Learner Corpora in Second Language Research: An Example Study in German as a Foreign Language[1]

Denisa Bordag
Herder Institute, Leipzig University
denisav@rz.uni-leipzig.de

Magdalena Sieradz
Herder Institute, Leipzig University
magdasieradz@yahoo.de

Abstract:
In this corpus based study we explored the acquisition of the L2 German perfect (Perfekt) and passive. These analytical forms exhibit structural analogies, which together with their usual acquisition sequence (first perfect, then passive) prepare challenging grounds for investigation of acquisition of these structures.

Contrary to existing studies on related topics (Blackshire-Belay, 1991; Dietrich et al., 1995; Clahsen, 1997; 1999; Attaviriyanupap, 2006), we explored the two structures in their morphosyntactic complexity and did not focus only on a one aspect of their acquisition.

The basis of our research was the FALKO corpus of German as an L2. A thorough quantitative and qualitative analysis of all target structures resulted in number of novel insights concerning, e.g. the interaction of the involved morphosyntactic features in the acquisition process, the role of type and token frequency especially in the acquisition of regular and irregular verbs, the factors affecting the determination of default forms, the influence of earlier acquired structures on later acquired related structure, discrepancies between mental and linguistic grammars especially with respect to defining boarders between "different" linguistic phenomena, and the modelling of lexical representations and processing of the investigated structures, especially reflecting the dynamic character of interlanguage.

Key words:
corpus linguistics, second language acquisition, German, perfect, passive

Introduction

Research in second language acquisition (SLA) is primarily possible by thoroughly examining the data produced by second language learners. There are two basic types of such data:

1) data collected in on-line or off-line experiments (including e.g. questionnaires or various types of tests);
2) authentic and semi-authentic data produced by learners in both natural and classroom settings.

1 The following article is based on data reported in greater detail also in: BORDAG, Denisa – SIERADZ, Magdalena (2012): Erwerb von Perfekt und Passiv bei DaF-Lernern: Eine Korpusstudie. *German as Foreign Language*, 1/2012, pp. 1–26. A German version of this paper was also presented at the FaDaF 2011 conference and is being prepared for publication in the corresponding proceedings.

In this study, we focus on the second type of data and its analysis. The most suitable way to store authentic L2 data is in the form of systematically prepared and well-documented learner corpora. Contrary to mere collections of errors, corpora also provide the context in which the error appeared, and allow for comparisons to be made between correct and deviating uses of the same structure by one learner and between learners, e.g. to assess variation within and between subjects. In addition, quantitative comparisons to L1 data and data of different learner groups (with another L1, of a different proficiency level, taught with a different method, etc.) are also possible. In recent years, a substantial increase of both the number of learner corpora and SLA studies based on them has been observed (cf. Lüdeling – Walter, 2009; 2010).

In the present paper, we introduce a corpus study concerned with the acquisition of perfect and passive in German and performed with data extracted from the FALKO corpus (FehlerAnnotiertes LernerKOrpus). Focusing on three particular types of findings, we want to demonstrate how corpus based L2 research can contribute to advances in L2 research methodology, theory and instruction.

Perfect and Passive in German

The term *Perfekt* denotes in German analytical past tense forms that consist of the auxiliary verbs *haben* or *sein* and of the past participle of another verb, e.g. *(ich) habe gespielt / (ich) bin gekommen*. The finite auxiliary verb stands in the present tense (cf. Duden Grammatik, 2006, p. 469).

The term *Passiv* comprises different language phenomena depending on the linguistic perspective. The term can be used narrowly to denote only those analytical forms with the auxiliary *werden*, i.e. only for the construction known also as *Vorgangspassiv* ("process passive", *werden* + past participle). This view, which is primarily based on the morpho-syntactic structure of the construction, is found particularly in older literature (e.g. Admoni, 1970). More recent works prefer the extended view of passive that also includes the so-called *Zustandspassiv* ("state, result passive"), i.e. the constructions consisting of the auxiliary *sein* + past participle. This approach, however, requires a combination of a morpho-syntactic perspective with a semantic approach. This is the only way to identify *Zustanspassiv* as passive and distinguish it from other, similarly constructed language structures (see Helbig – Kempter, 1997, pp. 32ff.). The third view extends the notion of passive even further, so that it also comprises structures with *bleiben*, *bekommen*, *kriegen*, *haben* etc. as auxiliaries. However, the linguistic categorization of these structures is not completed yet and they are treated differently by different authors. Helbig and Kempter (1997, p. 48) speak in this context of passive paraphrases (*Passiv-Paraphrasen*), which include the so-called *Dativpassiv* (*Adressatenpassiv* or *Rezipientenpassiv*).

In the present study, we explore both *Vorgangspassiv* and *Zustandspassiv*. Passive paraphrases are not considered, since they hardly ever occurred in the corpus.

