

**The ALeSKo learner corpus: Design – annotation – quantitative  
analyses**

[short title for running heads: The ALeSKo learner corpus]

Heike Zinsmeister

University of Konstanz

Linguistics Department

Box 185

78457 Konstanz

Germany

heike.zinsmeister@uni-konstanz.de

Margit Breckle

Lithuanian University of Educational Sciences

Department of German Philology and Didactics

Faculty of Philology

Studentų g. 39

08106 Vilnius

Lithuania

margit.breckle@gmx.de

## **Abstract**

The ALesKo learner corpus is a small-scale comparable corpus consisting of two subcorpora: annotated essays by advanced Chinese learners of German and comparable essays by German native speakers. The motivation for its compilation was the investigation of discourse-related phenomena such as local coherence in second-language acquisition of German. After introducing how the texts were compiled and annotated, the article focuses on quantitative studies at the token level. We discuss problems of tokenisation and part-of-speech tagging and compare the inventory of the two subcorpora in terms of frequently used N-grams and lexical richness, among other aspects. We conclude the article by describing possible applications of the study in foreign language acquisition research and language teaching.

## **1. Introduction**

The ALesKo corpus<sup>1</sup> (Breckle and Zinsmeister 2010; Zinsmeister and Breckle 2010), under construction since 2009, provides a linguistically

---

<sup>1</sup> ALesKo is an abbreviation for *Annotiertes Lernersprachenkorpus* ('annotated learner language corpus'). Its URL is <[ling.uni-konstanz.de/pages/home/zinsmeister/alesko.html](http://ling.uni-konstanz.de/pages/home/zinsmeister/alesko.html)> (accessed 30.11.2011).

annotated collection of texts. It is multilingual in the sense that it comprises texts written by learners of German as a foreign language and texts by German native speakers. These texts are not translations of each other but are comparable by belonging to the same text type, namely argumentative essays written on a controversial topic.

The aim of the ALeSKo corpus is to conduct studies on coherence phenomena.<sup>2</sup> Local coherence can be described as the fluency of a text on a sentence-by-sentence basis by which a sentence is linked to its textual context. All such phenomena have in common that they cannot be analysed on the level of an individual sentence alone but require that the textual context be taken into account.

All the studied texts were preprocessed in the same way and annotated according to the same guidelines, which allows users to compare the two subcorpora in a systematic way. The learner texts were collected from two different student groups, but the authors had very similar backgrounds: Chinese learners of German in their third or fourth year of learning German. It is important to note that the learner texts in ALeSKo do not document the

---

<sup>2</sup> Coherence phenomena are often studied using narrative texts. We opted to investigate argumentative writing instead, as it is a more commonly used text type in foreign language teaching at the university level and it generally includes a more varied set of discourse relations. The results of the studies are intended to provide insights into L2 learners' interlanguage. These insights should be applicable in foreign language teaching (see also Section 5).

process of individual second language (L2) acquisition; rather, they present one snap-shot per learner at a certain point in time and therefore create a cross-individual text basis. Comparing this text basis with essays written by native speakers (L1) taken from the Falko corpus<sup>3</sup> provides evidence of the overuse and underuse of certain linguistic items in the learner texts. Given that the learners all had the same L1 (Chinese) and the same L2 (English) before learning German, differences between the L2 and the L1 subcorpora may occur due to transfer effects, which arise when the learners apply Chinese grammar rules or lexical preferences in German. (Transfer effects of L2 English to L2 German have not yet been taken into account.)

Differences may also come about due to general patterns of L2 developmental stages. However, texts from learners with different L1s than Chinese have not been included in the ALeSKo corpus, since such research lies outside of our current investigative scope.

The article is organised as follows: Section 2 gives an overview of the compilation of the texts and their preprocessing. Section 3 presents the research questions that motivated the annotation of ALeSKo and briefly introduces the different annotation layers. It also discusses problems of tokenisation and part-of-speech tagging, which are relevant for the quantitative studies presented in Section 4. There, we compare the lexical

---

<sup>3</sup> Falko: < <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/research/falko/standardseite>> (accessed 30.11.2011).

content of the two subcorpora in terms of token N-grams and the grammatical content in terms of an approximation by part-of-speech N-grams. Finally, we model the complexity of the texts on different levels. Section 5 concludes the article by describing two possible applications for ALeSKo in foreign language acquisition research and language teaching.

## **2. Design of the corpus**

Contrastive interlanguage analysis assumes that learners speak an ‘interlanguage’ (cf. Selinker 1972) that systematically differs from the target language. Such analyses compare learner texts either with texts of native speakers or with data from different learner groups. By this method, the overuse and underuse of certain words and structures by learners that “contribute to the foreign-soundingness of perhaps otherwise error-free advanced interlanguage” (Granger 2008: 267) can be studied. In order to explore to what extent the L2 phenomena correspond to the L1 speakers’ language use and to determine the types of divergences that can be observed, the ALeSKo corpus has been designed as a comparable corpus: Its L2 texts have been complemented by comparable L1 German texts originating from the Falko corpus (Lüdeling et al. 2008). Both L2 and L1 texts are argumentative essays meant to discuss the pros and cons of a given thesis, concluding with the writer’s own opinion on the topic. This means

that the subcorpora are comparable at the level of text type. Table 1 gives an overview of the texts collected in the ALeSKo corpus (version 0.9<sup>4</sup>):

**Table 1.** ALeSKo L1 and L2 subcorpora and their thematic subcorpora.

Subcorpus	Authors	Size	Topic	Conditions
L2 <sub>all</sub>	Chinese L2 learners of German: students at HTWG Konstanz in the BA program	43 texts, 11,716 tokens (13,584 tokens with punctuation), Ø 272 tokens/text, Ø 25 sentences/text	Different topics	Different conditions
L2 <sub>holidays</sub> (wdt07)	<i>Business German and Tourism Management</i> in 2007 and 2008, German level ~B2, aged 18-23.	25 texts, 5,914 tokens (6,896 tokens with punctuation), Ø 237 tokens/text, Ø 23 sentences/text	<i>Are holidays an unsuccessful escape from everyday life?</i>	30–45 min, (hand-) written exam, no aids.
L2 <sub>tourism</sub> (wdt08)		18 texts, 5,802 tokens (6,688 tokens with punctuation), Ø 322 tokens/text, Ø 27 sentences/text	<i>Does tourism support understanding among nations?</i>	90 min, (hand-) written in-class task, dictionary permitted.
Falko Essays L1 0.5 (dhw)	German L1 speakers: Berlin high-school students, aged 16–19.	39 texts	Different topics	90 min, in-class task, typed in Notepad, no internet access, no spell-checker.
L1 <sub>all</sub> <sup>5</sup>		24 texts, 17,185 tokens (19,587 tokens with punctuation), Ø 716 tokens/text, Ø 48 sentences/text		
L1 <sub>feminism</sub>		4 texts, 2,929 tokens (3,319 tokens with	<i>Feminism has done more harm to the cause of</i>	

<sup>4</sup> The present version of the ALeSKo corpus is 0.9; there are some tokenisation issues not yet addressed that affect various levels of the annotation and lead to minor inconsistencies. Version 1.0 is intended to remedy this problem. The overall number of tokens in the next version will be slightly different from what is reported here.

<sup>5</sup> The subcorpus L1<sub>all</sub> and its thematic subcorpora are used in the current study.

		punctuation), Ø 732 tokens/text, Ø 44 sentences/text	<i>women than good.</i>
L1 <sub>salary</sub>		9 texts, 7,038 tokens (8,006 tokens with punctuation) Ø 782 tokens/text, Ø 54 sentences/text	<i>A man/woman's financial reward should be commensurate with their contribution to the society they live in.</i>
L1 <sub>crime</sub>		11 texts, 7,218 tokens (8,252 tokens with punctuation), Ø 656 tokens/text, Ø 45 sentences/text	<i>Crime does not pay.</i>

The core of the ALeSKo corpus consists of 43 argumentative essays written by Chinese L2 learners of German. The learners were students in their fourth semester at the Konstanz University of Applied Sciences in the program *Business German and Tourism Management*.<sup>6</sup> They had been learning German for 2.5 to 3.5 years and their level of German was about B2, according to the Common European Framework of Reference for Languages (CEFR) (cf. CEFR [online]). The essays were collected during the Winter Term 2007 (wdt07, ‘L2<sub>holidays</sub>’) and the Winter Term 2008 (wdt08, ‘L2<sub>tourism</sub>’), respectively.

In addition to the L2 texts, the ALeSKo corpus includes 39 essays by L1 German high school students (aged 16–19) from Berlin, which originate from the Falko corpus (Falko Essays L1 0.5). These essays were collected in

---

<sup>6</sup> In German: *Wirtschaftssprache Deutsch und Tourismusmanagement*; thus, the abbreviation for the subcorpora is ‘wdt’.

