computer-assisted error analysis*, in which divergences in the L2 texts are identified (and possibly also annotated) based on a target hypothesis.

The current project deals with the creation of the ALeSKo learner corpus [2], which contains texts from *Chinese L2 learners of German* and is complemented by comparable L1 German texts. Our main interest lies in the expression of *local coherence* – whether the learners acquire the linguistic means to express a smooth flow from one sentence to the next in German. In the project's current state, we carry out linguistic annotation to create a basis for a CIA of local coherence. Systematic error tagging is not yet performed.[3]

It is assumed that local coherence is mainly expressed at two levels cross-linguistically (e.g. Reinhart (1980): 168f.; (1982): 19): It is either supported by coreferential entities that play a role in a sequence of sentences (*entity-based coherence*) or it is supported by discourse relations that relate two clauses and also larger parts of the text semantically (*discourse relation-based coherence*). In the current study, we concentrate on entity-based coherence and on the question how it is expressed in the sentence beginnings since both languages – the learners' L1 Chinese as well as their L2 German – do not restrict the initial position in the sentence to a particular grammatical function (i.e. the subject). The position presents itself as an ideal position for linking a sentence to its preceding discourse and establishing local coherence.

Chinese is a *topic-prominent* language. Hence, its general word order and notably its left periphery is strongly determined by information-structural conditions: the topic always comes first which can either be a time or a locative phrase or a familiar referent, for example a referent that is known from the preceding discourse (Li and Thompson (1989): 15, 85f., 94f.). German is a *verb-second* language and provides an initial sentence position (*Vorfeld*), which precedes the finite verb. In contrast to Chinese, German is not strictly topic-prominent even though information

[1] For a comprehensive list of learner corpora see www.uclouvain.be/en-cecl-lcWorld.html.

[2] ALeSKo: ling.uni-konstanz.de/pages/home/zinsmeister/alesko.html.

[3] Multi-layer error tagging is discussed in Lüdeling et al. (2005). For a recent overview of error-tagged corpora see Hana et al. (2010).

structure influences the realisation of the Vorfeld (e.g. Molnár (1991); but see e.g. Frey (2004); Speyer (2007) for critical discussions).

Our working hypothesis is that the Chinese learners transfer rules of using the left periphery of a sentence in their L1 Chinese to their L2 German to assure local coherence and hence will show an overuse or an underuse of certain functions in comparison with L1-German speakers.

The rest of the paper presents the ALeSKo corpus and its (entity-based) discourse annotation. We conclude the paper by briefly discussing results from a contrastive interlanguage analysis of entity-based coherence.

2 Related Work

The linguistic annotation of learner corpora is a relatively recent development. The *International Corpus of Learner English* (ICLE)[4] is the largest project to date. It is responsible for creating a large database of comparable L2-English texts from speakers with a variety of L1s (currently of about 25 different L1s).

The multi-layer annotation of the German error-annotated FALKO corpus[5] is used as a prototype for the current project's annotation efforts.[6]

Albert et al. (2009) report on error tagging of a learner corpus of French L2 learners of English and a decision model for the best error correction derived from the annotation. The workshop series *Automatic Analysis of Learner Language* (AALL 2008, 2009) brought together various projects of L2-corpus developers and developers of Natural Language Processing applications for foreign language teaching.

The transfer of information structure between two verb-second languages and the filling of the *Vorfeld* is contrastively investigated by Bohnacker and Rosén (2008). However, their analysed data is not published as a reusable annotated corpus.

