This paper will appear in: Grandin, Karl (eds.) *Going Digital. Evolutionary and Revolutionary Aspects of Digitization.* [Nobel Symposium 147]. Science History Publications/USA, New York. [This is the last version I have copyright for.]

o

Corpora in Linguistics: Sampling and annotation[1]

Anke Lüdeling, Humboldt-Universität zu Berlin, anke.luedeling@rz.hu-berlin.de

1. Introduction

Gregory Crane said at the beginning of his talk at the Nobel symposium 'Going Digital' that in linguistics electronic corpora can be used in three ways: (a) they can help answer old research questions faster and more accurately, (b) they make it possible to formulate and answer new research questions through new ways of looking at and analyzing textual data, and (c) they make the data widely available. This paper deals with two aspects of corpora that pertain to all three issues: sampling and annotation. My focus is methodological – I want to describe *how* electronic corpora – if designed and annotated transparently – have contributed to and changed linguistic research.[2] I want to start with the following statement by Hermann Moisl (2009, 876).

> "Data is ontologically different from the world. The world is as it is; data is an interpretation of it for the purpose of scientific study. The weather is not the meteorologist's Data – measurements of such things as air temperature are. A text corpus is not the linguist's data – measurements of such things as average sentence length are."

What Moisl says is that 'the world' must be interpreted. The 'world' for many linguists is a given linguistic variety, be it Modern German, Old English, or the language of teenagers in a certain suburb of Stockholm that they want to analyze. One possible way to interpret the 'world' – here the linguistic variety under question - is to collect a corpus of the variety under question, according to specific design criteria – which then is a **sample** of the world. This corpus can then be interpreted further: Segments of the corpus such as words or phrases or sentences can be classified and analyzed. The classes could be grammatical (in Moisl's statement, for example, the strings <data>, <world>, <world>, <data>, <purpose>, <study>, … could be classified as nouns), rhetorical (in Moisl's statement we could classify the third sentence as exemplifying sentence; in a conversation we can classify the utterances

[1] I want to thank the members of the corpus linguistics group at Humboldt-University for many heated and fruitful discussions about corpus linguistics and no less for more direct contributions to this paper: Amir Zeldes and Florian Zipser gave valuable input to specific points, Hagen Hirschmann and Marc Reznicek read the pre-final draft and their constructive criticism and suggestions made this paper more coherent and accessible. Amir Zeldes and Marc Reznicek helped with the Falko data. I also want to thank the organizers and colleagues at the Nobel Symposium – especially Karl Grandin – for one of the most stimulating and interesting conferences I ever went to.

[2] For many years the dominating issue here was *whether* corpus data should be used to answer 'interesting' (often understood as 'generative') questions (cf. Fillmore's Nobel Symposium article from 1992, Karlsson 2008 and many other papers). Most linguists do not find this issue interesting anymore; it seems uncontroversial meanwhile that corpus data is relevant for many qualitative and quantitative research questions in many frameworks – in recent years the focus of the debate has shifted to the issue of *how* corpus data can be used in linguistic research (cf. among many others for example the Corpus Linguistics and Linguistic Theory Special Issue on Grammar and Grammaticality (2007) or the articles in Kepser & Reis 2005). There are, of course, many linguistics research questions which cannot be answered by using corpus data. In this article I focus only on those questions for which corpus data is helpful.

o

as statements, questions, answers etc.) or anything else one might want to study. If such classification is coded directly in the corpus we speak of **annotation**.
Only f we take both sampling and annotation seriously and make all the decisions involved in them explicit we can perform rigorous experiments with reproducible results. Here lies the real advantage of electronic corpora: In electronic corpora we can document every step of interpretation so that every user is able to see exactly how we arrived at the empirical base for our analyses. We can then use corpus data for exploratory studies in which we look for connections between variables, such as sentence type or part-of-speech category, or find categories, such as subordinate clause or verb, (see Section 4, Moisl 2009), as well as for testing hypotheses that come from theoretical considerations.

The paper is organized in three sections. First I want to show how pre-electronic corpora were used and where the empirical problems lay. The following sections then correspond to the two steps of interpretation mentioned above, sampling (Section 4) and annotation (Section 5). Although this paper is mainly an overview paper I shall illustrate each general point with specific results from a deeply annotated special corpus (a learner corpus called Falko) which is briefly introduced in Section 3.

2. Corpus linguistics: research questions and methods

Many of the research questions and methods in corpus linguistics today date back to pre-electronic times. Long before there were electronic corpora people collected real-life examples (in contrast to made up examples) to develop and illustrate grammars, define lexical entries, argue about different readings of a word etc. (this has often been described, for an overview see e.g. Meyer 2008). In some sense it can be said that (at least part of) linguistics developed to preserve an earlier language stage (the Sanskrit Vedas, as described in the works of Panini in India or Homer's writings in Greece, cf. the 'Ungrammatical Words' by Aristonicus of Alexandria) that was passed on as a body of text – a corpus. A typical use of pre-electronic corpora is illustrated in example (1), from the German Neogrammarian Hermann Paul. An example taken from a real text (a quote, a sentence, a word in context) is used to illustrate a grammatical or lexicographic fact. Paul here writes about a word formation phenomenon that he considers marginal. (Don't worry: The phenomenon itself – the addition of the 'linking element' *s* between the two stems in certain compounds – is not crucial here, I am only concerned with the way the evidence is presented.)

> „Sometimes the linking element *s* appears after feminine nouns, although this is not yet common in written language, compare e.g. *Gemeindsversammlung* 'council + s+ meeting' Hebel 452,24, *Huldszeichen* 'benevolence + s + sign' Heine 2, 111, *über Naturs Größe* 'about nature + s + greatness', *Sprachsverbesserer* 'language + s + improver' Leibniz, Unvorgreifl. Ged. 67,3, *Vernunftswahrheiten* 'sensibility + s+ truths' Le(ssing) 12, 434, 32 …" [3]

[3] My translation of: „Gelegentlich erscheint auch sonst ein *s* in der Kompositionsfuge nach Femininum, ohne daß es in die Schriftsprache durchgedrungen ist, vgl. z.B. *Gemeindsversammlung* Hebel 452, 24, *Huldszeichen* Heine 2, 111, *über Naturs Größe* Le. 11, 209, 5, *Sprachsverbesserer*, Leibniz, Unvorgreifl. Ged. 67,3, *Vernunftswahrheiten* Le. 12, 434, 32. ..." Hermann Paul (1959, Band V, 13).