Germany (*Deutschland*) at a high school in *Hermannswerder* (dhw) in June 2007 (cf. Reznicek et al. 2010). The reason for including comparable essays by German native speakers in ALeSKo was to ensure that comparisons would be possible with L1 essays that have been preprocessed and annotated in exactly the same way as the L2 texts.<sup>7</sup> Both the L2 group and the L1 group were fairly homogeneous: The L2 learners all had about the same level of language proficiency, and the L1 speakers were in the same grade at the same high school.

For reasons of comparability, the ALeSKo corpus has adhered to design criteria provided by the Falko project (Lüdeling et al. 2008) and the ICLE project (cf. ICLE [online]), e.g., text type (an essential aspect of the meta-data collected by Falko) and transcription.

The L2<sub>holidays</sub> essays were hand-written as part of an exam in which no aids were permitted. During the 90-minute exam, the students spent approximately 30–45 minutes on the essay. The introductory text and the task were formulated as follows:

In the text *Escape to the holidays* (Jost Krippendorf) and in the interview with Mr Hennig, you learned about two contrary opinions regarding the *escape hypothesis*.

Write a text (introduction – main body – conclusion) on the topic *Are holidays an unsuccessful escape from everyday life?*, in which you present arguments for and against

---

<sup>7</sup> ALeSKo v0.9 includes only a subset of 24 of the L1 texts (those that have been fully annotated).

the escape hypothesis and in which you make your own opinion clear in the conclusion.

(Hint: The text will be assessed with regard to *language* and *content*.)<sup>8</sup>

The preparation included a teaching unit (12 lessons) on the topic *Tourism and Travel* based on Lodewick (1999). During the teaching unit, the students had the option to hand in a practice essay for correction and comments.

The wdt08 essays were hand-written as a 90-minute in-class task during a teaching unit (6 lessons) on the topic *Tourism and Travel* based on Lodewick (1999). The use of a dictionary was permitted. The introductory text and the task were formulated as follows:

In travel companies' advertisements, tourism is recommended as a means for developing understanding among nations. Tourism offers the opportunity to come into contact with people from foreign cultures and to learn about their way of life. Therefore, tourism makes an important contribution toward reducing prejudices between nations. What do you think? Does tourism promote global understanding?

---

<sup>8</sup> In German:

In dem Text *Flucht in den Urlaub* (Jost Krippendorf) und in dem Interview mit Herrn Hennig haben Sie zwei gegensätzliche Standpunkte zur *Fluchtthese* kennen gelernt. Schreiben Sie einen Text (Einleitung – Hauptteil – Schluss) zum Thema *Ist Urlaub die vergebliche Flucht aus dem Alltag?*, in dem Sie Argumente für und gegen die Flucht-These darstellen und am Schluss Ihren eigenen Standpunkt deutlich machen.

(Hinweis: Der Text wird *sprachlich* und *inhaltlich* bewertet.)

Task:

Write a cohesive text (introduction – main body – conclusion) on the topic *Does tourism support understanding among nations?*, in which you present *arguments for and against* the hypothesis and in which you make your *own opinion* clear in the conclusion.<sup>9</sup>

The L1 essays (Falko Essays L1 0.5) were written as a 90-minute in-class task and were typed in Notepad. No aids were permitted. The students could choose to write their argumentative essay on one of the following statements:

Feminism has done more harm to the cause of women than good.<sup>10</sup>

Crime does not pay.<sup>11</sup>

---

<sup>9</sup> In German:

Der Tourismus wird in der Werbung von Reiseunternehmen oft als Mittel zur Völkerverständigung empfohlen. Er biete Gelegenheit, mit Menschen fremder Kulturen in Kontakt zu kommen und ihre Lebensbedingungen kennen zu lernen. Damit leiste der Tourismus einen wichtigen Beitrag zum Abbau von Vorurteilen zwischen den Völkern. Was meinen Sie dazu? Dient der Tourismus der Völkerverständigung?

Aufgabe:

Schreiben Sie einen zusammenhängenden Text (Einleitung – Hauptteil – Schluss) zum Thema *Dient der Tourismus der Völkerverständigung?*, in dem Sie *Argumente für und gegen* die These darstellen und am Schluss Ihren *eigenen Standpunkt* deutlich machen.

<sup>10</sup> In German: Der Feminismus hat den Frauen mehr geschadet als genutzt.

<sup>11</sup> In German: Kriminalität zahlt sich nicht aus.

A man/woman's financial reward should be commensurate with their contribution to the society they live in.<sup>12</sup>

These tasks come from suggested essay titles in the International Corpus of Learner English (cf. ICLE – titles [online]).

Two Chinese student assistants manually transcribed the hand-written L2 texts (see Figure 1) according to transcription guidelines.

**Figure 1.** Extract of hand-written text of wdt08\_07.

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

Only the students' final versions were transcribed; no corrections made by the students during the writing process were marked in the transcription. In the first version, the original line breaks were maintained (see Figure 2).

**Figure 2.** Extract of the first transcribed version of wdt08\_07.

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,

---

<sup>12</sup> In German: Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, der er/sie für die Gesellschaft geleistet hat.

fremde Kultur kennenzulernen, schöne Landschaft zu genießen usw. Aber einige [...]

The transcriptions were checked independently and corrected when necessary. A basic annotation was performed on the final transcript without line breaks: All texts (L2 and L1) were tokenised, lemmatised, and part-of-speech tagged (see Section 3).

In addition, all texts were labelled with metadata: an ID for the author, native language, year of birth, gender, study program, foreign language(s), length of L2 exposure, and essay topic.<sup>13</sup>

### **3. Annotation layers**

This section is concerned with the annotation layers of the ALeSKo corpus. It presents the research questions that motivated the individual layers, the tagsets, and the tools for automatic and manual annotation. In addition to general linguistic annotation layers, the corpus includes annotations specific to learner corpora, namely error tagging and target hypotheses. Thus far, explicit target hypotheses have only been included in the annotation of article selection for referential expressions. This investigation is described in detail in Breckle & Zinsmeister (2010).

---

<sup>13</sup> In the investigations, only a subset of the available metadata has been used. Features such as gender and additional foreign languages have not yet been taken into account.

### *3.1. Basic annotation layers: Parts of speech and lemmas*

All texts were automatically part-of-speech tagged and lemmatised using TreeTagger<sup>14</sup> (Schmid 1994, 1995) on the basis of the STTS tagset (Schiller et al. 1999) and were subsequently loaded into the EXMARaLDA Partitur Editor<sup>15</sup> (Schmidt 2004) for manual correction of the automatic tagging and for further annotation (see Figure 3).

Results from the tagger were similar to scores reported in the literature for tagging average newspaper texts.<sup>16</sup> When all texts were considered together as one corpus, we obtained a per-word accuracy<sup>17</sup> of 97.3% for the L2 texts and a slightly lower accuracy of 96.3% for the L1 texts.<sup>18</sup> Comparing the per-word accuracy of individual texts from the two subcorpora having the same number of equally sized text fragments (two times 11 texts with 361

---

<sup>14</sup> TreeTagger: <<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>> (accessed 30.11.2011).

<sup>15</sup> EXMARaLDA: <[http://www.exmaralda.org/en\\_index.html](http://www.exmaralda.org/en_index.html)> (accessed 30.11.2011).

<sup>16</sup> Other studies report systematic part-of-speech tagging errors in learner texts (cf. Díaz-Negrillo et al. 2010 for a recent discussion). A qualitative analysis of the tagging errors in ALeSKo is left for further research.

<sup>17</sup> The per-word accuracy denotes the probability for each individual token to be tagged correctly:  $\text{accuracy} = \text{number of correctly tagged tokens} / \text{number of all tokens}$ .

<sup>18</sup> Schmid (1995) reported a per-word accuracy of 97.53% for TreeTagger when tested on newspaper texts.

tokens each including punctuation) produced a non-significant difference (L2 accuracy: 97.0%, L1 accuracy: 96.3%, Wilcoxon test:  $W = 42$ ,  $p = 0.235$ ).

The tagger options were set to print only the most probable tag and to print *<unknown>* if the lemma was not known (rather than copying the token). This latter function made it easier to spot lemma errors arising from word formation errors (e.g., *menschlichzwischen Beziehung*, correct: *zwischenmenschlichen Beziehung* ‘interpersonal relations’ in wdt07\_20) or simple typos; it also singled out other non-standard forms, such as neologisms (e.g., *vollrichten* in the sense of *verrichten* ‘carry out, perform’ in dhw031) and colloquial language (e.g., *Weichei* ‘wimp’ in dhw005, *Egogang* ‘ego trip’ in dhw006). Student annotators corrected the tokenisation, the part-of-speech tagging, and the lemmatisation. To this end, they split or merged tokens when necessary and entered the correct part-of-speech tag or lemma into separate tiers of the Partitur Editor (see Figure 3). Each text was independently corrected by two annotators and was subsequently controlled by a linguistic expert. Finally, the corrected parts of speech and lemmas were merged with the rest of the tags to create two new layers that assigned each token its correct part of speech tag and lemma, respectively. The original correction layer became a layer that merely pointed to the corrected tokens with an *x* (see Figure 3).