Perfect and passive show structural analogies in German that provide a challenge when investigating their L2 acquisition. Unlike most similarly oriented studies (e.g., Blackshire-Belay, 1991; Dietrich – Klein – Noyau, 1995; Clahsen, 1997; 1999; Attaviriyanupap, 2006), which only deal with certain aspects of the acquisition of the two structures, we analyze the acquisition of perfect and passive in their morpho-syntactic complexity. In this particular

study, we focus on three findings that contribute to the L2 research methodology, theory and instruction, respectively.

Both perfect and passive consist of an auxiliary verb and a participle, which is typically constructed with the prefix *ge-* and, depending on whether it is a regular or irregular verb, the suffix *-t/-en*. The position of the prefix depends primarily on the presence of other prefixes and their status. If the prefix is separable, *ge-* is located between it and the stem (e.g. *ausgestorben – died out*). Participles of verbs with inseparable prefixes do not have *ge-* (e.g. *versprochen – promised*). The auxiliary in perfect is *haben/sein* and its choice is governed by syntactic and semantic properties of the verb. The auxiliary in passive is *werden/sein* depending on the properties of the verb and the meaning that the speaker intends to express (process/state). The word order is inverted in subordinate clauses, so that the finite form of the auxiliary follows the participle.

There are relatively few studies concerned with the acquisition of perfect and passive in German (see also Attaviriyanupap, 2006, or Wegener, 1998). Most studies on the acquisition of the inflectional paradigms of verbs focus on the selection of verbal categories like person, number, tense, etc. (see Dietrich – Klein – Noyau, 1995, and von Stutterheim, 1986). Others deal primarily with the acquisition of personal inflection as a manifestation of subject-verb agreement (e.g., Blackshire-Belay, 1991; Köpcke, 1987). Syntactic aspects such as the word order of the finite and infinite form or the selection of the auxiliary verb are not included.

In psycholinguistics, there are several studies by Clahsen and his colleagues (Clahsen, 1997; 1999; Clahsen et al., 2001; Clahsen – Rothweiler, 1993, and others), who conducted various experiments in order to explore how adults, children and L2 learners of German produce and understand the participle form, especially to gain insights about the mental representation of regular and irregular verbs.

The basis of our study is the error-annotated learner corpus of German as a foreign language FALKO, which was created at the Institute of German Language and Linguistics at the Humboldt University in Berlin. The corpus consists of three subcorpora: Falko Essays L2, Falko Georgetown L2, and Falko Summaries L2. The subject of this study is the subcorpus Falko Georgetown L2. This subcorpus includes texts from intermediate learners, who studied German as a foreign language at the German Department of the Georgetown University, Washington, DC. It includes 208 written texts (126 105 word forms) and covers the following genres: letters, fiction, newspaper articles, talks and book reviews.[2]

The Analysis of the Corpus in General

The past participle is a part of several constructions. In the Falko corpus, the past participle is annotated according to the Stuttgart-Tübingen Tagset with the abbreviation VVPP. A direct search for perfect or passive is not possible.

The query for VVPP in the Falko Georgetown subcorpus resulted in 2028 hits. The search results were examined manually and it was found that 240 (11.8 %) hits were false positives. These were typically present or infinitive forms that are homonym with perfect (such

2 The error annotation of Falko, which was intended and is now also available, had not yet been finished when the presented research was conducted.

as *studiert*, *versucht*, *verlassen*), forms that are similar to the past participle on the surface, especially because of starting with the syllable *ge-* (e.g. *gelingt*), or adjectives in combinations of a copula and an adjective (e.g. *sie sind verschieden*, *sie sind sich bewusst* etc.). 63 (3.1 %) cases could not be analyzed reliably and in 36 (1.8 %) cases, the past participle was mistakenly used instead of another form, often instead of the 3rd pers. sg.: **Man kann viel bei dieses Buch gelernt*, **wir können uns verloren*, **Jeden Tag sah er die Muslimen gebeten*, **weil es von einem wichtigen Thema gehandelt* and others.

The distribution of the remaining 1689 analyzable past participle forms in the various analytical constructions is summarized in Table 1.

	Hits	%	Number of errors (without inversion)
Perfect	951	56.3 %	146 (15.4 %)
Passive	439	26.0 %	145 (33 %)
Plusquamperfect	169	10.0 %	
Conjunctive Perfect	35	2.1 %	
Conjunctive Plusquamperfect	45	2.7 %	
Form with a modal verb[3]	50	3.0 %	

Table 1: Number of analyzable results for the individual analytical forms with past participle.

The results confirm that the past participle most commonly occurs as a part of perfect and passive (82.3 %).

Out of the 1390 participles that occurred in passive or perfect, 92 were built incorrectly (6.6 % of both investigated analytical verb forms). 66 participles were a part of the perfect tense (about 7 % of all participles that occurred in perfect), and 26 participles were a part of passive (5.9 % of all participles that occurred in passive). A detail report on all errors observed in the study can be found in Bordag and Sieradz (2012).