This paper will appear in: Grandin, Karl (eds.) *Going Digital. Evolutionary and Revolutionary Aspects of Digitization.* [Nobel Symposium 147]. Science History Publications/USA, New York. [This is the last version I have copyright for.]

o

Here, as in many other similar grammars, examples from literary works are used to provide evidence (and thereby lend authority) for the claims made by the grammarian. I do not mean to demean the enormous achievements of grammarians like Paul, but in the light of current empirical standards in linguistics quotations like these are methodologically problematic: Paul does not attempt to explain why in these cases (contrary to what he considers the correct written language) the linker *s* is being used. He does not give any context for his examples. He does not tell us whether the cited authors use the same words (or at least the same non-head words) without the linking element *s*. He does not even give us a definition of what he considers a word or tell us why *Naturs Größe* (which could be considered a phrase) is included in the list. In addition there are two problems with regard to the selection of the cited authors. First, Paul, like most of the other grammarians before (and often after) structuralism, cites only well known 'authoritative' authors. He therefore describes a written literary register without acknowledging this. Second, Paul wants to write a grammar for German at the beginning of the 20th century. The lives of the cited authors span two hundred years (from Leibniz' birth in 1646 to Heine's death in 1856). Language has certainly changed in those 200 years (and the roughly 70 years between Heine's death and Paul's grammar). If we think back to the desiderata formulated in Section 1 (making sampling and annotation explicit) we see that Paul fails in both respects.

The issues are very similar for other areas of linguistics, for example for pre-electronic lexicography (see e.g. Heid 2008). Typically, lexicographers manually collected quotations that consisted of the word to be described and its context. The quotations were then sorted and analyzed (there are impressive collections of paper slips in all lexicographic institutions). While it is amazing how much data some of the scholars processed and kept in mind we cannot say that either sampling method or search was fully systematic.[4] Inherently corpus-based areas like historical linguistics or language acquisition studies had their own issues with respect to availability, access and coverage, but also in these areas pre-electronic texts could not be systematically searched (manual searches often took months or years and added research questions meant that everything had to be read again) and results could not easily be reproduced. The method – producing a concordance of a key word with some context – is, however, still the research method most widely used in corpus studies today (only now the searches can be systematic).

The first electronic corpus was made in the 1940s (Busa 1974, 1980) but it took a long time – essentially until the 1980s or 1990s – until corpora became freely and widely available for linguistic research (computational linguistics used them earlier). For many languages and varieties there are still no electronic corpora. Many influential linguists in the 20th century did not use corpus data and so in many departments corpus construction and analysis were not on the agenda (see FN 2). This has changed considerably. More computing power, internet access, search tools, and standardization have also helped. In recent years we have seen both a tendency towards the development of increasingly larger corpora for linguistic and NLP purposes[5] and a tendency towards many more small dedicated corpora (dialect corpora, historical corpora, spoken corpora, learner corpora, etc.) designed for very specific research questions.

[4] This is not problematic if the purpose is simply to find any example to illustrate a statement. It is problematic if the purpose (as it was explicitly stated for the Oxford English Dictionary) is to cover the range of constructions or meanings and also if the purpose is to discover co-variation. There are a few pre-electronic dictionaries based on a systematic and quantitative study of a corpus (see for example the manually accumulated frequency dictionary by Käding 1897). There are also a few pre-electronic quantitative studies on language data (see e.g. Köhler 2005 for an overview), some of them extremely far reaching and interesting.

[5] One extreme are Web-based corpora (see e.g. Baroni et al. 2009 or Pomikalek/Rychly/Kilgarriff 2009).

o

3. The Falko Corpus and Overuse/Underuse Statistics

I want to illustrate my general points using one of these special corpora, the learner corpus Falko (Lüdeling et al. 2008). Falko contains written essays from advanced learners of German as a foreign language as well as native speaker control data.[6] Learner data is studied in order to find acquisition patterns (see e.g. Granger 2008 for an overview). Learner data often contains errors and these can be indicative of the learners' hypothesis of their target language. The learner in Example (1), for instance, uses the wrong preposition, the verb *vorbereiten* 'to prepare' needs *auf* instead of *für*. Single errors such as this one are not interesting in themselves – they are merely anecdotal. But if many students make the same kind of error again and again this could tell us something about acquisition.

(1) […] *dass ihr Studium sie nicht für* (should be: *auf*) die *wirkliche Welt und ihre berufliche Zukunft vorbereitet* (fk_006_2006_08)
„[…] that their course of studies does not prepare them for the real world and their future job"
.
In a learner corpus we could now see whether the verb *vorbereiten* is often used with the wrong preposition. We could then see whether this is dependent on the learner's native language (if it is English there could be a transfer from *prepare for*) or whether generally the polysemy of prepositions leads to these problems etc. What is important for us here is that we see the learner language that we collected as a sample for the learner language of a given population and that errors need to be found and classified. Many learner corpus studies concentrate on finding and classifying errors (error analysis, EA). Other studies compare the learner corpus quantitatively with another corpus – this can be a native speaker corpus or a different learner corpus (contrastive interlanguage analysis, CIA). In the following I will speak about a combination of EA and CIA.
Before I introduce the method used here I need to say a few words about learner language[7]. The language produced by learners is not random. It follows an internal production grammar, just like the language of native speakers does (this is called interlanguage, following Selinker 1972). This internal grammar is, of course, different from native speaker grammar. The corpus data is an expression of the internal grammar of the learners and can be used to discover the internal grammar of the learners. One source of evidence that points to the interlanguage is the kinds of errors that learners make – I will say more on this in Section 5.2.
One other source of evidence is quantitative. Everybody knows the phenomenon: Learner varieties – such as, for example, scientific papers written in English by non-English speakers like myself – are often recognizable as non-native even if they do not contain any outright errors – they are too 'formal' in some situations or too 'informal' or just 'strange'. This perception is based on quantitative differences between learner varieties and native varieties; the learners use certain words, phrases, sentence types or other constructions not in the same