**Figure 3.** The basic annotation layers of automatic pos-tagging, lemmatisation, and their manual correction, extract from wdt08\_07.

	324	325	326	327	328	329	330	331	332	333	334	335
<b>Text</b>	Die	Reisende	müssen	die	andere	Kultur	und	die	Unterschiede	respekieren	,	und
<b>Wortart</b>	ART	NN	VMFIN	ART	ADJA	NN	KON	ART	NN	VVFIN	,\$	KO
<b>POS-Fehler</b>										x		
<b>Korrigierte Wortart</b>	ART	NN	VMFIN	ART	ADJA	NN	KON	ART	NN	VVINF	,\$	KO
<b>Lemma</b>	d	Reisende	müssen	d	ander	Kultur	und	d	Unterschied	[unknown]	,	und
<b>Lemma-Fehler</b>										x		
<b>Korrigiertes Lemma</b>	d	Reisende	müssen	d	ander	Kultur	und	d	Unterschied	respektieren	,	und

At first glance, it may seem surprising that part-of-speech tagging of L1 texts had a lower accuracy (96.3%) than that of L2 texts (97.0% / 97.3%). However, the L1 texts were written by high school students without further editing, which means that they do not necessarily contain standard language and may differ from the newspaper texts on which TreeTagger was trained. Furthermore, it holds true that the longer the sentence, the higher the probability that TreeTagger will mis-tag a token. Given that the average sentence length of the L1 texts (in terms of words) is longer than that of the L2 texts, it is therefore likely that the L1 texts will contain more tagging errors. Finally, the L1 authors misspell nouns by using lower-case letters and omit punctuation between adjacent sentences more often than the L2 authors, both of which tend to produce tagging errors.

### 3.2. Starting a sentence in learner German

One of the research questions that motivated the annotation of the ALeSKo corpus was whether L2 authors start their sentences in the same way that L1 authors do. The answer to this question should shed light on a variety of aspects: (i) The syntactic competence of the learners – do they use the same syntactic categories and the same grammatical functions as L1 authors when starting a sentence? (ii) Their competence in structuring information – do they present old and new information with the same linguistic means as L1 authors do? (iii) The learners’ knowledge of local coherence – do L2 authors start a sentence in a contextually adequate way in comparison to L1 authors?

The ‘starting region’ was operationalised as the *Vorfeld* (‘prefield’) in accordance with the model of topological fields (e.g., Höhle 1986), which organises clauses in German into different fields. Figure 4 depicts the fields of a declarative main clause.

**Figure 4.** Topological field model of a German main clause (adapted from wdt07\_03).

<b>Vorvorfeld</b>	<b>Vorfeld</b>	<b>Linke Satzklammer</b>	<b>Mittelfeld</b>	<b>Rechte Satzklammer</b>	<b>Nachfeld</b>
‘pre-prefield’	‘prefield’	‘left sentence bracket’	‘middle field’	‘right sentence bracket’	‘post field’
<i>Aber</i>	<i>aufgrund von Erschöpfung</i>	<i>Müssen</i>	<i>sie</i>	<i>abschalten</i>	<i>obwohl</i> ...

'But	due to exhaustion	have_to	they	unwind	although'
------	----------------------	---------	------	--------	-----------

A German declarative main clause is characterised by the sequence *prefield* – *left sentence bracket*, i.e., a single constituent (not necessarily the subject) preceding the finite verb.<sup>19</sup> The pre-prefield<sup>20</sup> is an optional position for sentence initial coordinators and left-dislocated constituents. We adopted the topological field annotation of the Falko corpus (cf. Doolittle 2008), which makes use of a limited set of field tags: VF (*Vorfeld* ‘prefield’), MF (*Mittelfeld* ‘middle field’), NF (*Nachfeld* ‘post-field’), and LSK and RSK (*linke* and *rechte Satzklammer* ‘left and right sentence bracket’). We marked the spans of matrix clauses (*Matrixsatzbeginn*) in one annotation layer in EXMARaLDA and the individual topological fields of the clause in another. Additional layers for spans and fields of embedded clauses (*Konstituenten-Sätze* KS) were added in a recursive way when required (see Figure 5 for an example sentence). There is only one explicit difference in comparison to the Falko annotation: Coordinating elements in the pre-prefield are considered part of the prefield span in Falko (Doolittle 2008: 37), whereas

<sup>19</sup> For the occurrence of multiple constituents in the prefield, see Müller (2003).

<sup>20</sup> Our ‘pre-prefield’ is a common field for various phenomena. Höhle (1986: 329), for example, distinguishes KOORD/PARORD and K<sub>L</sub> for left-dislocated constituents.

these elements are excluded from the prefield in the ALeSKo annotation.<sup>21</sup>

This design decision was made to allow easy access to the prefield tokens in the quantitative evaluation without further filtering. Erroneous field structures were marked with an error tag at the clause level:  $f_{<clause>}$ .

For instance,  $f_{KS}$  in Figure 5 denotes that the embedded clause could not be properly divided into fields (it is a relative clause with an erroneous verb position). Error analysis at the field level has been left for further research.

**Figure 5.** Recursive layers of topological field annotation in EXMARaLDA (wdt08\_18).

	301	302	303	304	305	306	307	308	309	310	311	312	313
Text	Sie	beobachten	den	Lebensstil	der	Einheimische	,	welchen	finden	sie	sehr	komisch	.
[KorrPOS_POS]	PPER	VVFIN	ART	NN	ART	NN	\$,	PRELS	VVFIN	PPER	ADV	ADJD	\$.
[KorrLem_Lem]	sie	beobachten	d	Lebensstil	d	Einheimische	,	welch	finden	sie	sehr	komisch	.
Matrixsatzbeginn	x												
Matrixsatzfelder	VF_MS	LSK_MS	MF_MS										
Konstituentensatzbeginn 1								x					
Konstituentensatzfelder 1								f_KS					

Further annotation of the prefield constituents required a relational annotation in order to model, for example, coreference or discourse relations. For this purpose, MMAX2<sup>22</sup> (Müller and Strube 2004) is a suitable

<sup>21</sup> According to the Falko scheme, the example in Figure 4 has a prefield that contains both the coordinating conjunction (*Aber*) and the prepositional phrase (*aufgrund ...*), whereas the ALeSKo scheme differentiates between prefield proper and material in the pre-prefield:

Falko: [Aber aufgrund von Erschöpfung]<sub>VF</sub> müssen sie abschalten, obwohl ...

ALeSKo: [Aber]<sub>VVF</sub> [aufgrund von Erschöpfung]<sub>VF</sub> müssen sie abschalten, obwohl ...

<sup>22</sup> MMAX2: <<http://mmax2.sourceforge.net/>> (accessed 30.11.2011).