650 (46.8 %) verbs were irregular and 740 (53.2 %) regular. The proportions are different for perfect and for passive and are therefore presented separately.

As Table 2 reveals, out of 951 verb forms in perfect, 505 (53.1 %) were irregular and 446 (46.9 %) regular. However, only 131 different irregular verbs (lemmas) were used compared to 183 regular. If we compare the number of different irregular and regular stems (i.e. without prefixes), then this becomes 75 to 155. Thus, the ratio stem : token is 1 : 6.7 with irregular verbs and only 1 : 2.9 with the regular ones.

From a total of 439 verb forms in passive, 145 (33 %) were irregular and 294 (67 %) regular. With respect to the number of different verbs, there were 59 different irregular verbs and 167 regular ones and only 39 different irregular stems vs. 144 regular). The stem : token ratio was thus 1 : 3.7 with irregular verbs and 1 : 1.8 with regular verbs.

3 "Forms with a modal verb" are mostly passive forms in which the auxiliary is in infinitive and the modal verb is in its finite form (e.g. *Der Brief muss geschrieben werden*). These forms were coded separately, since they are not relevant for the error analysis of the finite forms of the auxiliaries.

	token	error	lemma/stem	ratio stem : token
perfect 951				
regular	446 (46.8 %)	29 (6.5 %)	183/155	1 : 2.9
irregular	505 (53.2 %)	26 (5.1 %)	131/75	1 : 6.7
passive 439				
regular	294 (67 %)	16 (5.4 %)	167/144	1 : 1.8
irregular	145 (33 %)	10 (6.9 %)	59/39	1 : 3.7

Table 2: The frequency of regular and irregular forms in perfect and passive.

The different proportions of regular and irregular verbs in perfect and passive are due to the fact that a significant number of irregular verbs, especially those that build perfect with *sein*, are unable to form passive. Some of these verbs were very frequent in perfect and their absence in passive results in less pronounced differences in the stem : token ratio.

Although the formation of the participle of irregular verbs usually involves a vowel alternation, stem errors on irregular verbs did not occur more frequently than stem errors on regular verbs. There are at least two reasons for this. As the quantitative analysis has shown, the irregular verbs occur as tokens to individual stems more frequently (each stem is repeated on average 6.7 times in perfect and 3.7 times in passive) than regular verbs and the large number of their occurrences (both in output and probably in the input) enables these forms to be stored correctly. They are also practiced extensively in the classroom and the learner's attention is directed to them explicitly.

In the following sections, we will present in detail data from our corpus analysis that addresses the L2 research methodology, theory and instruction.

L2 Research Methodology: Acquisition sequences

In order to explore acquisition sequences, either cross-sectional data from learner groups at various proficiency levels, or longitudinal data of one (usually rather small) group of learners is typically collected and analyzed with respect to which structures (correct or erroneous) appear at which point of time during acquisition. In our study, we attempt to extract information about the acquisition process by comparing the written production of two constructions with structural analogies that are acquired subsequently, i.e. perfect before passive.

The subcorpus Falko Georgetown L2 is partially longitudinal and includes texts from intermediate learners at the A2–B2 level. At the time that we conducted the research, it was not possible to work with subcorpora of Falko Georgetown L2 that would correspond to particular proficiency levels. However, our approach enabled us to draw conclusions about the details of the acquisition sequence of particular language phenomena. Since perfect is acquired before passive, the differences between the errors in these two structures can reveal interesting aspects about the course of acquisition of their components. This can be demonstrated with three examples: The acquisition of the prefix *ge-,* the suffixes *-en*/*-t* and the inversion.

Errors on the prefix *ge-* were very rare and occurred only in 9 cases. The prefix was omitted four times (e.g., **ankommen* instead of *angekommen*) and it was used incorrectly with verbs that form the past participle without it five times. The small number of errors on the prefix *ge-*shows, that the prefix is very salient and its acquisition relatively easy. However, it is worth

noting that eight errors occurred in perfect and only one error in passive. This means that at the point when the learners start to use passive, the acquisition of the prefix *ge-* has already been successfully completed.

The analyses also revealed that both the regular suffix *-t* and the irregular suffix *-en* were overgeneralized (more in the next section), however, the *-en* overgeneralizations were more frequent in the earlier acquired perfect (14, 3.1 %) than in the later acquired passive (2, 0.7 %). The *-t* overgeneralizations were also rarer (5, 3.4 %) in passive than in perfect (22, 4.3 %), which is in agreement with the general developmental tendency towards less overgeneralizations. However, the decline was smaller than in the case of the overgeneralizations of the irregular *-en* suffix. This suggests that at the time when the learners start to use passive, the information about the irregular status of the *-en* suffix (and on the productivity of the suffix *-t*) is already stored in their mental lexicon rather firmly and its processing is more reliable.