[6] Falko is freely available at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en. It contains several subcorpora of learner data and comparable native speaker data in two genres. Here I only use the essay data (122778 tokens L2 and 68491 tokens L1, version 2.0; 'L1' is used for first (native) language and 'L2' is used for second/foreign language). The learners represented in the Falko corpus are advanced adult learners (university students) who learned German in a classroom setting.
[7] When I speak about learner language I refer to foreign (tutored/classroom setting) and second (non-tutored) language learners (the differences are not relevant for the purpose of this paper, but see e.g. Dietrich 2004 or Kroll & Sunderman 2003). There are, of course, also numerous corpora and corpus-based studies of first language learners (see e.g. MacWinney 1996, Behrens 2008, Diessel 2009).

o

distribution that native speakers would in that situation: Our learner English contains not enough variety in sentence types or perhaps too many passive sentences or too few adverbs etc. This perception implies an implicit quantitative comparison – it is not ungrammatical at any point in the article to use a passive sentence – native speakers would just not do it as often. There are many quantitative methods to compare learner language and native language (or any other two corpora, see Section 4). In this paper I want to concentrate on just one method, namely overuse/underuse statistics. The idea behind this is that learners of a language tend to (consciously or unconsciously) avoid certain constructions which, for some reason, are difficult for them – here we find an *underuse* when we compare the learner corpus to the native corpus – and compensate by using alternative constructions too often – here we find an *overuse*. Overuse and underuse can show different things: It might be the case that the learner is not aware of the 'native' distribution of the construction under question because he or she has not had enough exposure to the variety. It might also be the case that a certain construction is underused because it is 'difficult'.

Table 1 shows how underuse and overuse can be visualized.[8] The table shows underuse and overuse of the most frequent words in the corpus. The German data (deu) is used as a reference. Falko contains data from learners with 48 different native languages (L1s) but sometimes the number of texts written by learners of a given language is too small to produce significant results. Here, we therefore only look at the five largest learner groups (with L1s Danish, English, French, Russian, and Uzbek). The colours are an easy way to see whether a given item is underused or overused. Cold colours signal underuse, warm colours signal overuse. The intensity of the colours signals the strength of the over- or underuse[9]. We see, for example, that the definite article *die* is overused by all learners and the conjunction *und* is underused by all learners. The other words do not behave uniformly across the learner groups.

word	deu	dan	eng	fra	rus	uzb
die (form of the definite article	0,028296	0,03748	0,037161	0,040397	0,03881	0,035178
und "and"	0,023288	0,022804	0,022409	0,019756	0,020837	0,020638
der (form of the definite article)	0,019667	0,019585	0,021245	0,022803	0,024495	0,023218
es "it"	0,013827	0,010638	0,013025	0,011991	0,01352	0,004925
nicht (negation)	0,013695	0,013857	0,014414	0,012679	0,012725	0,015478
zu "to"	0,013608	0,012166	0,015352	0,015038	0,013679	0,009146
ist "is"	0,012615	0,011075	0,014189	0,014645	0,015588	0,011726
in "in"	0,012308	0,014239	0,014489	0,01553	0,015429	0,011257

Table 1: Visualization of overuse and underuse for the most frequent words in the Falko corpus. The reference numbers are in the native German (deu) column; the other languages are Danish (dan), English (eng), French (fra), Russian (rus), and Uzbek (uzb). Overuse is signalled by warm colours, underuse by cold colours.

After this short introduction of the corpus and the overuse/underuse method I will now come to the first general issue: sampling. Some of the general issues are then illustrated using the Falko data in Section 4.2.

[8] Overuse and underuse can be computed for any two corpora. The visualization Add In for Excel was written by Amir Zeldes and is freely available at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/-mitarbeiter-innen/amir/. Lüdeling/Hirschmann/Zeldes (under review) show how the same method can be used in detecting language change.

[9] All corpus counts are normalized. All cells that are coloured in this and the following tables show a statistically significant difference to the German data.

o

4. Sampling – corpus design

Ideally, the research question determines the data to be used, so that all variables can be controlled. A corpus is almost always a sample of the linguistic variety to be analyzed. Only in very few cases – when the entire population is finite, small and accessible - is it possible to work without sampling. If, for example, one wanted to analyze all of Nobel's letters to his mistress it could, in principle, be possible to take all of them (if they were available). But usually the linguistic variety to be analyzed ('Modern English', 'Modern British English', 'the English of boys between 10 and 15 that live in Manchester', …) is too large to be completely analyzed and one has to take a sample. Sampling, of course, depends on many parameters. One has to decide on a sampling algorithm, size etc. All of this plays a role for the possibilities to generalize from observed data (the corpus) to unseen data (the variety in question). Here I can only speak about variation as one of the issues involved, but see e.g. Biber (1993) or Hunston (2008) for further thoughts on sampling.

4.1 Variation

A speaker can express the same thought in many different ways. He can use active or passive voice, different lexical items, be more or less specific etc. It has long been recognized that external factors (such as modality, age/sex/education of producer, text type and probably many more that we have not yet identified) influence language production on all linguistic levels. There are also language-internal factors such as information structure or previous discourse. While some differences seem obvious (but might really be due to faulty observation, cf. Mair 2009) many such differences can only be observed and really tested by using corpus material (often together with elicited material[10]). The idea behind all of these studies is that a speaker has several options (variants) to realize a variable. While there might be cases of genuine free variation, many areas of variation can be explained (or at least one can find co-varying variables), and these are the areas we want to detect (see for example Labov 2001, 2008).
Although there might be categorical differences between varieties (feature A appears in texts of type Y but never in texts of type Z), most of the differences are quantitative (feature A is more frequent in texts of type Y than in texts of type Z; see e.g. Biber 2009). There are, of course, many different ways to design experiments that test variation. Biber & Jones (2009) distinguish between two types of studies: Type A studies test whether and how the distributions of variants of the same variable in the same corpus might be explained. Type B studies test differences between two corpora. There are simple descriptive methods but these are limited in scope. „Statistical inference is necessary because any sample from a language is subject to random variation" (Baroni & Evert 2009, 778). Statistical modelling and testing is necessary, especially because corpus studies have shown that the distributions of different categories differ widely and thus that not each statistical model can be applied to each type of

[10] In many cases it is necessary to complement corpus data with other data (elicited data, data from psycholinguistic or sociolinguistic experiments etc.). Corpus linguistics as an empirical method uses many of the same methods used in sociolinguistics and psycholinguistics (see e.g. Nevalainen & Raumolin-Brunberg 2003 for historical sociolinguistics or Romaine 2008 for an overview of corpora in sociolinguistics, or Gilquin & Gries 2009 for an overview of the combination of corpus methods and experimental methods).