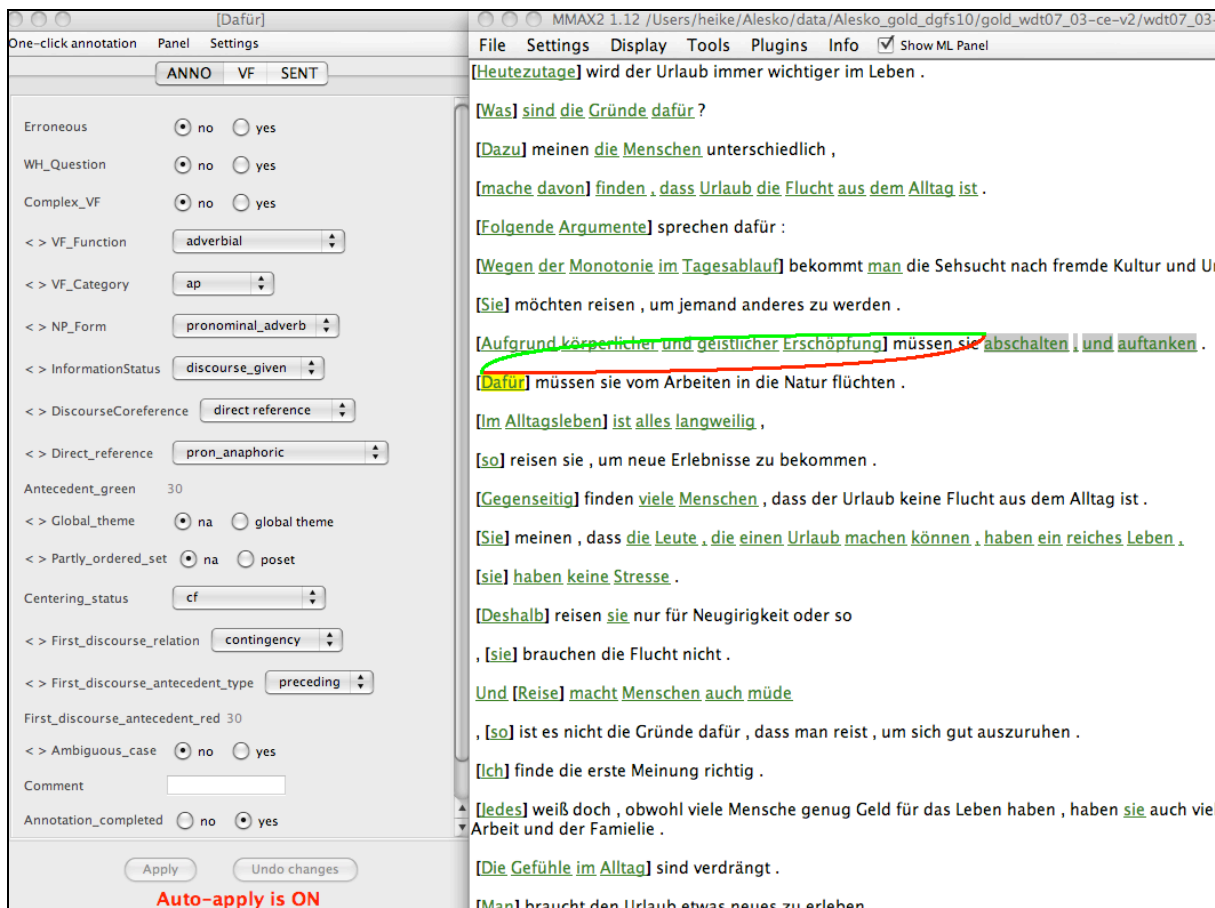
annotation tool.<sup>23</sup> We therefore converted the annotated EXMARaLDA XML files into MMAX2 XML files. The pre-annotation of prefields and sentences facilitated a user-friendly display of the texts. Figure 6 depicts the annotation window, in which each clause is printed on its own line; prefields are marked in green and are framed by bold brackets. The following list briefly summarises the annotation categories applied to the prefield category and their functions both within the clause and with respect to the context.

- *General description*: Erroneous, *wh*-question, complex prefield (default is the simple prefield)
- *Constituent level*: Function, category, surface form (e.g., definite NP, proper name)
- *Referential level*: Information status (cf. Prince 1999), coreference, and anaphora (direct and indirect, cf. Poesio 2000), reference to the global theme, partly ordered contrast sets (cf. Prince 1999, Speyer 2007)
- *Text level*: Centering status (cf. Grosz et al. 1995), discourse relations (cf. Prasad et al. 2007).

---

<sup>23</sup> MMAX2 uses a relation-based data model. EXMARaLDA, on the other hand, uses a time-based data model, which represents the annotation in tiers. The annotation refers to segments defined with respect to a common timeline. The MMAX2 model provides pointers for encoding relations in a structured way and visualising the annotation of relations. For further discussion, see Dipper et al. (2004). Other suitable tools are PALinkA (Orăsan 2003) and Serengeti (Diewald et al. 2008).

**Figure 6.** Annotation of sentence initial properties in MMAX2.



The annotation is described in detail in the ALeSKo annotation guidelines (Breckle & Zinsmeister 2009). For the time being, we have not applied a detailed error tagging at this level of annotation. The annotators only marked an error when a declarative main clause did not contain the sequence *prefield – finite verb* in a target language-like way. For this study,

erroneous field structures were not taken into account.<sup>24</sup> If a coordinate conjunction or a left-dislocated element preceded the prefield, the annotators marked the sentence with the label ‘complex prefield’. We evaluated these instances separately from ‘bare prefields’.

Studies that investigate the use of prefields in ALeSKo have been published in Breckle & Zinsmeister (forthcoming) and Zinsmeister & Breckle (2010). Ongoing work addresses the annotation of non-prefield phrases.

### *3.3. Accessing the ALeSKo annotation*

In order to ensure sustainable availability of the annotated data independent of the corpus creators, the final ALeSKo version will be integrated into the Falko corpus. ALeSKo version 1.0 will be available in PAULA XML (cf. Dipper 2005) and will also be converted to RelANNIS to make the data accessible for multi-layer search in ANNIS2 (cf. Zeldes et al. 2009).

## **4. Quantitative descriptive analyses**

---

<sup>24</sup> There are only ten ‘erroneous’ sentences in the L2 texts: seven cases in which more than one constituent erroneously precedes the finite verb (as in the example (i) below) and three cases of main-clause structure in embedded sentences that required the finite verb in final position.

(i) \*Vielleicht vorhin glaubst du, dass ... . (literal: Maybe before believe you that ..., wdt08\_13).

In this section, we present quantitative descriptive analyses of the corpus and its parts, comparing the L2 subcorpus with the L1 subcorpus and the thematic subcorpora within the two subsets. The analyses characterise the subcorpora with respect to three dimensions: Lexical Content (4.1), Grammatical Context (4.2), and Complexity (4.3). The statistics allow us to compare the subcorpora and will serve as a basis for drawing conclusions as to the learners' competence in German.

#### *4.1 Lexical Content*

Language use is a combination of creative elements and formulaic sequences; this holds for both L1 and L2. Concerning the lexical content, we analyse to what extent the learners reproduce (memorised) formulaic sequences (so-called *chunks*) in a target language-like way.

The term *chunk* was introduced by Miller (1956). Because of the limited capacity of the working memory, information is combined into information units. This process is referred to as *chunking*; the product of chunking is called a *chunk*. According to Newell et al. (1989: 125), chunking plays a significant role not only in perception and memorisation but also in the acquisition of new knowledge. Newell (1990: 328) summarises the role of chunking as follows: "Much experimental evidence exists that chunking goes on all the time."

Chunks can be defined as formulaic more-morphemic sequences (e.g., idioms or collocations but also patterns / sentence frames, cf. Aguado 2002: 28) that are memorised as a whole. According to Andersson (2000: 207f.), chunks are also recalled *en bloc*: If one part of the chunk is recalled, the rest of the chunk is also activated.

Based on Myles et al. (1998) and Peters (1983), the following characteristics (amongst others) of spoken language determine whether a sequence counts as formulaic: (i) frequency, (ii) invariance, (iii) speed, (iv) phonological coherence, (v) comparably high degree of complexity, (vi) comparably high degree of correctness, and (vii) situation-specific appropriateness. However, it is obvious that chunks also appear in written language. In this case, characteristics (i), (ii), (v), and (vi) seem especially relevant.

In language acquisition, formulaic sequences and creative language use appear to co-exist (cf. Myles et al. 1998; Wray 1992); both contribute to language use in their own specific ways and interact with each other.

According to Raupach (1984), even advanced adult L2 learners make extensive use of formulaic sequences. Hickey (1993: 34) comes to the conclusion that an L2-specific use of chunks can be observed, demonstrating either an overuse or an underuse compared to L1 speakers' usage. When analysing chunks, the learner's perspective plays an important role, since sequences – although erroneous in the target language – might function as chunks from the learner's perspective (cf. Aguado 2002: 34).

For our study, we analyse N-grams (i.e., bi- and trigrams) in the L1 and L2 subcorpora as well as in the thematic subcorpora in order to observe formulaic sequences (chunks) used by the L1 and L2 authors. Bigrams are defined as sequences of two successive words; trigrams are strings of three successive words in a text. Our analysis includes token bigrams and trigrams as well as part-of-speech bigrams and trigrams. It should be noted that our analysis is entirely frequency-based.<sup>25</sup>

In order to decide whether an N-gram counts as a chunk, one must consider both N-word-phrases in which the entire phrase counts as a chunk (e.g., the bigram *viel Spaß* ‘much fun’) and longer phrases of which the N-gram is one part (e.g., the bigram *Meinung nach* as part of *meiner Meinung nach* ‘in my opinion’). The more words the phrase contains, the easier it is to decide whether it qualifies as a chunk (see Footnote 28). In this article, the focus is on the analysis of trigrams, but our findings also hold for bigrams.

The N-grams were analysed on the basis of the following categorisation: all – the whole N-gram is part of the title or the task description (i.e., probably not memorised); w – the N-gram contains one word of the title or the task description; t – the N-gram contains a (part of a) chunk that was taught in the teaching unit in class prior to the essay writing (only applicable for L2; i.e., memorised for at least some time), e.g., *Verarmung der*

---

<sup>25</sup> In addition to mere frequency, other measures can also be used for the analysis of lexical (and grammatical) content (cf., e.g., Stefanowitsch and Gries 2003, Baroni and Lenci 2010). These have not been implemented because of the small size of the ALeSKo corpus.