Inversion is another feature whose acquisition process can be observed through the parallel examination of the two constructions. 37 % (351) of the contexts in which a verb in perfect occurred required inversion. In about 89 % (312) of these cases the learners performed the inversion correctly. In the remaining 11 % of the cases they wrongly located the finite form of the auxiliary in front of the infinite form of the lexical verb. On the other hand, in 2 % (18) of cases the learners performed the inversion when the uninverted word order would have been correct.

In passive, the inversion was obligatory in 177 (40.3 %) contexts, and the unmarked word order in the remaining 262 contexts. The learners made an error in only 12 (2.7 %) cases. In 7 cases, the inversion should have been performed, but was not, in 5 cases it was the other way around. This means that at the time when the learners start to use passive constructions, they already use the word order of the finite and infinite part of the analytical predicate correctly: They hardly make any inversion errors and almost never leave out the auxiliary verb.

The simultaneous examination of both constructions shows which features are easiest for the learners to acquire and which are more persistent as a learning problem. With respect to the participle formation, the salient prefix *ge-* is the easiest, while the detection of the marked status of the *-en* suffix takes longer. Also the acquisition of the inversion at the time of passive acquisition is very advanced and the learners only seldom make mistakes.

Theory of L2 acquisition: Identification and storage of defaults as a learning problem

In this section, we show how corpus data can contribute to enhancing and specifying the theories of second language acquisition.

Most models of L2 processing assume the existence of defaults. The term default is typically used for forms and processes which are unmarked, frequent, regular, salient (and the like, according to the theory) and therefore are usually easier to acquire as well and more frequently overgeneralized than other forms and processes. Overgeneralizations of particular stems and suffixes are viewed as one piece of evidence for the modeling of the representation and processing of e.g. regular (default) and irregular forms (e.g. Wunderlich, 1996, or the dual-route models, e.g. Clahsen et al., 2001; Clahsen et al., 2002; Pinker – Ullman, 2002).

Clahsen et al. (2002) assume that stem without alternation, as it appears in infinitives or in 1st pers. sg. present, is the unmarked default, which is most frequently overgeneralized.

The analysis of their longitudinal data on L1 acquisition of German revealed that overgeneralization of the unmarked stems was eight times more frequent than of the marked ones. Experimentally elicited data from children aged 3; 6 and 8; 10 showed that as many as 90 % of errors on the past participle are due to the overgeneralization of unmarked stems (**gehelft* instead of *geholfen*, etc.). Clashen et al. (2002) explain this phenomenon based on the way that irregular verbs are stored in the mental lexicon. Following Wunderlich (1996), the authors assume that the morphological relations between the stems of an irregular verb are represented in a form of a non-monotonic default inheritance hierarchy, in which the stem variants correspond to subnodes in a hierarchically structured lexical entry. On the highest level of the hierarchical entry is the unmarked base form, as it occurs in the infinitive and most present forms. The lower a stem/subnode is in the hierarchy, the more marked and specified it is. A stem error appears when the specification of the subnode is missing or cannot be accessed. In this case, the least specified form of the structured lexical entry is retrieved, i.e. the unmarked base form. This form does not have all of the necessary specifications, but it also does not involve any that would be in contradiction with the required characteristics.

The data from the corpus investigated here, however, shows that the model does not offer an adequate explanation for all data.

1. Incorrectly formed regular and irregular participles occur in the corpus equally often. 35 (5.3 %) participles of irregular verbs and 41 (5.5 %) participles of regular verbs were formed incorrectly. The model would predict that fewer errors should appear in the default regular formation.
2. As mentioned in the previous section, both the regular suffix *-t* and the irregular suffix *-en* are overgeneralized. The suffix *-t* more frequently (27, 4.2 % of all irregular verbs) than the suffix *-en* (8, 16, 2.1 % of all regular verbs). The *-en* overgeneralizations, however, are still relatively common, which contrasts with the assumptions of the model and also with the findings from L1 and L2 that are usually reported in this context (see e.g. Attaviriyanupap, 2006, for conflicting evidence).
3. Stem errors on the forms with the correct suffix are more prevalent in regular (24) than irregular verbs (8), though actually the opposite would be expected. More than two thirds of the errors on regular verbs were due to the absence of the umlaut (this confirms that umlauts have low salience). It is noteworthy that some of the verbs that lacked an umlaut were related to words from other word classes that also have no umlaut (Tot – **getotet* and **getoten* instead of *getötet*, *kalt* – **erkaltet* instead of *erkältet*, Traum – **getraumt* instead of *geträumt*). One of the reasons for these stem errors can thus be the intralingual interference from related forms without an umlaut.
4. No reliable evidence could be found to support the model's prediction that the infinitive stem (default) would be more frequently overgeneralized than the marked stem of the 3rd pers. sg., e.g. **abgenimmt*, **gegibt*, **gehilft*. In 15 incorrectly formed participles, the stem corresponds to the form of the 3rd pers. sg. With 6 of them (**gefält*, **abgenimmt*, **gegibt*, **gehilft* (2×), **entwirft*) it is also obvious that the participle form was derived from the 3rd pers. sg., since these verbs do not have a homonymous present stem (such as **geleidet*, **geschreibt*, **überwindet*).