o

corpus data.[11] But how much variation is there and is it really important to study variation? Some people argue that variation is not interesting because some 'core' or 'standard' grammar always applies and it is only important to find this core grammar. However, without understanding variation it is not possible to understand language acquisition, language use and language change. The fact that speakers predictably produce different varieties shows that this awareness is part of the general knowledge of language. The differences are subtle but can reliably be detected. This has been shown in numerous studies, for example in the register studies by Douglas Biber and colleagues (for an overview cf. Biber 2009). They use a multidimensional method where they annotate the different corpora on many linguistic levels. This creates a multidimensional space which is very difficult to interpret. The dimensions are then reduced by calculating co-occurrences[12] and then interpreted functionally. Using this method, Biber is able to find linguistic differences within scientific articles – introductions can be reliable distinguished from the rest of the article, for example. There are many similar results[13]: people vary their phonology (Labov 1996, Medoza Denton 2008), their word formation behaviour (Plag/Dalton Puffer/Baayen 1999), syntactic expressions (Manning 2003, Ford & Bresnan 2010), etc. Fields that research variation include dialectology, sociolinguistics, pragmatics, historical linguistics, language acquisition, and many more. All of these use corpus data in some way or other.

4.2 Variation in Falko

In the following I want to use the overuse/underuse statistics introduced in Section 3 to show which effects variation can have in our learner corpus. Consider the Uzbek row in Table 2: all three cells show underuse. If there were no other corpora one could conclude that Uzbek speakers have problems with respect to the complementizer *dass* 'that' in as well as with respect to the reflexive (and indeed there are many studies that use exactly this one-to-one comparison, but see Granger/Hung/Petch-Tyson 2002). However, when we look at the cells for the other languages we see that the situation is completely different for the two cases. The complementizer is overused in all other subcorpora, the reflexive is underused in all other corpora as well. For the complementizer we could assume that the L1 of the learners influences their acquisition behaviour.[14] For the reflexive we could assume that there is a problem in target language that is independent of the L1 of the learner (as indeed the different L1s of the learners use the reflexive in completely different ways). The underuse of *Frauen* 'women' has yet a different reason. The learners were free to choose between four topics for

[11] Just briefly: Words are distributed in a Zipfian way which means that some words are very frequent but many words are rare. In essence this shows even large corpora have not yet come close to sample all the words in a language (this is due to morphological productivity, borrowing, creativity, and other factors, which lead to the fact that the vocabulary of a language is, for all intents and purposes, infinite). For Zipfian (or LNRE) distributions only certain statistical models can be used (see Baayen 2001). Closed class categories such as part-of-speech categories often show a normal distribution and can be analyzed using 'regular' statistical tests (Baroni 2008). It is necessary to know which kind of distribution one is looking at. For more on these issues see also Evert (2006) and Kilgarriff (2005).

[12] There are several ways of doing this, for example a Principal Components Analysis or a Factor Analysis (see e.g. Baayen 2008).

[13] There is so much work in each of these areas that I can only point to a few examples.

[14] This is called transfer (after Selinker 1969). There are many studies that show that transfer happens on all linguistic levels, can be either positive or negative, and is much more complex and difficult than initially thought (Ellis 2008).

o

their essays, one of them about feminism, and the Uzbekian learners happened to like the other topics better and simply didn't write about women.

word	deu	dan	eng	fra	rus	uzb
sich (reflexive)	0,012294	0,005892	0,005255	0,006389	0,004613	0,00469
dass "that"	0,007709	0,012602	0,009797	0,008748	0,011293	0,004221
Frauen "women"	0,003051	0,006438	0,008408	0,005111	0,006362	0,001173

Table 2: Overuse/underuse of *sich, dass* and *Frauen* for five Falko subcorpora.

The overuse/underuse tables can be used as a diagnostic to find interesting cases. The real linguistic analysis still has to follow.

While it is not surprising that the native language of a learner influences the interlanguage it might come as a surprise that the gender of a learner influences his/her linguistic behaviour. Table 3 shows the differences between native (L1) men and women as well as between L2 men and women. Because we have more essays written by women we chose the women as the 'standard' (this is, in essence, arbitrary). Again, there are clear differences in content words which, of course, reflect the topic and might be gender specific (unsurprisingly again *Frauen* 'women' is underused by all men), but more interesting are the grammatical words like prepositions etc. that are fairly independent of the topic matter. Again, two areas of comparison are interesting. First, one could look for those cases where all men differ from all women (like *die, das* 'the' or *Frauen*). Then one could look at those cases where learners differ from native speakers (most of the other cases in the table).

word	L1_f	L1_m	L2_f	L2_m
die (form of definite article)	0,02847173	0,02773356	0,03873803	0,03793173
ist "is"	0,01309508	0,01108118	0,01296901	0,01558681
sich (reflexive)	0,01263493	0,01120362	0,00552113	0,00629799
das (form of definite article)	0,01090937	0,01138729	0,00895775	0,00963391
für "for"	0,00718983	0,00887719	0,00704225	0,00701694
Frauen "women"	0,0036045	0,00128566	0,00802254	0,00382481
wir "we"	0,00105451	0,00140811	0,00376338	0,00316337
um "in order to"	0,00375788	0,00177544	0,00307606	0,00350847
was "which, what"	0,00370037	0,00281621	0,00345915	0,00370977
aus "out (of)"	0,00325939	0,00367332	0,00161127	0,00135162
nach "after"	0,00262668	0,00330599	0,00305352	0,00238691

Table 3: Overuse/underuse comparison of some frequent words between men and women in Falko. L1_f: native German speakers female, L1_m: native German speakers male, L2_f: learners female, L2_m: learners male.

Again, it would be necessary to investigate further. What I wanted to show is that gender – just like native language – is a relevant parameter for this kind of corpus, even for seemingly content independent grammatical items. The Falko L2 corpus has more than 2/3 female authors. So whatever we say about 'the learners' might be just a fact about female learners.

4.3 Metadata and Subcorpora

This poses a question: If so much variation exists, what is the status of a given corpus datum? The answer depends, of course, on the research question one wants to answer. But very much

o

research in linguistics does not take variation into account as much as it should – there are still many opportunistic corpora or corpora unsuitable for a given research question. Often we see analyses and models without having access to the data. But how do we then know whether a given analysis is valid? If it seems plausible we assume that it is if it doesn't seem plausible we assume that it isn't. Those cases are in essence like the situation in Hermann Paul's time.
Two points become very clear: (1) the data should always be provided with any analysis[15] and (2) the use of (standardized) metadata is important so that subcorpora can be constructed at any point in the analysis. Both issues rely on electronic corpora, common coding standards and strategies for data preservation (Romary, this volume) and powerful search tools.

