VIENNA ENGLISH WORKING PAPERS

VOLUME 17 NUMBER 2 DECEMBER, 2008

INTERNET EDITION AVAILABLE ON:

HTTP://WWW.UNIVIE.AC.AT/ANGLISTIK/ANG_NEW/ONLINE_PAPERS/VIEWS.HTML

CONTENTS

LETTER FROM THE EDITORS

Dear Readers,

After the frequently hectic and stressful Christmas season, we would like to invite you all to relax with an inspiring new issue of VIEWZ – an issue packed with material (quite literally so, as it is concerned with corpus studies) and, indeed, views (on corpus building and uses of corpora, corpus studies and corpora in general).

Ursula Lutzky's contribution provides an example of the use (and indeed the creation!) of corpora in diachronic linguistics, analysing the use of the

pragmatic marker *marry* as it presents itself in selected plays of the Early Modern English period.

The established team of Marie-Luise Pitzl, Angelika Breiteneder and Theresa Kimpflinger offer not only a further view on the VOICE corpus but also present both synchronic and diachronic views, as they address matters of language variation and word-formation in English as a lingua franca in a paper that exemplifies how today's synchronic studies are tomorrow's historical linguistics.

Last but certainly not least, Barbara Schiftner provides the synchronic side to corpus studies, giving an in-depth analysis on the status quo and the future of learner corpora of both English and German as a foreign language and supplying the reader with a description of existing corpora and their availability.

We hope you will enjoy the stimulating contributions of this year's winter issue and would be happy to include your comments in form of a reply in the next issue. We wish you all the best for 2009 - may it be a successful and happy year!

THE EDITORS

The discourse marker marry – a sociopragmatic analysis

*Ursula Lutzky, Vienna**

1. Introduction

The discourse marker *marry*, first attested in c1350 according to the *OED* and, according to Fischer (1998: 38), dropping out of use in the eighteenth and early-nineteenth centuries, goes back to the name of the Virgin Mary, diverging from its source in both pragmatic-functional and semantic terms in the course of its development (*OED*: s.v. *marry*, *int*., 19/08/2008; see also Kellner 1922: 190; Schmidt 1875: 696).[1] Although this discourse marker has not received a great deal of attention in (socio)pragmatic research to date (but see Fischer 1998; Jucker 2002),[2] it has been claimed that *marry* has "potential sociolinguistic relevance" (Fischer 1998: 43). Quoting personal communication with Hans-Jürgen Diller, Fischer (1998: 43) notes that "in Shakespeare it seems to be used above all (though by no means exclusively) by characters of lower social rank".

Using a corpus-based approach, the study at hand tries to shed more light on this hypothesis, which has so far not been tested in empirical analyses. Instead of focusing on Shakespeare's plays alone, a wider perspective will be taken by examining a variety of Early Modern English (EModE) drama text samples in the sociopragmatically annotated 'Drama Corpus' (see section 3, below). This corpus combines the drama files of the *Corpus of English Dialogues 1560-1760* (*CED*), the *Sociopragmatic Corpus 1640-1760* (*SPC*) and the *Penn-Helsinki Parsed Corpus of Early Modern English* (*PPCEME*), resulting in a total of 238,751 words and a time span which reaches from 1500 to 1760.

* The author's e-mail for correspondence: ursula.lutzky@univie.ac.at.

1 Note that in the present study discourse markers are defined as forms with little or no propositional content that are syntactically and semantically optional but have important pragmatic functions on the level of discourse. For a discussion of discourse markers' formal features see Lutzky 2006.

2 For a large-scale corpus analysis of the discourse marker *marry* in EModE see Lutzky forthcoming.

The analysis of the data will involve three different parameters: first, the attestations of *marry* are analysed with regard to social status in order to discover whether *marry* may indeed have been a marker of the lower social ranks in the EModE period; second, its directions of use are investigated to find out whether *marry* is primarily attested in discourse among social equals or mainly appears with a socially upward or downward direction of use; finally, in the analysis of the discourse marker for the parameter 'gender', its distribution in male and female data is taken into account so as to discern possible tendencies in the use of *marry* by the two genders.[3]

2. The *Sociopragmatic Corpus*

One of the most recently established corpora for historical sociopragmatic research is the *Sociopragmatic Corpus* (*SPC*), a sociopragmatically annotated corpus of 219,970 words which is based on a subsection of the *Corpus of English Dialogues 1560-1760* (*CED*). It comprises the time span 1640-1760, i.e. the subperiods 3-5 of the original *CED*, and includes two of the five *CED* text types – drama and trial proceedings – which provide constructed and authentic "interactive, face-to-face, speech-related data, which has only a minimum of narratorial intervention" (Archer & Culpeper 2003: 43; Culpeper & Archer 2007: 4; see also Archer 2005: 107).