Eight other participles might appear to be derived from the infinitive (such as **beigetragt*, **gelauft*, **geschlagt*, **verlasst* and **vorgeschlagt*). In most cases, however, the infinitive stem of these verbs differs from the 3rd pers. sg. form only in that the latter involves an umlaut. Since the learners often omit the umlaut, it is not clear whether the participles were actually derived from the infinitive or from the stem of the 3rd pers. sg.[4] Except for the forms *gelest** (3 times in the corpus) and **vertretet*, no unambigous overgeneralizations of the infinitive stem were observed.

The data from our study shows that the hierarchical model in its present form cannot explain all characteristics of the interlanguage. The main problem seems to be the fact that the model is static and makes assumptions (e.g. the default form is the highest node of the hierarchy) that do not necessarily have to be valid, or not for all acquisitional stages. The recognition and the appropriate storage of the default form, be it the stem or the suffix, should not be considered as a given, but should be viewed as a learning task in itself, which is more or less demanding depending on other properties of the target language system. Learners can also have incorrectly structured lexical entries whose highest node is e.g. the stem of the 3rd pers. sg. If this is the case, then all other assumptions of the hierarchical model would remain valid, while the data from our study would be explained at the same time: It would be the highest node which is overgeneralized and there would also be other overgeneralizations beyond just the overgeneralizations of the infinitive stem.

The frequency of occurrences of irregular (53.1 %) and regular (46.9 %) participles in perfect, whose acquisition precedes the acquisition of passive, is also an indication that at the given acquisitional stage, the default status of the regular suffix for the participle formation does not have to be already recognized: Although the number of regular verbs is much larger than the number of irregular ones, both suffixes occur approximately equally often, if the number of tokens is considered. Moreover, both suffixes are also homonymous and appear regularly in other verb forms (*-t* in the 3rd pers. sg. pres., *-en* even in more forms, e.g. in the infinitive, 3rd pers. pl. pres., etc.).

The above analyses show the importance and the potential of corpus-based studies, which provide information about actual language use. Contrary to the information in grammars and textbooks (e.g. small number of irregular verbs vs. large number of regular ones), corpora make it obvious that e.g. the number of regular and irregular forms in the output and presumably also input is basically equally large. This makes the task of developing the right mechanisms for processing them more difficult (e.g. establishing two routes for processing regular and irregular forms as assumed in the dual-route model, see, e.g. Clahsen, 1997).

The learner's mental language system is constantly reorganized and differs from the target language system not only in that information may be missing in it, but also because it may be organized differently and this organization is more dynamic than an adult native speaker system. The L2 models should be able to stand up to these characteristics of the L2 system.

4 These forms pose a problem for studies that base their conclusions on the overgeneralizations of infinitive stems. Only a direct comparison with errors made on the regular stems, as was done it in our study, makes it obvious that subsuming such forms under the category "derived from an infinitiv stem" is very questionable.

L2 Instruction: Mental vs. linguistic grammar

Theoretically and didactically interesting conclusions can be drawn from our data on frequency of (correctly and incorrectly used) auxiliary verbs in perfect and passive.

The corpus analysis showed that from 951 verb forms in perfect, 76.8 % were formed with the auxiliary *haben* and only 23.2 % with the auxiliary *sein*. In the passive, the formation with *werden* is more frequent (73.1 %) than the formation with *sein* (see Table 3). This data would support the assumption made in most grammars and textbooks, namely that the default formation in perfect is with the auxiliary *haben* and in passive with *werden*.

	Total	Overgeneralized
Perfect	951	
haben	730 (76.8 %)	47 (4.9 %)
sein	221 (23.2 %)	7 (0.7 %)
Auxiliary omitted		24 (2.5 %)
Passive	439	
werden	321 (73.1 %)	4 (0.9 %)
sein	118 (26.9 %)	62 (14.1 %)

Table 3: The frequency of the auxiliary verbs and their overgeneralizations in perfect and passive.

The error analysis shows that the auxiliary *werden* is replaced by the auxiliary *sein* more often than the other way around, although *Vorgangspassiv*, with 73.1 %, is more frequent than *Zustandspassiv* and although *Zustandspassiv* is more restrictive than *Vorgangspassiv* (all verbs which form *Zustandspassiv* can also form *Vorgangspassiv*, but not vice versa). The proportion of auxiliary verbs in perfect was similar (76.8 % took the auxiliary *haben*), and it was this frequently used auxiliary verb which was overgeneralized – this supports the expectations and is consistent with the markedness theory and the overgeneralization of defaults.

The overgeneralization of the less frequent auxiliary *sein* in passive is conspicuous particularly in the context of other overgeneralizations, where it is always the case that the more frequent and typical/regular/unmarked member of the pair is overgeneralized:

1. plural tends to be replaced more often by singular than vice versa;
2. the preterite form of the auxiliary would be replaced with its present form, but never the other way around;
3. active voice was overgeneralized, passive never was.