5. Annotation – interpretation of the data.

Section 4 showed that the sample we chose for a linguistic study crucially influences the results. This section is concerned with further interpretation of the sample. It is impossible to use data without interpreting it in some way. Even deciding which surface forms are to be analyzed is a way of interpretation. Interpretation is category building or abstraction and abstraction entails a necessary loss of information: One does not want to look at individual cases but at classes of cases.

5.1 Necessary Loss of Information

In corpus studies, interpretation happens at two levels – at the level of the primary data, and at the level of annotating the primary data. First, one has to decide on a reading or an analysis of the primary data which can be ambiguous syntactically, semantically etc. In many cases linguists concur in the interpretation of a datum and do not notice the ambiguity.[16] This might have implications for an analysis. The issue is even more interesting for data that is not covered by established grammatical analyses (learner data, historical data, spoken data etc., see Section 5.2).
Here I want to focus on the second level of interpretation, namely annotation. In Hermann Paul's time – and very often today – a linguist used categories (e.g. feminine nouns with linker *s*) for his or her analysis without assigning them directly to the data – the researcher works in his own spreadsheet, file-card system or whatever. If it were completely clear which categories there are and how these were to be assigned, this would be okay. But linguists almost never agree on a category. There are scores of articles on how to assign even the most basic part-of-speech categories (just try to decide on a definition of 'adverb'). This is where electronic corpora are crucial: In multi-layer corpus architectures (see Section 5.3) it is possible to assign any item to a pre-defined category so that the categories can be searched just like the primary data. This could be part-of-speech categories, syntactic information, co-reference information, semantic information, information about rhetorical structure, narrative structure or whatever else one wants to study. There are two steps involved: First one needs to decide on the relevant categories (the set of relevant categories is often called tagset) and then one needs to decide on the guidelines for assigning each category. All the decisions involved in the assignment can be made explicit so that a user of the corpus knows exactly what is in

[15] There are well-known legal and (less well-known) ethical issues involved sometimes. See XXX in this collection (XXX Suber? XXX Rausing?).
[16] See Wasow/Perfors/Beaver (2005) on different aspects on ambiguity in natural language.

o

the corpus. And if she disagrees with the categories or the guidelines she can open up a new annotation layer and provide her own interpretation of the data.

There has been a lot of discussion about the status of annotation. Many corpus linguists want to work in a corpus-driven paradigm (following Sinclair 1991) and some argue against annotation. But *every* qualitative analysis that goes beyond a single instance and every quantitative endeavour relies on some kind of categorization (be it only word forms) and some kind of linguistically motivated question. In 1991 multi-layer annotation was not possible and very often the only annotation layer was perceived as some kind of 'true' interpretation of the data. I believe that we can take Sinclair's arguments against such annotation seriously but rather than *not* annotating I think we need to have possibly conflicting independent layers of annotation and thereby make our assumptions visible.[17]

I want to exemplify why this is important again using the Falko data. I will come back to multi-layer corpus architectures in Section 5.3.

5.2 Falko: Target Hypotheses

In Section 3 I said that one possible way of studying learner's interlanguage is error analysis: Errors in the learner output are identified, classified and analyzed. From the errors one can hypothesize what kind of a 'rule' or 'regularity' the learner might have about the target language. This sounds straightforward but is extremely difficult to implement because it is unclear what an error might be.[18] I want to illustrate this with an example. The learner utterance in (2) contains orthographic, word formation, case and number errors – all of these are at first glance uncontroversial.

(2) Die politiker die in Korruption aktivitäten sind wird im Gefängnis gehen. (kne21_2006_09)
„ ≈ The politicians that are involved in corrupt activities will go to prison."

But even with the seemingly uncontroversial example we have to note that an error analysis is not possible without at least implicitly assuming a correct version of the sentence. And if we look more closely there are different possible 'correct' versions (which we call target hypotheses). This is illustrated in Table 4. Target hypothesis 1 is possible and very close to the learner utterance but target hypothesis 2 is probably a more idiomatic German sentence. It becomes immediately clear that the target hypothesis determines which errors are found. There are a number of errors common to both target hypotheses (capitalization of *politiker* 'politician', commas around the relative clause, compounding of *Korruptionsaktivitäten* 'corrupt activities') but there are also errors that concern only one of the target hypotheses (addition of verb in the relative clause and using a more idiomatic verb in the main clause for target hypothesis 2). These problems arise in almost every sentence and can have a potentially huge effect on the analysis (see Lüdeling 2008 for an experiment on the effects of target hypotheses).

[17] Although it is possible to have several annotation layers that code the same linguistic level (say: part-of speech) I am aware of only very few projects that have tried to do this. One prominent example is the AMALGAM project that compared part-of-speech tags and grammatical coding schemes for English (http://www.comp.leeds.ac.uk/amalgam/amalgam/amalghome.htm).

[18] The status of learner errors in language learning and teaching has been extensively and controversially discussed (among many others see Corder 1981, Ellis 2008). Here I cannot recapitulate this discussion but want to limit myself to problems of error detection and annotation.

o

learner utterance	gloss	target hypothesis 1	gloss of th1	errors th1	target hypothesis 2	gloss of th2	errors th2
Die	the	Die	the		Die	the	
politiker	politicians	Politiker	politicians	orth	Politiker	politicians	orth
		,	,	orth	,	,	orth
die	that	die	that		die	that	
in	in	in	in		in	in	
Korruption aktivitäten	corruption activities	Korruptions-aktivitäten	corrupt activities	word formation	Korruptions-aktivitäten	corrupt activities	word formation
					verwickelt	involved	lexicon
sind	are	sind	are		sind	are	
		,	,	orth	,	,	orth
wird	will (sing.)	werden	will	number	werden	will	number
im	in+the (dative)	ins	into+the (accusative)	case	im	in+the (dative)	
Gefängnis	prison	Gefängnis	prison		Gefängnis	prison	
gehen	go	gehen	go		landen	end	lexicon

Table 4: A learner utterance and two different target hypotheses. The common errors are marked in blue, the errors that only refer to one of the hypotheses are marked in red.