There have been various research efforts concerning the discourse annotation of L1 corpora. The current project adapts the annotation guidelines for coreference annotation and bridging by MATE (Poesio, 2000), information structure as applied to the Potsdam Commentary Corpus (Götze et al., 2007) and the implicit guidelines of centering annotation from Speyer (2005; 2007).[7]

3 Data

3.1 Collection

The corpus consists of 43 argumentative essays of Chinese L2 learners of German in which they discuss pros and cons of a given subject and state their own opinion. The learners were students at the Konstanz University of Applied Sciences, studying in the program *Business German and Tourism Management*[8] with a German level of about B2. In addition to the L2 texts, the ALeSKo corpus contains essays by L1 German high school students (aged 16–19) from Berlin, which originally belong to the FALKO corpus. In sum, the Alesko subcorpora include the following texts:

- **wdt07**: 25 L2 texts on the topic *Are holidays an unsuccessful escape from every-day life?* (6,902 tokens, 30–45 min, written exam, no aids)
- **wdt08**: 18 L2 texts on the topic *Does tourism support understanding among nations?* (6,685 tokens, 90 min., dictionary permitted)
- **Falko Essays L1 0.5**: 39 essays on different topics (34,155 tokens, typed in notepad, no internet access, no spell-checker).

The metadata for each individual text provides information about the person's ID, the L1, the year of birth, the gender, the study programme, the foreign language(s), the length of L2 exposure – if applicable – and the essay topic.

3.2 Preprocessing

The hand-written L2 learner texts were manually transcribed. All texts (both L2 and L1) were tokenized, lemmatized and part-of-speech tagged with the TreeTagger (Schmid, 1994). We used EXMARaLDA (Schmidt, 2004) for annotating topological fields in the tagged data. The annotation output of this annotation was converted into

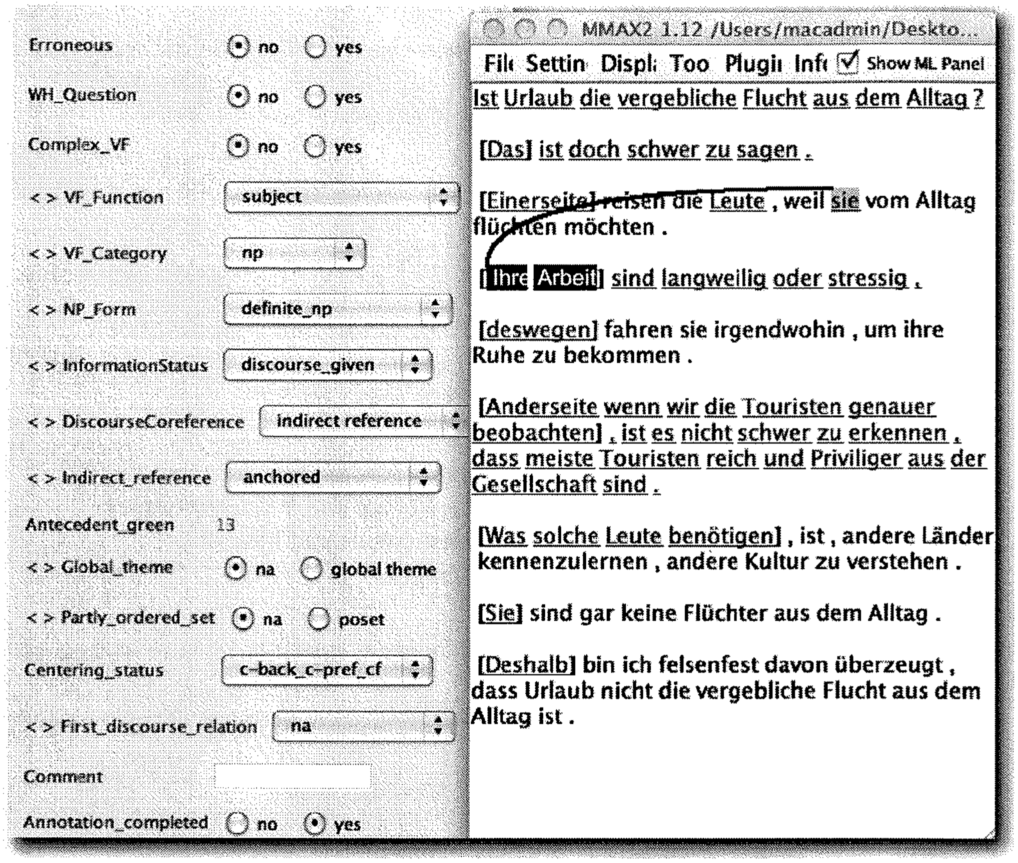
[4] ICLE: `cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm`

[5] FALKO: `http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko`

[6] The German L1 texts that we report on belong to the FALKO corpus (Falko Essays L1 0.5) and are enriched with additional layers of discourse annotation in the current project.