*zwischenmenschlichen Beziehungen* (‘impoverishment of interpersonal relations’); o – other chunks (idioms, collocations, patterns, and sentence frames that are memorised, rather stable, and possibly target language-like, e.g., *meiner Meinung nach* ‘in my opinion’); and na – not applicable (i.e., sequences that are not formulaic).<sup>26</sup>

In order to illustrate our findings, we ranked the 17 most frequent trigrams in the thematic subcorpora L1<sub>salary</sub><sup>27</sup> and L2<sub>holidays</sub><sup>28</sup> (see Table 2); the table provides absolute frequencies rather than proportions, since it is rank order that is of interest here.

**Table 2.** Ranking (1 to 17) of trigrams in the subcorpora L1<sub>salary</sub> and L2<sub>holidays</sub>.

L1 <sub>salary</sub>			L2 <sub>holidays</sub>		
	N	Chunk		N	Chunk
1. für die gesellschaft	22	all	1. aus dem alltag	64	all
2. den ganzen tag	8	o	2. flucht aus dem	57	all
3. die finanzielle entlohnung	5	all	3. vergebliche flucht aus	35	all
4. sie für die	5	all	4. dem alltag ist	31	all

<sup>26</sup> Two independent student annotators attained substantial agreement of  $\kappa = .8$  in the classification of the about 30 highest-ranked chunks (in the L2 texts the agreement was almost perfect with  $\kappa = .88$  (due to title strings); in the L1 texts they attained  $\kappa = .72$ ).

<sup>27</sup> Essay title in German: *Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, den er/sie für die Gesellschaft geleistet hat.*

<sup>28</sup> Essay title in German: *Ist Urlaub die vergebliche Flucht aus dem Alltag?*

5.	doppelt so viel	4	o	5.	die vergebliche flucht	30	all
6.	meiner meinung nach	4	o	6.	urlaub die vergebliche	22	all
7.	ist es nicht	4	na	7.	meinung nach ist	10	o
8.	beitrag für die	4	all	8.	urlaub keine flucht	9	w
9.	die gesellschaft und	3	w	9.	keine flucht aus	9	w
10.	nicht so viel			10.	meiner meinung		
		3	o		nach	9	o
11.	eines menschen dem	3	all	11.	dass urlaub die	8	w
12.	er für die	3	all	12.	in den urlaub	7	all
13.	es gibt viele	3	na	13.	nach ist urlaub	7	w
14.	entlohnung eines			14.	urlaub zu machen		
	menschen	3	all			6	o/w
15.	und kann nicht	3	na	15.	urlaub kann man	6	w/na
16.	was ist mit	3	o	16.	ist urlaub die	6	all
17.	für seine arbeit	3	na	17.	kann man sich	6	t

The trigram rankings of the thematic subcorpora  $L1_{\text{salary}}$  and  $L2_{\text{holidays}}$  give an overview of the use of chunks in L1 and L2. In  $L2_{\text{holidays}}$ , *aus dem alltag* is in first place, whereas in  $L1_{\text{salary}}$ , *für die gesellschaft* is highest ranked. Ranks 1 to 6 in  $L2_{\text{holidays}}$  are trigrams in which all of the words have been taken directly from the essay title; three out of six  $L1_{\text{salary}}$  trigrams in ranks 1 to 6 contain ‘other’ chunks (ranks 2, 5, and 6). The first ‘other’ chunk in  $L2_{\text{holidays}}$  (*meiner meinung nach*) is in seventh place. In  $L2_{\text{holidays}}$ , the first trigram with a chunk that had been taught in class prior to the essay writing

(*kann man sich* as part of the phrase *kann man sich erholen* ‘can one recover’) comes in at rank 17.<sup>29</sup>

In this context, it is worth noting that the detailed task description for L2<sub>wdt</sub> has probably influenced the text pattern (cf. Skiba 2009 for text patterns by Chinese learners of German in their L1 and L2) and consequently the use of word sequences taken from the task description.

However on the whole, the findings show that L2<sub>holidays</sub> contains extremely high frequencies of trigrams with title words, which can be interpreted as an overuse of these sequences; certainly, this usage differs from the L1<sub>salary</sub> findings.

#### *4.2 Grammatical Content*

We model the grammatical content as bigrams and trigrams of part-of-speech tag sequences (based on the STTS tagset). Table 3 lists the eleven most frequent trigrams of the L2<sub>all</sub> corpus plus two very frequent trigrams of the L1<sub>all</sub> corpus, in a comparison of the ranking  $R_{L2}$  of L2<sub>all</sub> with the ranking  $R_{L1}$  of L1<sub>all</sub>. The two most frequent trigrams in both subcorpora are unsurprisingly “APPR ART NN” (preposition – article – common noun)

---

<sup>29</sup> The frequencies of the top-most chunks in the L1 subcorpus seem to conform to Zipf’s law; that is, frequency =  $C/\text{rank}$ , with  $C$  being the frequency of the most frequent item. The top-most ranks in the L2 subcorpus, however, are less smoothly distributed, tending to form frequency clusters.

and “ART ADJA NN” (article – attribute adjective – common noun), corresponding to a prepositional phrase and a modified noun phrase. Each pattern is illustrated by an example in parenthesis. The fourth column in the table interprets the comparison: “↑” marks an overuse in L2<sub>all</sub> when a pattern is more than ten ranks higher in L2<sub>all</sub> than in L1<sub>all</sub>. “L2 ↓” indicates the corresponding underuse.

**Table 3.** Comparison of the most frequent part-of-speech trigrams.

POS-trigram	R <sub>L2</sub>	R <sub>L1</sub>	Comment
APPR ART NN (e.g., <i>aus dem alltag</i> )	1	1	
ART ADJA NN (e.g., <i>den ganzen tag</i> )	2	2	
NN APPR ART (e.g., <i>flucht aus dem</i> )	3	4 (tie)	
NN ART NN (e.g., <i>tourismus der völkerverständigung</i> )	4	3	
ART NN ART (e.g., <i>der tourismus der</i> )	5	7	
ART NN APPR (e.g., <i>die flucht aus</i> )	6	6	
ART NN VVFIN (e.g., <i>die rechnung bezahlt</i> )	7	8	
ART NN VAFIN (e.g., <i>der urlaub ist</i> )	8	20	↑ L2 overuse
ADJA NN APPR (e.g., <i>vergebliche flucht aus</i> )	9	14	
NN KON NN (e.g., <i>sitten und gebräuche</i> )	10	24	↑ L2 overuse
VVFIN ART NN (e.g., <i>dient der tourismus</i> )	11	52	↑ L2 overuse
ADV ART NN (e.g., <i>auch der staatsanwalt</i> )	21	4 (tie)	↓ L2 underuse
ART NN ADV (e.g., <i>die gesellschaft überhaupt</i> )	24	9	↓ L2 underuse

There is a remarkable overuse of the sequence “VVFIN ART NN” (finite verb – article – common noun) in the L2 texts. This trigram comes in at rank

11 in the  $R_{L2}$  and only at rank 52 in the  $R_{L1}$ . However, this seems to be an artefact of the learners' strategy of repeating parts of the title or the task description, e.g., *Dient der Urlaub* ....

The most striking difference is that none of the top-ranked L2 patterns in  $R_{L2}$  contains an adverb (ADV), while the patterns at rank 4 (tie) and rank 9 in  $R_{L1}$  do. The most frequent L2 trigram using an adverb is "ADV ART NN" at rank 21. The trigram "ADV ADV ADV" (a sequence of three adverbs, e.g., *dann auch noch* 'then also (still)') comes in at place 19 on the L1 list (not depicted here). However, in the L2 list, this trigram is almost at the bottom of the ranking: it is tied at rank 719 with 287 other patterns.

In examining part-of-speech bigrams, a similar picture is obtained: The combination "ADV ADV" (e.g., *auch schon* '(also) already') is at rank 5 in the L1 bigram list; it is also the highest-ranked L1 bigram containing at least one ADV. In contrast, this bigram is only at rank 55 in the L2 list. The highest-ranked L2 bigram containing an ADV is "NN ADV" (e.g., *Urlaub nur* 'holidays only') in place 17 (rank 12 in  $R_{L1}$ ), followed by "ADV ADJD" (adverb – predicatively or adverbially used adjective, e.g., *denn wirklich* '(then) really') at rank 24 (rank 13 in  $R_{L1}$ ). This underuse of adverbs is also reported in the Falko corpus, in which sequences of adverbs are significantly underused in the Falko subcorpora, independent of the learner's L1. Zeldes et al. (2008) explain this by the L2 strategy of learning language with the help of chunks (see also Section 4.1). They assume that elements such as adverbs that occur in highly variable topological contexts

are less easily acquired than more regularly occurring ones. The learners are more uncertain about their use and thus tend to avoid them. This is supported by the fact that in Falko, more regularly occurring sub-classes of adverbs (or adverb sequences) are less underused than average adverbs (sequences).