If the target hypothesis is so important for all further analysis and if there are potentially always several possible target hypotheses it becomes clear that the target hypothesis should be coded as an annotation layer in the corpus. There can be no single "correct" version; every target hypothesis is an interpretation of the data. (This is true for every error analysis, independent of whether the target hypothesis is explicitly coded in the data or not.) Only if the hypotheses are annotated *in* the corpus one always knows what the basis for a given analysis is (and one can choose to define a different target hypothesis if necessary).

The Falko corpus is annotated with two different target hypotheses (the decisions behind them are described in Reznicek et al. 2010). The first one is minimal and only corrects grammatical errors even if the sentence does not make sense or does not fit into the narrative. The second target hypothesis looks at the sentence within the text and also corrects unidiomatic expressions or anaphoric errors.[19] This means that Falko can be used as an example in which we can show what effects different interpretations can have on the analysis. I want to exemplify this with an experimental question where we can use the same overuse/underuse method that we used before. Only now we do not calculate overuse or underuse on the primary data but on the annotations.

Above we saw that learner language differs quantitatively from native language. One reason is the errors: If, for example, many learners use incorrect prepositions (as in Example 1), there could be an underuse of more difficult or rarer prepositions. Since both target hypotheses are 'correct' German, this effect should be gone. But the target hypotheses are still close to the learner utterances. We could use the annotated data to see how close the target hypotheses are to the native speaker data. In other words: Are the target hypotheses more similar to the L1 corpus than the learner data? How strongly does the learner data 'shine through'[20]? We could, of course, look at each word in the corpus separately. But for questions like these it is more interesting to look at classes of words. Table 5 shows a comparison of different inflectional categories of main verbs. Main verbs constitute an open class, which means that it is not easily possible to look at all items separately. We can therefore use the part-of-speech annotation. The 'standard' against which everything is compared is the German (deu) column. The 'paler' the other cells are the more similar to the original they are. If target hypothesis 2 is

[19] The learner text and both target hypotheses are the automatically annotated with part-of-speech categories (using the TreeTagger, Schmid 1994). In addition there is an automatic error coding.

[20] The term ‚shining through' is used in translation studies, cf. Teich (2003).

o

closer to the L1 data than target hypothesis 1 and target hypothesis 1 is closer to the L1 data than the original data Table 5c should be paler that Table 5b and Table 5b should be paler than Table 5a.

The expectation is borne out, except for the Uzbek data. Looking at the Danish, English, French and Russian corpora we see that all main verb categories are underused. The underuse is less strong in the target hypotheses. The Uzbek corpus differs markedly. We would now have to analyze where these differences come from.

pos	deu	dan	eng	fra	rus	uzb
VVFIN	0,045553	0,044081	0,046582	0,03735	0,048513	0,054409
VVINF	0,028383	0,025259	0,028415	0,029487	0,03038	0,033537
VVIZU	0,001898	0,000927	0,001051	0,001376	0,000954	0,000469
VVPP	0,020543	0,01713	0,015915	0,015333	0,013361	0,008443

Table 5a: Overuse/underuse table of main verbs in Falko, original learner utterances.

pos	deu	dan	eng	fra	rus	uzb
VVFIN	0,045553	0,045485	0,042596	0,0354	0,042345	0,026316
VVINF	0,028383	0,024555	0,029251	0,029326	0,025733	0,031579
VVIZU	0,001898	0,001154	0,001379	0,001466	0,001954	0,002632
VVPP	0,020543	0,017029	0,016398	0,017386	0,014984	0,015789

Table 5b: Overuse/underuse table of main verbs in Falko, target hypothesis 1.

pos	deu	dan	eng	fra	rus	uzb
VVFIN	0,045553	0,045829	0,042658	0,03617	0,043803	0,031008
VVINF	0,028383	0,024453	0,029999	0,03042	0,026282	0,036176
VVIZU	0,001898	0,001319	0,001527	0,001464	0,001622	0,005168
VVPP	0,020543	0,016266	0,016452	0,018085	0,016223	0,015504

Table 5c: Overuse/underuse table of main verbs in Falko, target hypothesis 2.
pos: part-of-speech category[21], VVFIN: finite main verbs, VVINF: infinite main verbs, VVIZU infinite main verbs with particle *zu*, VVPP: participles

This study shows that different interpretations (here the different target hypotheses) lead to different results and that annotations strongly depend on the interpretation. Since it is not possible to choose the one 'correct' interpretation the only way out is to provide as much transparency as possible. This can only be done through annotation.

5.3 Multi-layer Annotation

I argued that one of the crucial improvements electronic corpora made for empirical work in linguistics is transparency. In this section I want to say just a few words about the technical background – multi-layer corpus architectures and standardized exchange formats – that make this possible.

[21] Part-of-speech tagging is usually done automatically. Most taggers use a mixture of lexicon-based and statistical techniques to assign a part-of-speech category to each word in the corpus. A tagset determines which part-of-speech tags can be used. Here we used the Stuttgart-Tübingen tagset which can be found at http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html. As you can see this tagset combines genuine part-of-speech information (VV – verb) with morphological information (FIN – finite, INF – infinite etc.).

o

As mentioned above, many linguists do their analysis away from the corpus in separate files or spreadsheets instead of directly in the corpus. There are several reasons for this. First, many linguists are not aware of the usefulness of direct annotation. This will hopefully change with time. Second, many corpora are not available in a format which can be annotated by users, but only via e.g. a Web interface with limited context. This is a hindrance for research, many people have realized that and the situation is changing (again: provided that the legal and ethical issues can be solved).[22] The third obstacle used to be technical – for a long time it was technically not possible to annotate the same primary data with different kinds of annotation, such as token-based annotation, syntactic annotation, pointing relations etc. But in the past ten years or so multi-layer annotation has developed[23] and today there are free and accessible tools to annotate a file in many different ways. Many of these tools output a standardized and well-described format (usually in XML) and in recent years several integrative frameworks have been developed (Romary, this volume). Using these it is possible for researchers that are far from each other and have different theoretical notions and different research questions to annotate the same data. The annotation layers can be integrated and each annotation layer can be searched separately or annotation layers can be combined for the search.