[7] In addition, we annotate discourse relations adapting the guidelines of the Penn Discourse Treebank (Prasad et al., 2007) which is not discussed in this paper.

[8] German: *Wirtschaftssprache Deutsch und Tourismusmanagement* (WDT).



MMAX2 1.12 /Users/macadmin/Deskto...
Fil Settin Displ Too Plugi Inf
☑ Show ML Panel

Erroneous ● no ○ yes
WH_Question ● no ○ yes
Complex_VF ● no ○ yes
< > VF_Function subject
< > VF_Category np
< > NP_Form definite_np
< > InformationStatus discourse_given
< > DiscourseCoreference indirect reference
< > Indirect_reference anchored
Antecedent_green 13
< > Global_theme ● na ○ global theme
< > Partly_ordered_set ● na ○ poset
Centering_status c-back_c-pref_cf
< > First_discourse_relation na
Comment
Annotation_completed ○ no ● yes

Ist Urlaub die vergebliche Flucht aus dem Alltag ?

[Das] ist doch schwer zu sagen .

[Einerseite] reisen die Leute , weil sie vom Alltag flüchten möchten .

[Ihre Arbeit] sind langweilig oder stressig .

[deswegen] fahren sie irgendwohin , um ihre Ruhe zu bekommen .

[Anderseite wenn wir die Touristen genauer beobachten] , ist es nicht schwer zu erkennen , dass meiste Touristen reich und Priviliger aus der Gesellschaft sind .

[Was solche Leute benötigen] , ist , andere Länder kennenzulernen , andere Kultur zu verstehen .

[Sie] sind gar keine Flüchter aus dem Alltag .

[Deshalb] bin ich felsenfest davon überzeugt , dass Urlaub nicht die vergebliche Flucht aus dem Alltag ist .

Figure 1: The MMAX2 annotation window of ALeSKo

the MMAX file format and served as the input for the current discourse annotation.

3.3 Annotation

For the investigation of the L2 learners' use of the sentences' left periphery as compared to its use by native speakers, both the L2 texts and the L1 texts are annotated with functional labels. The annotation is performed manually, using the tool MMAX2 (Müller and Strube, 2006), see Figure 1.

Our annotation guidelines define the labels, illustrate them with examples and discuss problematic cases (Breckle and Zinsmeister, 2009). The annotation proceeds in two steps: after a primary annotation by three student annotators or one of the authors, the two authors agree on a gold annotation for each *Vorfeld.*

Table 1 illustrates the relevant layers of annotation with examples.[9] Figure 1 shows a snapshot of the annotation of example (2) in the annotation tool MMAX2.

Information status (*new, deictic, discourse given*, cf. Götze et al. (2007)): The referent of the definite NP *Die Leute, die viele Reise machen* in example (1) is mentioned for the first time and is annotated *new* even though the term *Leute* as such occurs earlier in the text. The referent of *Ihre Arbeit* in (2) is also mentioned for the first time. However, it is a case of indirect reference in which the newly introduced referent is anchored by the possessive pronoun *Ihre* which refers to a discourse-old referent. *Ihre Arbeit* is therefore annotated *discourse_given.*

Partly-ordered set relation (*poset*, cf. Speyer (2005); Prince (1999)): In example (3) *jeden Morgen* (every morning') and *jeden Abend* ('every evening') form a set of *Tageszeiten* ('times of the day').[10]

Centering (*forward-looking center, preferred center, backward-looking center*, cf. Grosz et

[9] The *Vorfeld* constituent is underlined in the English translation.