The overgeneralization of *sein* in passive cannot be explained by the absolute frequency of the verb in the language – *sein* is more frequent in German than both *werden* and *haben*. If the decisive factor would be the absolute frequency, then the auxiliary *sein* would have to be overgeneralized both in perfect and in passive.

The pattern of results of error analysis would rather support the assumption that the default form of the passive is the formation with the auxiliary *sein* and/or that there are other factors that may play a stronger role than un/markedness. One possible explanation is that learners do not interpret *Zustandspassiv* as a predicate construction. A similar approach is also taken by some researchers, who exclude *Zustandspassiv* from their investigations of passive for this reason (e.g. Wegener, 1998). From this perspective, the overgeneralization of *sein* could

be explained in accordance with the markedness theory: It is not the less frequent and more restricted form of passive which is overgeneralized, but the frequent, early acquired copula construction with the auxiliary *sein*.

Reports from the L1 language acquisition and acquisition by adults in natural setting[5] support this hypothesis. According to Fritzenschaft (1994), copula constructions that are similar in morphological and syntactic respects to the passive, but are less complex and more frequent, are acquired earlier and represent precursors and bridges to passive for children. Mills (1985, p. 201) reports that the most common mistake in a sentence repetition test with pre-school children was the replacement of *werden* by *sein*. According to the author, this indicates that *sein* is available earlier and that the pre-school children had not yet recognized the difference between the two auxiliaries. Wegener (1998, p. 157) stipulates a gap of several weeks between the acquisition of *sein*-passive and *werden*-passive in children's L2 acquisition in natural setting.[6]

The assumption that is typically presented in grammars and textbooks, namely that *Zustandspassiv* and *Vorgangspassiv* form together one subsystem (see Figure 1), is not supported by the data in this study. An overgeneralization of a less frequent and marked feature (*sein*-passive) would be a very rare phenomenon and would go against the principles of the development of the L1 and L2 language systems.

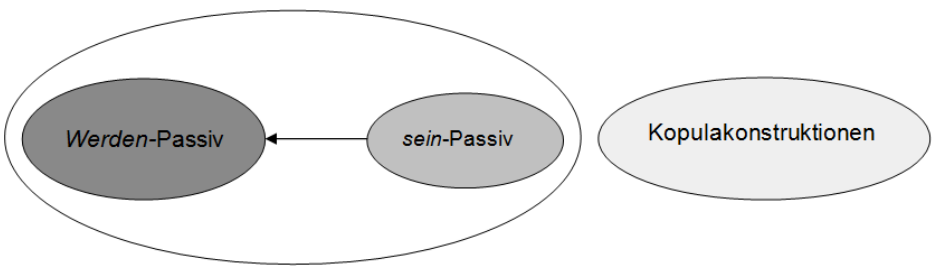

Figure 1: Map of the linguistic grammar.

More probable is the hypothesis that the learner's mental grammar is different from the linguistic grammar and that it is the *sein*-passive form and the copula constructions that together form one subsystem (see Figure 2). Under this assumption, *sein* overgeneralizations in passive would be expected because it is the more frequent and earlier acquired construction.

5 Zweitspracherwerb.

6 Another factor that most likely also contributed to the overgeneralization of the *sein*-passive in our particular subcorpus is the interlingual interference. The L1 of most Georgetown subcorpus subjects was English, which formally does not differentiate between *Vorgangspassiv* and *Zustandspassiv* (at least not the way German does) and both types of passive are expressed with the auxiliary *be + past participle* (e.g. *the letter is/was sent*), i.e. equivallently to the formation of *Zustandspassiv* in German.

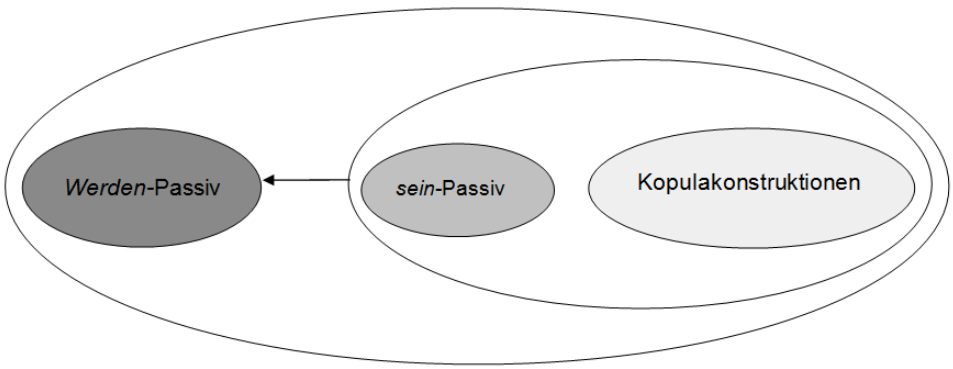

Figure 2: Map of the learner's mental grammar.