6. Summary: Transparency and Availability

Corpus data is one of the empirical bases for linguistic research. It has been used long before electronic times and many of the research questions and methods were there before computers came along. In this paper I illustrated how the possibilities for systematic sampling and transparent interpretation that electronic corpora provide have changed the understanding of linguistic work and empirical methods:

Corpora are always samples of a given variety. Through more systematic search and analysis tools – especially for quantitative analysis – we are beginning to understand how much variation there is between different varieties of a language and how many external factors influence language use. Careful sampling techniques and especially the use of metadata allows us now to see exactly which factors are relevant.

Every linguistic analysis is an interpretation of the data. In electronic corpora it is possible to explicitly store the interpretation with the primary data so that it is visible to the user. If the data is stored in an appropriate format many scholars can work in parallel on the same data.

Corpora are just data. The research questions must be formulated within a given theory or model using linguistic knowledge. The results of the analysis must be integrated into the theory or model. In between question and results we need to choose the appropriate empirical bases and suitable research methods. Corpora are one type of data with possibilities and limitations and many interesting qualitative and quantitative methods for corpus analysis have been developed. We have learned a lot about language from using corpora and there is still so much more that we do not yet understand. We also learned a lot about the empirical work and

[22] Many linguists are organized in large infrastructure projects like CLARIN (http://www.clarin.eu/external/) and DARIAH (http://www.dariah.eu/) which promote common coding standards and accessibility. A well-known and widely used standard is formulated by the Text Encoding Initiative (http://www.tei-c.org/index.xml).

[23] Multi-layer annotation was first developed for multimodal corpora which comprise audio data, video data, and text (see Wittenburg 2008 for an overview). There are several models for standoff multi-layer corpora today (Carletta et al. 2003, Chiarcos et al. 2009) and suggestions for converting annotated data out of and into several formats (see e.g. Zipser & Romary 2010). The Falko corpus is stored in such a format and can be searched using the freely available search tool Annis2 (Zeldes et al. 2009).

o

– again – there is so much more that we do not yet understand. But electronic corpora enable us to make every step of interpretation and analysis transparent and reproducible.

References

Baayen, R. Harald (2001) *Word Frequency Distributions*. Kluwer, Dordrecht.

Baayen, R. Harald (2008) *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge.

Baroni, Marco (2008) Distributions in Text. In: Lüdeling, Anke & Kytö, Merja (eds), 803-822.

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano & Zanchetta, Eros (2009) The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora. In: *Journal of Language Resources and Evaluation 43 (3)*, 209-226.

Baroni, Marco & Evert, Stefan (2009) Statistical Methods for Corpus Exploitation. In: Lüdeling, Anke & Kytö, Merja (eds), 777-803.

Behrens, Heike (ed) (2008) *Corpora in Language Acquisition Research. History, Methods. Perspectives*. John Benjamins, Amsterdam.

Biber, Douglas (1993) Representativeness in Corpus Design. In: *Literary and Linguistic Computing 8*, 243-257.

Biber; Douglas (2009) Multi-dimensional approaches. In: Lüdeling, Anke & Kytö, Merja (eds), 822- 855.

Biber, Douglas & Jones, James K. (2009) Quantitative Methods in Corpus Linguistics. In: Lüdeling, Anke & Kytö, Merja (eds), 1286-1304.

Busa, Roberto (1974) *Index Thomisticus*. Stuttgart: Frommann-Holzboog.

Busa, Roberto (1980) The Annals of Humanities Computing: The Index Thomisticus. In: *Computers and the Humanities 14*, 83–90.

Carletta, Jean; Kilgour, Jonathan; O'Donnell, Tim; Evert, Stefan & Voormann, Holger (2003) The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets. In: *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003).*

Chiarcos, Christian; Dipper, Stefanie; Götze, Michael; Leser, Ulf; Lüdeling, Anke; Ritz, Julia & Stede, Manfred (2009) A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. In: *Traitement Automatique des Langues* 49(2), 271-291.

Corpus Linguistics and Linguistic Theory. Special Issue on 'Grammar without Grammaticality' 3(1), 2007.

Corder. Stephen Pit (1981) *Error Analysis and Interlanguage*. Oxford, Oxford University Press.

Dietrich, Rainer (2004) Zweitsprache – Fremdsprache. In: Ammon, Ulrich; Dittmar, Norbert; Mattheier, Klaus & Trudgill, Peter (eds) *Soziolinguistik. Ein Internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*. 2nd edition. Walter de Gruyter, Berlin, 311-313.

Diessel, Holger (2009) Corpus Linguistics and Language Acquisition. In: Lüdeling, Anke & Kytö, Merja (eds), 1197-1212.

Ellis, Rod (2008) *The Study of Second Language Acquisition*. Oxford, Oxford University Press.

Evert, Stefan (2006) How Random is a Corpus? The Library Metaphor. In: *Zeitschrift für Anglistik und Amerikanistik 54(2)*, 177-190.

o

Fillmore, Charles (1992) 'Corpus linguistics' vs. 'Computer-aided Armchair Linguistics'. In: *Directions in Corpus Linguistics*. Proceedings from Nobel Symposium 82, 4-8 August, 1991. Mouton de Gruyter, Berlin, 35-60.

Ford, Marilyn & Bresnan, Joan (2010) Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. In: *Language* 86(1), 186-213.

Gilquin, Gaëtanelle & Gries, Stefan Th. (2009) Corpora and experimental methods: A state-of-the-art review. In: *Corpus Linguistics and Linguistic Theory 5(1)*, 1-26.

Granger Sylviane (2008) Learner Corpora, In: Lüdeling, Anke & Kytö, Merja (eds), 259-275.

Granger, Sylviane, Hung, Joseph & Petch-Tyson, Stephanie (eds) (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. John Benjamins, Amsterdam.

Heid, Ulrich (2008) Corpus Linguistics and Lexicography. In: Lüdeling, Anke & Kytö, Merja (eds), 131-153.

Hunston, Susan (2008) Collection strategies and design decisions. In: Lüdeling, Anke & Kytö, Merja (eds), 154-168.

Käding, Friedrich Wilhelm (1897) *Häufigkeitswörterbuch der deutschen Sprache*. Privately published, Berlin.

Karlsson, Fred (2008) Early generative linguistics and empirical methodology. In: Lüdeling, Anke & Kytö, Merja (eds) (2008), 14-32.

Kepser, Stefan & Reis, Marga (2005) *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives.* Mouton de Gruyter, Berlin.