[10] The poset relation is similar to the concept of contrastive topic (cf. Büring (1999)) which should be taken into account in future revisions of the corpus. Thanks to one of the reviewers for pointed this out to us.

Information status

(1) *[Die Leute, die viele Reise machen,]$_{new}$ haben immer mehr Geld als die, die selten reisen.*
'The people who travel a lot always have more money than those who seldom travel.'

(2) *Einerseite reisen die Leute$_1$, weil sie$_1$ vom Alltag flüchten möchten.*
[Ihre$_1$ Arbeit]$_{given}$ sind langweilig oder (...).
'On the one hand people travel because they want to escape every day life.
Their job is boring or (...)'

Partly-ordered set relation

(3) *[Jeden Morgen]$_{element\ 1}$ stehen wir auf, um pünktlich zur Arbeit zu sein. (...)*
[Jeden Abend]$_{element\ 2}$ bleiben wir zu Hause, sehen sinnlose Serien im Fernsehn.
'Every morning, we get up for being at work in time. (...)
Every evening, we stay at home, watch the senseless shows on TV.'

Centering

(4) *Durch Reisen können sie$_1$ auch andere Kultur und Lebenstile kennenlernen.*
[Sie]$_{1\ backward-looking\ center}$ können auch ihre Kenntnisse durch Reisen erweitern.
'By travelling, they can become acquainted to other culture and lifestyles.
They can also broaden their knowledge by travelling.'

Internal functions (frame-setting)

(5) *[Heutzutage]$_{frame-setting(temporal)}$ gelangt es in hoher Konjunktur, einen Urlaub zu machen.*
'Nowadays many people go on holidays.'

(6) *[In den Attraktionspunkten]$_{frame-setting(local)}$ werden (...) notwendige Einrichtungen konzentriert angeboten.*
'Necessary facilities are especially offered at the attraction sites.'

Table 1: Examples of discourse annotation in ALeSKo

al. (1995)): In example (4) *Sie* in the second sentence is a *backward-looking center* – the referential expression that corefers with the most salient expression in the previous sentence according to a saliency hierarchy (in comparison to other antecedents): subject is more salient than object(s), object is more salient than other functions.

In addition, **sentence-internal functions** are marked (*frame*: frame-setting topic (Götze et al. (2007): 167f.) and others): Example (5) and (6) present two *frame-setting* elements (temporal and local). They do not contribute to local coherence but they set the frame for the interpretation of the current sentence and are frequently used in the *Vorfeld* in L1 German (cf. Speyer (2007)).

4 Results and Conclusion

We performed a contrastive interlanguage analysis on the basis of the discourse annotation described in section 3. To this end, we compared the relative frequencies of the different functions in the *Vorfelds* of all 43 L2 essays with those in 24 of the Falko L1 essays. With respect to information status (including *poset*) and frame-setting elements, there is no statistical significant difference between L1 and L2 speakers. However, L2 speakers use the function *backward-looking center* significantly more often in the *Vorfeld* than L1 speakers.[11]

A more detailed discussion of the analysis is given in Breckle and Zinsmeister (in preparation). Under the (simplifying) assumption that the backward-looking center corresponds to the sentence topic we analyse the observed preference as a transfer effect from the topic-prominent L1 Chinese to the L2 German.

Acknowledgments

Heike Zinsmeister's research was financed by Europäischer Sozialfonds in Baden-Württemberg

[11] 287 out of 884 (32 %) Vorfelds constituents in the L2 essays function as backward-looking center vs. 207 out of 764 (27 %) in the L1 essays; χ^2=5.61, df=1, $p<0.05$. The conclusion is still valid when the scores are normalised by the lengths of the texts.

References

Camille Albert, Laurie Buscail, Marie Garnier, Arnaud Rykner and Patric Saint-Dizier. 2009. Annotating language errors in texts: investigating argumentation and decision schemas. In *Proceedings of the Third Linguistic Annotation Workshop* (LAW III), ACL 2009, 130–133. Singapore, Singapore.