### *4.3 Complexity*

In order to measure the relative complexity of the L1 and L2 subcorpora, the following aspects have been examined: 1. Type-Token Ratio (TTR) and Vocabulary Growth Rate, 2. length of sentences, 3. depth of clausal embedding, and 4. complexity of the prefields and linguistic material preceding the prefields.

#### *4.3.1 Type-Token Ratio and Vocabulary Growth Rate: Lexical variation*

In the literature, several measures of lexical complexity have been discussed, including Lexical Originality, Lexical Density, Lexical Variation, and Lexical Sophistication (cf. Laufer and Nation 1995: 309f.). For the present study, we focus on lexical variation (i.e., vocabulary variation), for which the *Type-Token Ratio* (TTR) and the *Vocabulary Growth Rate* are two measures frequently applied (cf. Baayen 2008). *Types* are defined as the total number of different words  $V_i$  used (i.e., the vocabulary  $V$ ); *tokens* are the total number of words used. The TTR provides evidence for the lexical

richness of a learner's lexicon and is calculated as follows:

Type-Token Ratio (TTR) = number of types / number of tokens

By this formula, a ratio of 0.5 indicates a frequent repetition of words, while a ratio of 1 means that no words have been repeated; for example, a TTR of 0.2 would indicate that there is one different word (*type*) for every five words used (*tokens*). Table 4 shows the Type-Token Ratio (TTR) for the subcorpora L1<sub>all</sub> and L2<sub>all</sub>. To adequately compare the ratios of the two corpora, it was first necessary to normalise the data (see below).

**Table 4.** Type-Token Ratio in L1<sub>all</sub> and L2<sub>all</sub> (not normalised).

Corpus	Tokens	Types	TTR
L1 <sub>all</sub>	17185	3343	0.195
L2 <sub>all</sub>	11716	1936	0.165

Table 4 indicates that the TTR of the L1<sub>all</sub> corpus is 0.195, compared to 0.165 for the L2<sub>all</sub> corpus. This would mean that the L1<sub>all</sub> corpus is lexically richer than the L2<sub>all</sub> corpus. However, TTR is strongly influenced by corpus size (cf., e.g., Baayen 2008: 224f.). Since L1<sub>all</sub> has more tokens than L2<sub>all</sub>, the difference in lexical richness might be an artefact of the corpus sizes. To test this, the two corpora must be cut down to the same size.

Another method of comparing lexical richness is the measurement of vocabulary growth. This is based on the number of *hapax legomena*, i.e., the

types  $V_I$  in the vocabulary  $V$  that occur only once in the corpus, and the corpus size  $N$ , i.e., the number of tokens (cf. Baayen 2008: 224):

$$\text{Vocabulary Growth Rate} = V_1 / N$$

The Vocabulary Growth Rate predicts the probability that the next word  $N+1$  will be of a new type.

The comparison is based on two test corpora  $L1_{\text{growth}}$  and  $L2_{\text{growth}}$  that consist of the same number of equally sized text fragments: the eleven texts in  $L1_{\text{crime}}$  and the eleven longest texts in  $L2_{\text{all}}$  (independent of topic). Each text was cut down to a fragment of 353 tokens, so that each test corpus consisted of 3883 tokens. Table 5 summarises the results (using the  $R$  function *compare.richness.fnc*, cf. Baayen 2001, Baayen 2008: 225).

**Table 5.** Comparing the lexical richness of  $L1_{\text{growth}}$  and  $L2_{\text{growth}}$ .

Subcorpus	Tokens	Types	TTR	$V_1$	Growth rate
$L1_{\text{growth}}$	3883	1114	0.29	723	0.18620
$L2_{\text{growth}}$	3883	956	0.25	575	0.14808

The  $L1_{\text{growth}}$  corpus has a slightly higher TTR than the corpus  $L2_{\text{growth}}$ , i.e., there are about two different types for every seven tokens in  $L1_{\text{growth}}$  (TTR = 0.29) and for every eight tokens in  $L2_{\text{growth}}$  (TTR = 0.25). The probability that the 3884<sup>th</sup> token in the corpus would be a previously unused word is about 0.18 in  $L1_{\text{growth}}$  and slightly lower (about 0.14) in  $L2_{\text{growth}}$ . The differences in Vocabulary Size (and hence in the Type-Token Ratio) and in

the Vocabulary Growth Rate are significant (according to a two-tailed Chi-Square test, Vocabulary Size:  $Z = 4.8047$ ,  $p = 0$ ; Vocabulary Growth Rate:  $Z = 4.5629$ ,  $p = 0$ ).<sup>30</sup>

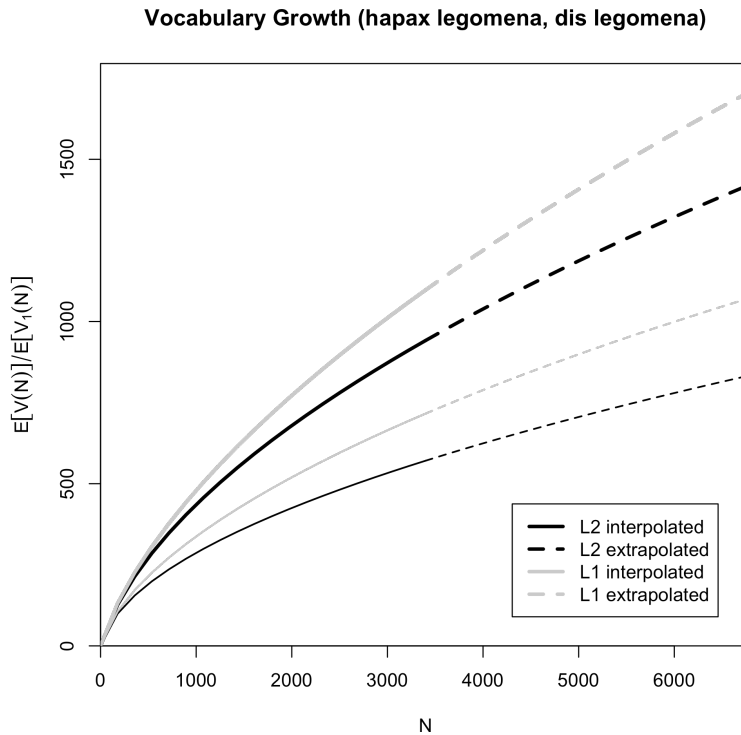
Figure 7 visualises the Vocabulary Growth Rates for *hapax legomena*  $V_1$  and for *dis legomena*<sup>31</sup>  $V_2$  in  $L1_{\text{growth}}$  and  $L2_{\text{growth}}$ . The continuous lines depict interpolations of observed ratios. The dashed lines extrapolate the ratios for unseen corpus sizes according to a fitted LNRE model (cf. Baayen 2008: 235).

**Figure 7.** Vocabulary Growth in  $L1_{\text{growth}}$  and  $L2_{\text{growth}}$ .

---

<sup>30</sup> The growth rate is estimated on the basis of a statistical LNRE model (cf. Evert and Baroni 2007); in our case, the finite Zipf-Mandelbaum model.

<sup>31</sup> *Dis legomena* are the types  $V_2$  in the vocabulary  $V$  that occur exactly twice in the corpus.



The curves in Figure 7 show that  $L1_{\text{growth}}$  contains more *hapax legomena* (upper grey curve) and *dis legomena* (lower grey curve) than  $L2_{\text{growth}}$  (corresponding black curves) at any point in the corpus. Furthermore, the L1 curves are steeper than the L2 curves, i.e., the probability of encountering a new type (or a type seen only once before, in the case of *dis legomena*) is larger in  $L1_{\text{growth}}$  than  $L2_{\text{growth}}$ .

#### 4.3.2 Syntactic variation: Sentence length

We assume that L1 texts contain longer sentences in terms of words than the L2 texts. To evaluate this hypothesis, we measured the length of the matrix

clauses (in words) as annotated in EXMARaLDA; punctuation was excluded from the analysis (see Table 6).

**Table 6.** Length of matrix clauses in L1<sub>all</sub> and L2<sub>all</sub> (in words).

Corpus	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
L1 <sub>all</sub>	1	9	13	14.87	19	78
L2 <sub>all</sub>	1	7	10	11.12	14	38

The values given in the table are the minimum (Min.), the value of the first quartile (i.e., the value of the first quarter of all values, 1st Qu.), the median, the mean, the value of the third quartile (3rd Qu.) and the maximum (Max.). The average sentence in the L1 texts is significantly longer than those in the L2 texts (according to a non-parametric Wilcoxon rank sum test with continuity correction:  $W = 775098$ ,  $p < 0.001$ ).