Kilgarriff, Adam (2005) Language is never ever ever random. *Corpus Linguistics and Linguistic Theory* 1 (2): 263-276.

Köhler, Reinhard (2005) Gegenstand und Arbeitsweise der quantitativen Linguistik. In: Köhler, Reinhard/Altmann, Gabriel & Piotrowski, Rajmund G. (eds) *Quantitative Linguistics / Quantitative Linguistik. An International handbook / Ein internationales Handbuch.* Berlin, Mouton de Gruyter, 1-16.

Kroll, Judith F. & Sunderman, Gretchen (2003) Cognitive Processes in Second Language Learners and Bilinguals: The Development of Lexical and Conceptual Representations. In: Doughty, Catherine J. & Long, Michael H. (eds) *The Handbook of Second Language Acquisition.* Blackwell, Oxford, 104-129.

Labov, William (1966) *The Social Stratification of English in New York City*. The Center for Applied Linguistics, Washington (2nd edition 2006, Cambridge University Press, Cambridge).

Labov, William (2001) *Principles of Linguistic Change. Social Factors*. Blackwell, Oxford.

Labov, William (2008) Quantitative Analysis of Linguistic Variation. In: Ammon, Ulrich; Dittmar, Norbert; Mattheier, Klaus & Trudgill, Peter (eds) *Sociolinguistics/-Soziolinguisik.* Vol 1. de Gruyter, Berlin, 6-21.

Lüdeling, Anke (2008) Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Maik Walter & Patrick Grommes (Hrsg.) *Fortgeschrittene Lernervarietäten*, Niemeyer, Tübingen, 119-140.

Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin & Walter, Maik (2008) Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache 2(2008)*, 67-73.

Lüdeling, Anke; Hirschmann, Hagen & Zeldes, Anke (under review) Variationism and Underuse Statistics in the Analysis of the Development of Relative Clauses in German. In: Kawaguchi, Yuji, Minegishi, Makoto & Viereck, Wolfgang (eds) *Corpus Analysis and Diachronic Linguistics*. John Benjamins, Amsterdam/Philadelphia.

Lüdeling, Anke & Kytö, Merja (eds) (2008) *Corpus Linguistics. An International Handbook.* Vol 1. Mouton de Gruyter, Berlin.

o

Lüdeling, Anke & Kytö, Merja (eds) (2009) *Corpus Linguistics. An International Handbook*. Vol 2. Mouton de Gruyter, Berlin.

MacWhinney, Brian (1996). The CHILDES system. In: *American Journal of Speech-Language Pathology 5*, 5-14.

Mair, Christian (2009) Corpora and the Study of Recent Change in Language. In: Lüdeling, Anke & Kytö, Merja (eds), 1109-1125.

Manning, Christopher D. (2003) Probabilistic Syntax. In: Bod, Rens; Hay, Jennifer & Jannedy, Stefanie (eds) *Probabilistic Linguistics*. MIT Press, Cambridge MA, 289-341.

Mendoza Denton, Norma (2008) *Homegirls: Language and Cultural Practice among Latina Youth Gangs*. Wiley/Blackwell, London.

Meyer, Charles (2008) Pre-electronic Corpora. In: Lüdeling, Anke & Kytö, Merja (eds), 1-14.

Moisl, Hermann (2009) Exploratory Multivariate Analysis. In Lüdeling, Anke & Kytö, Merja (eds), 874-899.

Nevalainen, Terttu & Raumolin-Brunberg, Helena (2003) *Historical Sociolinguistics. Language Change in Tudor and Stuart England*. Pearson Education, London.

Paul, Hermann (1959 [1920]) *Deutsche Grammatik*. Niemeyer, Halle.

Plag, Ingo; Dalton-Puffer, Christiane & Baayen, R. Harald (1999) Morphological productivity across speech and writing. In: *English Language and Linguistics 3*, 209-228.

Pomikalek, Jan; Rychly, Pavel & Kilgarriff, Adam (2009) Scaling to Billion-plus Word Corpora. In: *Advances in Computational Linguistics. Special Issue of Research in Computing Science Vol 41*, Mexico City.

Reznicek, Mark; Walter, Maik; Schmid, Karin; Lüdeling, Anke; Hirschmann, Hagen; Grommes, Cedric (2010) Das Falko-Handbuch. Korpusaufbau und Annotationen. Online at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/standardseite

Romaine, Suzanne (2008) Corpus Linguistics and Sociolinguistics. In: Lüdeling, Anke & Kytö, Merja (eds) (2008), 96-111.

Romary, Laurent (this volume) Stabilizing Knowledge through Standards - a Perspective for the Humanities. XXX

Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing. September 1994.* Online at http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

Selinker, Larry (1969) Language Transfer. In: *General Linguistics 9(2)*, 67-92.

Selinker, Larry (1972) Interlanguage. In: *IRAL 10/1972*, 209-231 (reprinted in: Richards, J. C. (ed.) *Error Analysis. Perspectives on Second Language Acquisition.* London: Longman, 31-54).

Sinclair, John M. (1991) *Corpus Concordance Collocation*. Oxford University Press, Oxford.

Teich, Elke (2003) *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. De Gruyter, Berlin/New York.

Wasow, Thomas; Perfors, Amy & Beaver**,** David (2005) The Puzzle of Ambiguity. In Orgun, C. Orhan & Sells, Peter (eds) *Morphology and The Web of Grammar: Essays in Memory of Steven G. Lapointe*. CSLI Publications, Stanford.

Wittenburg, Peter (2009) Preprocessing Multimodal Corpora. In: Lüdeling, Anke & Kytö, Merja (eds), 664-685.

Zeldes, Amir; Ritz, Julia; Lüdeling, Anke & Chiarcos, Christian (2009) ANNIS: A Search Tool for Multi-Layer Annotated Corpora. In: *Proceedings of Corpus Linguistics 2009*, July 20-23, Liverpool, UK.

Kommentar [l1]: Pages to be added

This paper will appear in: Grandin, Karl (eds.) *Going Digital. Evolutionary and Revolutionary Aspects of Digitization.* [Nobel Symposium 147]. Science History Publications/USA, New York. [This is the last version I have copyright for.]

o

Zipser, Florian & Romary, Laurent (2010) A Model-Oriented Approach to the Mapping of Annotation Formats using Standards. In: *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010. Malta*, 7-18. Online at http://www.lrec-conf.org/proceedings/lrec2010/workshops/W4.pdf.