The data in this study provides an important indication that the mental subsystems do not always overlap with the subsystems as they are defined linguistically. The distinction between *Zustandspassiv* and copulative constructions may be a useful linguistic construct, but it does not correspond to the way the learner's mental grammar is organized. Under this assumption, the set of linguistic structures, based on which e.g. the learner (unconsciously) derives which auxiliary verb is the unmarked default for a certain construction, form a homogeneous subsystem – however not in a grammar book, but in the mental grammar of the learner. It is important to consider which, if any consequences teachers and authors of textbooks should draw from it. So far, *Vorgangspassiv* has been considered a "prototypical" passive and most lessons focusing on passive start with it. However, the present data and the data from the L1 and German as a second language in natural setting acquisition point out that the acquisition of *Zustandspassiv* is easier for learners and also that *Zustandspassiv* is acquired first in the natural order of the acquisition.

The parallel investigation of the two analytical forms enables one not only to formulate hypotheses and draw conclusions about the differences between the linguistic and the mental grammars, but also to specify how exactly frequency affects language acquisition and to demonstrate that the term frequency itself allows different interpretations. Because the auxiliary verb *sein* is overgeneralized in passive, but not in perfect, it is possible to better determine the scope of influence of the frequency effects: It is not always the absolute frequency of a lexeme in the language (*sein* is more frequent than *haben*), which exerts the decisive influence, but also not always the frequency within a particular subsystem, as it is specified and defined in grammars (in our case passive, in which the auxiliary verb *sein* is much more frequent). The existence and frequency of forms with similar structures and similar functions (such as copula constructions) may prove to be more important.

Conclusion

The availability of annotated electronic corpora of learner language allows us to work with large amounts of data and process it relatively fast. Such an approach enables us to discover correlations between linguistic phenomena or gain other types of knowledge that would not

have become obvious if we had studied mere error collections, small unannotated corpora or if we had conducted controlled experimental studies (see also Lüdeling – Walter, 2009).

The findings from the present study have challenging implications for L2 research methodology, theory and instruction. The study has also shown the advantages and strengths of corpus linguistic analysis. Only corpus-based data provide reliable information about different forms and structures and, above all, about how often they occur in actual usage. As we could see e.g. from the example of lemma vs. token frequency of regular and irregular verbs, some phenomena in the learner's language can be understood better and more fully if not only systemic properties of the language system are considered, but also their usage.

The analysis of the corpus data has shown that the acquisition of perfect and passive is very complex and that the simultaneous observation of both constructs can reveal relationships that would otherwise stay undiscovered. The opportunity to work with large amounts of electronically stored data proved to be very productive, although much manual effort was needed to prepare the data for the analyses.[7]

Nevertheless, it also has to be said clearly that corpus-based studies are primarily hypothesis generating and that their results (including the results of this study) should be tested with other methods, e.g. in psycholinguistic experiments, before definite or far-reaching conclusions are made.

References:

ADMONI, Wladimir (1970): *Der deutsche Sprachbau*. München: Beck.

ATTAVIRIYANUPAP, Korakoch (2006): The acquisition of verb inflection by Thai migrant women in Switzerland. An inventory. *Linguistik Online*, 29(4).

BLACKSHIRE-BELAY, Carol (1991): *Language Contact: Verb Morphology in German of Foreign Workers*. Tübingen: Gunter Narr Verlag.

BORDAG, Denisa – SIERADZ, Magdalena (2012): Erwerb von Perfekt und Passiv bei DaF Lernern: Eine Korpusstudie. *German as a Foreign Language*, 1/2012, pp. 1–26.

CLAHSEN, Harald (1997): The representation of participles in the German mental lexicon: Evidence for the dual-mechanism model. In: Geert Booij – Jaap Van Marle (eds.), *Yearbook of Morphology 1996*. Dordrecht – Boston, MA – London: Kluwer Academic Publishers, pp. 73–96.

CLAHSEN, Harald (1999): Lexical entries and rules of language: a multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, 22, pp. 991–1060.

CLAHSEN, Harald – EISENBEISS, Sonja – HADLER, Meike – SONNENSTUHL, Ingrid (2001): The mental representation of inflected words: an experimental study of adjectives and verbs in German. *Language*, 77, pp. 510–543.

CLAHSEN, Harald – PRÜFERT, Peter – EISENBEISS, Sonja – CHOLINE, Joana (2002): Strong stems in the German mental lexicon: Evidence from child language acquisition and adult processing. In: Ingrid Kaufmann – Barbara Stiebels (eds.), *More than Words: a Festschrift for Dieter Wunderlich*. Berlin: Akademie Verlag, pp. 91–112.

7 Here we would like to thank the students of Herder-Institute of the Leipzig University who attended the corpus linguistics seminars in 2010 and partly helped with the collection and annotation of the data.