Ute Bohnacker and Christina Rosén. 2008. The clause-initial position in L2 German declaratives: Transfer of information structure. *Studies of Second Language Acquisition*, 30: 511–538.

Margit Breckle and Heike Zinsmeister. 2009. Annotationsrichtlinien Funktion des Vorfelds. Manuscript. December 2009. Pedagogical University Vilnius and University of Konstanz.

Margit Breckle and Heike Zinsmeister. In preparation. A corpus-based contrastive analysis of local coherence in L1 and L2 German. In *Proceedings of the HDLP conference* Frankfurt/Main [a.o.]: Peter Lang.

Daniel Büring. 1999. Topic. In Peter Bosch and Rob van der Sand (eds.) *Focus – Linguistic Cognitive and Computational Perspectives*, 142–165. Cambridge: Cambridge University Press.

Werner Frey. 2004. A medial topic position for German. *Linguistische Berichte* 198. 153–190.

Michael Götze, Cornelia Endriss, Stefan Hinterwimmer, Ines Fiedler, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, Ruben Stoel and Thomas Weskott. 2007. Information structure. In S. Dipper, M. Götze and S. Skopeteas (eds.) *Information Structure in Cross-Linguistic Corpora: Annotation guidelines for phonology, morphology, syntax, semantics, and information structure* (Working Papers of the SFB 632, Interdisciplinary Studies on Information Structure (ISIS) 7), 94–137.

Sylvaine Granger. 2008. Learner corpora. In Anke Lü deling and Merja Kytö (eds.) *Corpus Linguistics. An International Handbook*, 259–275. Berlin / New York: de Gruyter.

Barbara Grosz, Arvind Joshi and Scott Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21, 203–225.

Jirka Hana, Alexandr Rosen, Svatava Škodová and Barbora Štindlová. 2010. Error-Tagged Learner Corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop* (LAW IV), ACL 2010, 11–19. Uppsala, Sweden.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar.* Berkeley and Los Angeles, CA: University of California Press.

Anke Lüdeling, Maik Walter, Emil Kroymann and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics* 2005, Birmingham, Great Britain.

Valéria Molnár, 1991. *Das TOPIK im Deutschen und Ungarischen.* Stockholm: Almquist & Wiksell International.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn and Joybrato Mukherjee (eds.) *Corpus Technology and Language Pedagogy*, 197–214. Frankfurt/Main [a.o.]: Peter Lang.

Massimo Poesio. 2000. Coreference. In Andreas Mengel et al. (eds.) *MATE Dialogue Annotation Guidelines*, 134–187.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo and Bonnie Webber. 2007. The Penn Discourse Treebank 2.0 Annotation Manual. Technical report. University of Pennsylvania [a.o.].

Ellen Prince. 1999. How not to mark topics: 'Topicalization' in English and Yiddish. 8 *Texas Linguistics Forum.*

Tanya Reinhart. 1980. Conditions for text coherence. *Poetics Today* 1(4): 161–180.

Tanya Reinhart. 1982. *Pragmatics and linguistics: An analysis of sentence topics.* Reprint of an earlier publication in 1981, Indiana University Linguistics Club.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In D. H. Jones and H. Somers (eds.) *New Methods in Language Processing*, UCL Press, 154–164.

Thomas Schmidt. 2004. EXMARaLDA – ein Modellierungs- und Visualisierungsverfahren für die computergestützte Transkription gesprochener Sprache. In *Proceedings of Konvens.* Vienna, Austria.

Augustin Speyer. 2005. Competing Constraints on Vorfeldbesetzung in German. In *Proceedings of Constraints in Discourse Workshop*, 79–87. Dortmund, Germany.

Augustin Speyer. 2007. Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. *Zeitschrift für Sprachwissenschaft*, 26: 83–115.