#### *4.3.3 Syntactic variation: Clausal embedding complexity*

In order to describe another form of syntactic variation, we also analysed clausal embedding complexity, i.e., the depth of clausal embedding. Based on impressions we got from reading the L1 and L2 texts, we expect that the clausal embedding complexity will be lower in the L2<sub>all</sub> subcorpus than in L1<sub>all</sub>. A matrix clause shows a clausal embedding of 1; a constituent clause on the third hierarchical level is marked as 4 (one matrix clause + three

constituent clauses). Table 7 shows the per-text clausal embedding complexity in L1<sub>all</sub> and L2<sub>all</sub>.

**Table 7.** Per-text clausal embedding complexity in L1<sub>all</sub> and L2<sub>all</sub>.

Corpus	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
L1 <sub>all</sub>	2	3	4	3.667	4	5
L2 <sub>all</sub>	2	2	3	2.721	3	4

Table 7 indicates that the minimum clausal embedding for both L1<sub>all</sub> and L2<sub>all</sub> is 2 (i.e., one matrix clause and one constituent clause); the maximum values for L1<sub>all</sub> and L2<sub>all</sub> are 5 and 4, respectively. The mean values are 3.667 (L1<sub>all</sub>) and 2.721 (L2<sub>all</sub>), respectively. The L2 values for clausal embedding are significantly lower than the L1 values (Wilcoxon rank sum test with continuity correction:  $W = 206$ ,  $p < 0.001$ ). This means that the clausal embedding complexity is lower in the L2 texts than in the L1 texts, i.e., they contain less deeply embedded constituent clauses.

#### *4.3.4 Syntactic variation: Length of the prefield and linguistic material preceding the prefield*

In the model of topological fields (e.g., Höhle 1986), declarative main clauses in German are organised into different fields, as illustrated by Figure 4 in Section 3.2. As stated in the model, the *prefield* is the constituent that precedes the finite verb. To measure syntactic complexity, the length of the

*prefield* in the L1 and L2 subcorpora was also measured.<sup>32</sup> Sentences that also contained a pre-prefield were not taken into account in this analysis. Table 8 shows the length of prefields in L1<sub>all</sub> and L2<sub>all</sub>, measured in words.

**Table 8.** Length of the prefields in L1<sub>all</sub> and L2<sub>all</sub> (in words).

Corpus	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
L1 <sub>all</sub>	1	1	1	2.476	2.250	42
L2 <sub>all</sub>	1	1	1	2.293	3	28

According to Table 8, the median prefield length for both L1<sub>all</sub> and L2<sub>all</sub> is only one word. The maximum of L1 is higher than that of L2 (maximum: 42 words vs. 28 words) and correspondingly also its mean (mean: 2.476 words vs. 2.293 words). However, the maximum of 42 is due to an extreme outlier in the L1 distribution (the second-longest L1 prefield is 23 words long). In addition, the third quartile of L1 is lower than that of L2. This means that most of the prefields of L1<sub>all</sub> are shorter in terms of words than those in L2<sub>all</sub>, even though some prefields in L1<sub>all</sub> are much longer than those in L2<sub>all</sub>. However, this difference is not statistically significant (Wilcoxon rank sum test with continuity correction:  $W = 41357$ ,  $p = 0.5466$ ). In conclusion, there

---

<sup>32</sup> Grammatical function could be used as an additional parameter to measure the complexity of the prefield; see Breckle and Zinsmeister (in press) for a study that reveals no substantial variation in the data.

is no significant difference in the length of the prefields in the L1 subcorpus and the L2 subcorpus.

As another measure for complexity, we analysed the linguistic material that precedes the prefield VF in the pre-prefield VVF:<sup>33</sup> these include coordinating conjunctions (such as *doch* ‘but’) and left dislocations (cf. Example 1, wdt 07\_07).

- (1) [Doch]<sub>VVF</sub> [während des Urlaubs]<sub>VF</sub> kann man weit vom Alltag sein.  
 But during the holidays can one far\_away from everyday life be  
 ‘But one can be far away from everyday life during holidays.’

The ratio between sentences with a prefield and those with an additional pre-prefield was calculated for L1<sub>all</sub> and L2<sub>all</sub> (see Table 9).

**Table 9.** Ratio of pre-prefields to prefields in L1<sub>all</sub> and L2<sub>all</sub>.

Corpus	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
L1 <sub>all</sub>	0.11	0.1675	0.22	0.2392	0.2650	0.44
L2 <sub>all</sub>	0	0.045	0.07	0.09209	0.14	0.32

Table 9 shows that the ratio of sentences with and without pre-prefields in the two subcorpora differs greatly: The minimum and maximum ratios in

<sup>33</sup> We have labeled this the ‘complex prefield’. It should be differentiated from cases in which more than one constituent occurs in the prefield proper (cf. Müller 2003).

L1<sub>all</sub> are 11% and 44%, respectively, in comparison to 0% and 32% in L2<sub>all</sub>. In the first quartile, L1<sub>all</sub> exhibits a ratio of pre-prefields four times higher than that of L2<sub>all</sub> (16.75% vs. 4.5%); the median ratio is three times higher (22% vs. 7%), and the mean value 2.5 times higher (23.92% vs. 9.2%). These results reveal a significant difference between the L1<sub>all</sub> and L2<sub>all</sub> distributions (Wilcoxon rank sum test with continuity correction:  $W = 118.5$ ,  $p < 0.001$ ), indicating that the L2 learners use significantly less linguistic material prior to the prefield proper.

#### *4.4 Conclusions*

The quantitative studies show that the L2 learners overuse sequences from the title and the task description, and that they often use chunks that had previously been introduced in class. They use grammatical patterns similar to those of the L1 authors; however, they underuse adverbs. With respect to complexity, the L2 texts are lexically poorer than the L1 texts. They also contain on average shorter sentences, which are also less deeply embedded than those in the L1 texts. However, the L2 texts and the L1 texts do not differ significantly with respect to the length of their prefields, although there is more variance in the L1 prefields. As found for the depth of embedding, the L2 texts contain significantly fewer pre-prefields than the L1 texts.

## 5. Applications for the corpus

One application scenario for the ALeSKo corpus is in teaching German as a foreign language. Most applications of corpora for teaching have thus far been developed for teaching English as a foreign language (cf., e.g., Mukherjee 2002, 2008 and Römer 2008). A discussion focussing on corpus linguistics and German as a foreign language was initiated by Fandrych and Tschirner (2007) in the journal *Deutsch als Fremdsprache*, and has been continued by a series of articles on various aspects of this topic (e.g., Meißner 2008 and Lüdeling et al. 2008). Lüdeling and Walter (2009, 2010) propose a variety of approaches for the use of corpora in language teaching (for teachers, educationalists, and learners) as well as in language acquisition research. While their suggestions for language teaching are based on L1 corpora containing both qualitative aspects (e.g., concordances) and quantitative aspects (e.g., frequency lists), their focus for corpora in language acquisition research is on error analysis and contrastive (interlanguage) analysis (cf. Granger 2008).

We would like to advocate a combination of those two approaches by briefly illustrating how the ALeSKo corpus could be integrated into the teaching of German as a foreign language and into teacher training and development. Both with learners in class and with teachers in training, the results of the quantitative descriptive analyses presented in Section 4 of this

article could be addressed – namely, the significant differences in lexical content (chunks), lexical richness, length of sentences, depth of clausal embedding, and linguistic material preceding the prefield. All of these phenomena have shown an over- or underuse in the L2 subcorpus in comparison to the L1 data. Data-driven learning, which requires corpus literacy (i.e., corpus training), seems to be an appropriate approach for both advanced learners of German and teachers. In this type of exercise, students and teachers would receive carefully chosen input (e.g., text sections) from the L1 and L2 subcorpora, in which they would try to discover the phenomenon in question. The phenomenon would not only be illustrated by concrete examples, but should also be supported by the results of the quantitative descriptive analyses in order to allow visualisation of the differences in L1 and L2 usage.

### **Acknowledgements**

We would like to thank our Chinese student participants at HTWG Konstanz and our student annotators. Heike Zinsmeister's research was financed by Europäischer Sozialfonds in Baden-Württemberg.

### **References**

- Aguado, K. 2002. "Formelhafte Sequenzen und ihre Funktionen für den L2-Erwerb". *Zeitschrift für Angewandte Linguistik* 37: 27-49.
- Anderson, J. 2000. *Cognitive Psychology and its Implications*. Fifth Edition.

New York: Worth Publishers.

Baayen, R. H. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. 2008. *Analyzing Linguistic Data: A practical introduction to statistics*. Cambridge: Cambridge University Press.

Baroni, M., and Lenci, A. 2010. "Distributional Memory: A General Framework for Corpus-Based Semantics". *Computational Linguistics*, 36, 673-721.

Breckle, M., and Zinsmeister, H. 2010. "Zur lernersprachlichen Generierung referierender Ausdrücke in argumentativen Texten". In *Textmuster: schulisch – universitär – kulturkontrastiv*, D. Skiba (ed.), 79-101. Frankfurt/Main: Peter Lang.