CLAHSEN, Harald – ROTHWEILER, Monika (1993): Inflectional rules in children's grammars: Evidence from the development of participles in German. In: Geert Booij – Jaap Van Marle (eds.), *Yearbook of Morphology 1992*. Dordrecht – Boston, MA – London: Kluwer Academic Publishers, pp. 1–34.

DIETRICH, Rainer – KLEIN, Wolfgang – NOYAU, Colette (1995): *The Acquisition of Temporality in a Second Language*. Amsterdam – Philadelphia, PA: John Benjamins.

Duden. Die Grammatik (2006). Mannheim: Bibliographisches Institut – F. A. Brockhaus.

FALKO : *Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache*. Accessible from WWW: <http://korpling.german.hu-berlin.de/falko/index.jsp>.

FRITZENSCHAFT, Agnes (1994): Activating passives in child grammar. In: Tracy Rosemarie – Elsa Lattey (eds.), *How Tolerant is Universal Grammar, Essays on Language Learnability and Language Variation*. Tübingen: Niemeyer, pp. 155–184.

HELBIG, Gerhard – KEMPTER, Fritz (1997): *Das Passiv. Zur Theorie und Praxis des Deutschunterrichts für Ausländer*. Berlin: Langenscheidt.

KÖPCKE, Klaus-Michael (1987): Der Erwerb morphologischer Ausdrucksmittel durch L2-Lerner am Beispiel der Personalflexion. *Zeitschrift für Sprachwissenschaft*, 6(2), 186–205.

LÜDELING, Anke – WALTER, Maik (2009): Korpuslinguistik für Deutsch als Fremdsprache. Sprachvermittlung und Spracherwerbsforschung. Stark erweiterte Fassung von Lüdeling/Walter (erscheint) Korpuslinguistik. In: *HSK 19, Deutsch als Fremdsprache*. Berlin – New York, NY: Mouton de Gruyter.

LÜDELING, Anke – WALTER, Maik (2010): Korpuslinguistik. In: Christian Fandrych – Britta Hufeisen – Hans-Jürgen Krumm – Claudia Riemer (eds.), *Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch*. Berlin – New York, NY: Mouton de Gruyter.

MILLS, Anne E. (1985): *The Acquisition of German*. New Jersey, NY: Lawrence Erlbaum.

PINKER, Steven – ULLMAN, Michael T. (2002): The past and future of the past tense. *Trends in Cognitive Sciences*, 6, pp. 456–463.

VON STUTTERHEIM, Christiane (1986): *Temporalität in der Zweitsprache. Eine Untersuchung zum Erwerb des Deutschen durch türkische Gastarbeiter*. Berlin – New York, NY: Mouton de Gruyter.

WEGENER, Heide (1998): Das Passiv im DaZ-Erwerb von Grundschulkindern. In: Heide Wegener (ed.), *Eine zweite Sprache lernen. Empirische Untersuchungen zum Zweitspracherwerb*. Tübingen: Gunter Narr Verlag, pp. 143–172.

WUNDERLICH, Dieter (1996): Minimalist morphology: The role of paradigms. In: Geert Booij – Jaap van Marle (eds.), *Yearbook of Morphology 1995*. Dordrecht – Boston, MA – London: Kluwer Academic Publishers, pp. 93–114.

Abstrakt:
Tato korpusová studie se zabývá akvizicí perfekta a pasiva při osvojování němčiny jakožto cizího jazyka (L2). Oba tyto analytické slovesné tvary mají analogickou strukturu a ustálené pořadí osvojování (nejprve začíná osvojování perfekta, teprve pak pasiva), což z nich činí z hlediska jejich akvizice velmi zajímavou oblast výzkumu. Na rozdíl od jiných studií zaměřených na podobné otázky (viz např. Blackshire-Belay, 1991; Dietrich et al., 1995; Clahsen, 1997; 1999; Attaviriyanupap, 2006) se u obou těchto tvarů nesoustředíme pouze na jeden aspekt jejich akvizice, ale zkoumáme je s ohledem na všechny jejich morfosyntaktické vlastnosti.

Východiskem pro náš výzkum byl korpus němčiny jakožto cizího jazyka s názvem FALKO. Na základě důkladné kvantitativní a kvalitativní analýzy všech relevantních tvarů formulujeme řadu nových poznatků, jež se týkají např. interakce morfosyntaktických vlastností perfekta a pasiva v průběhu jejich osvojování; vlivu frekvence na úrovni typů a tokenů na osvojování pravidelných a nepravidelných sloves; faktorů, které mají vliv na výběr defaultních tvarů; vlivu dříve osvojených tvarů na příbuzné tvary osvojované později; rozporů mezi mentální a lingvistickou gramatikou, a to zejména s ohledem na vymezení hranic mezi „odlišnými“ lingvistickými jevy; a modelování mentálních reprezentací a zpracování sledovaných tvarů, především s ohledem na dynamický charakter osvojovaného jazyka.