Breckle, M., and Zinsmeister, H. (forthcoming). "A corpus-based contrastive analysis of local coherence in L1 and L2 German". In *Discourse and Dialogue Studies between Theory, Research Methods, and Application / Diskurs- und Dialogforschungen zwischen Theorie, Methodik und Anwendung*, M. A. Varga, V. Karabalic, and L. Pon (eds.). Frankfurt/Main: Peter Lang.

Breckle, M., and Zinsmeister, H. 2009. *Annotationsrichtlinien "Funktion des Vorfelds"*. Draft. December 2009. Pedagogical University Vilnius and University of Konstanz.

CEFR [online]:

[http://www.coe.int/t/dg4/linguistic/Source/Framework\\_EN.pdf](http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf)

(accessed 30.11.2011).

Díaz-Negrillo, A., Meurers, D., Valera, S., and Wunsch H. 2010. "Towards interlanguage POS annotation for effective learner corpora in SLA and FLT". *Language Forum* 36 (1-2): 139-154.

Diewald, N., Stührenberg, M., Garbar, A., and Goecke, D. 2008. "Serengeti – Webbasierte Annotation semantischer Relationen". *Journal for Language Technology and Computational Linguistics (JLCL)* 23, 74-93.

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*.  
Magisterarbeit, Humboldt-Universität zu Berlin.

Ellis, N. 1996. "Sequencing in SLA: Phonological memory, chunking, and points of order". *Studies in Second Language Acquisition* 18: 91-126.

Evert, S., and Baroni, M. 2007. "zipfR: Word frequency distributions in R".  
In *Proceedings of the 45th Annual Meeting of the ACL, Posters and Demonstrations Sessions*, 29-32, Prague, Czech Republic.

Fandrych, C., and Tschirner, E. 2007. "Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivenwechsel". *Deutsch als Fremdsprache* 44 (4): 195-204.

Granger, S. 2008. "Learner corpora". In *Corpus Linguistics: An International Handbook* [Handbooks of Linguistics and Communication Science 29], A. Lüdeling and M. Kytö (eds.), 259-275. Berlin et al.: de

Gruyter.

Hickey, T. 1993. "Identifying formulas in first language acquisition".

*Journal of Child Language* 20: 27-41.

Höhle, T. 1986. "Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie

der topologischen Felder". In *Textlinguistik contra Stilistik? –*

*Wortschatz und Wörterbuch – Grammatische oder pragmatische*

*Organisation von Rede?*, W. Weiss, H. E. Wiegand, and M. Reis (eds.),

329-340. Tübingen: Niemeyer.

ICLE [online]: <http://www.uclouvain.be/en-cecl-icle.html> (accessed

30.11.2011).

ICLE – titles [online]: <http://www.uclouvain.be/en-317607.html> (accessed

30.11.2011).

Laufer, B., and Nation, P. 1995. "Vocabulary size and use: Lexical richness

in L2 written production". *Applied Linguistics* 16 (3): 307-322.

Lodewick, K. 1999. *Gegensätze Neu. Ein Programm für die Mittelstufe*

*Deutsch als Fremdsprache*. Göttingen: Fabouda.

Lüdeling, A., and Walter, M. 2009. "Korpuslinguistik für Deutsch als

Fremdsprache. Sprachvermittlung und Spracherwerbsforschung".

Extended version of Lüdeling, A., and Walter, M. 2010.

<https://www.linguistik.hu->

[berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-](https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-)

[innen/anke/pdf/LuedelingWalterDaF.pdf](https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/pdf/LuedelingWalterDaF.pdf) (accessed 30.11.2011).

Lüdeling, A., and Walter, M. 2010. "Korpuslinguistik". In *Handbuch*

- Deutsch als Fremd- und Zweitsprache* [Handbooks of Linguistics and Communication Science 35], H.-J. Krumm, C. Fandrych, B. Hufeisen, and C. Riemer (eds.), 315-322. Berlin et al.: de Gruyter.
- Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K., and Walter, M. 2008. "Das Lernerkorpus Falko". *Deutsch als Fremdsprache* 45 (2): 67-73.
- Meißner, C. 2008. "Eine gebrauchtorientierte Beschreibung des Sprachsystems mit Hilfe der Korpuslinguistik – das Beispiel der Synonyme *ewig* und *unendlich*". *Deutsch als Fremdsprache* 45 (1): 8-13.
- Miller, G. 1956. "The magical number seven, plus or minus two: Some limits on our capacity for processing information". *Psychological Review* 63 (2): 81-97.
- Müller, S. 2003. "Mehrfache Vorfeldbesetzung". *Deutsche Sprache* 31, 29-62.
- Mukherjee, J. 2002. *Korpuslinguistik und Englischunterricht: Eine Einführung*. Frankfurt/Main: Peter Lang.
- Mukherjee, J. 2008. *Anglistische Korpuslinguistik: Eine Einführung*. Berlin: Erich Schmidt.
- Myles, F., Hooper, J., and Mitchell, R. 1998. "Rote or rule? Exploring the role of formulaic language in classroom foreign language learning". *Language Learning* 48 (3): 323-363.
- Newell, A. 1990. *Unified theories of cognition*. Cambridge, MA: Harvard

University Press.

Newell, A., Posner, P., and Laird, J. 1989. "Symbolic architecture for cognition". In *Foundations of cognitive science*, M. Posner (ed.), 93-131. Cambridge, MA: MIT Press.

Orăsan, C. 2003. "PALinkA: A highly customizable tool for discourse annotation". In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, 39-43.

Peters, A.M. 1983. *The units of language acquisition*. New York: Oxford University Press.

Raupach, M. 1984. "Formulae in second language speech production". In *Second language productions*, H. W. Dechert, D. Moehle, and M. Raupach (eds.), 114-137. Tübingen: Narr.

Reznicek, M., Walter, M., Schmid, K., Lüdeling, A., Hirschmann, H., and Krummes C. 2010. *Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 1.0*. [http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch\\_Korpusaufbau\\_und\\_Annotationen\\_v1.0.1](http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau_und_Annotationen_v1.0.1) (accessed 30.11.2011).

Römer, U. 2008. "Corpora and language teaching". In *Corpus Linguistics. An International Handbook* [Handbooks of Linguistics and Communication Science 29], A. Lüdeling and M. Kytö (eds.), 112-131. Berlin et al.: de Gruyter.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. 1999. *Guidelines für*

- das Tagging deutscher Textcorpora mit STTS*. Technischer Bericht.  
Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.
- Schmid, H. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In *New Methods in Language Processing*, D. H. Jones and H. Somers (eds.), 154-164. London: UCL Press.
- Schmid, H. 1995. "Improvements in part-of-speech tagging with an application to German". In *Proceedings of the ACL SIGDAT Workshop*, March 1995.
- Schmidt, T. 2004. "Transcribing and annotating spoken language with EXMARaLDA". *Proceedings of the LREC-Workshop on XML based richly annotated corpora*, Lisbon.
- Selinker, L. 1972. "Interlanguage". *International Review of Applied Linguistics in Language Teaching* 10 (3): 209-231.
- Skiba, D. 2009. *Schriftliches Argumentieren in der Fremdsprache. Eine explorativ-interpretative Untersuchung von Interimstexten chinesischer Deutschlerner*. Tübingen: Narr.
- Stefanowitsch, A., and Gries, S. 2003. "Collostructions: Investigating the interaction of words and constructions". *International Journal of Corpus Linguistics* 8 (2): 209-243.
- Wray, A. 1992. *The focusing hypothesis*. Amsterdam: J. Benjamins.
- Zeldes, A., Lüdeling, A., and Hirschmann, H. 2008. "What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data". In *Proceedings of Quantitative Investigations in Theoretical*

*Linguistics 3 (QITL-3)*, Arppe, A., Sinnemäki, K., and Nikanne, U.

(eds.), 74-77. Slides:

[http://www.ling.helsinki.fi/sky/tapahtumat/qitl/Presentations/Zeldes\\_et\\_al.ppt](http://www.ling.helsinki.fi/sky/tapahtumat/qitl/Presentations/Zeldes_et_al.ppt) (accessed 30.11.2011).

Zeldes, A., Ritz, J., Lüdeling, A., and Chiarcos, C. 2009. “ANNIS: A search tool for multi-layer annotated corpora”. In *Proceedings of Corpus Linguistics 2009*.

Zinsmeister, H., and Breckle, M. 2010. “Starting a sentence in L2 German – Discourse annotation of a learner corpus”. In *Semantic Approaches in Natural Language Processing: Proceedings of the Conference on Natural Language Processing 2010*, M. Pinkal, I. Rehbein, S. Schulte im Walde, and A. Storrer (eds.), 181-185. Saarbrücken: unversaar.