As its annotation accounts for (among other factors) the gender and status of both speakers and addressees, the *SPC* would seem to lend itself to a study of the sociolinguistic potential of discourse markers in EModE.[4] However, with regard to the discourse marker *marry*, it turned out that its density of occurrence in the *SPC* is too low to allow for representative results, which relates to the fact that this discourse marker was presumably slowly dropping out of use in the course of the EModE period. As Table 1 shows, *marry* is not very prominently represented in the drama files and is not attested at all in the trial proceedings that form part of the *SPC*. Thus, a sociopragmatic analysis of *marry* based on the *SPC* alone would not lead to any far-reaching conclusions as far as its use by speakers of different status and gender is concerned. Instead, a wider time span which also comprises the first part of the EModE period would need to be investigated.

---

[3] The study at hand is based on the paper "*Marry* – a lower social rank feature?" held at the International Conference on English Historical Linguistics 15 in August 2008 in Munich and forms part of a PhD project on discourse markers in EModE.

[4] For information about the annotation scheme and the individual tag fields and values, see Archer and Culpeper (2003: 43ff.), Archer (2005: 107ff.), or Culpeper and Archer (2007: 5ff.).

Text type	1640-1679	1680-1719	1720-1760
drama	5 (1.3)	1 (0.3)	1 (0.3)
trials	0	0	0

Table 1: The distribution of *marry* in the *SPC*[5]

3. Extending the *SPC* – the data

As the *SPC*'s time frame turned out to be too restricted, I devised a supplement to the drama section of the *SPC* for the present study. This supplement covers the time span 1500-1639 and includes the drama text samples of subperiods 1 and 2 of the *CED*[6] as well as those of the *PPCEME*. The extension of the drama section of the *SPC* by this supplement resulted in a corpus of 238,751 words, which I refer to as the 'Drama Corpus'.[7] Thus, the number of words of the *SPC* drama files could be more than doubled (cf. Table 2) and the time span could be enlarged to also include the first half of the EModE period, now reaching from 1500 to 1760.

SPC	drama	trials	total number of words
	115,800[8]	103,980	219,780
'Drama Corpus'	SPC drama	new drama samples	total number of words
	101,7898	136,962	238,751

Table 2: Word counts for the *SPC* and the 'Drama Corpus'

The text type 'drama' was selected because the claim that *marry* may have been a lower rank feature has primarily been made with reference to EModE drama, in particular to Shakespearean texts (cf. Fischer 1998: 43, quoted

5 Due to the different numbers of words constituting the drama texts in each of the three subperiods of the *SPC*, token frequencies were weighted per 10,000 words. The raw token numbers are followed by the weighted frequencies in brackets.

6 Following the design of the *SPC*, which is based on 12 of the 15 drama samples of subperiods 3-5 of the *CED*, four of the five text samples representing subperiods 1 and 2 respectively were selected for the supplement (the text files D1CLYLY and D2CBARRE were not included).

7 While the use of the term 'Drama Corpus' is intended to facilitate reference to the data set used, it needs to be borne in mind that the 'Drama Corpus' itself draws on text files from other, published corpora (i.e. *CED*, *SPC*, *PPCEME*).

8 Note that the total number of words cited in the *SPC* guide (Culpeper & Archer 2007: 5) for the *SPC* drama files deviates from the word count arrived at for the same files in the 'Drama Corpus'. This is due to the fact that different word count programmes were used in each case and that the *SPC* compilers included, for instance, speaker identifications and stage directions in their word counts, which were excluded in the 'Drama Corpus'.

above). Besides, drama offers several practical advantages over other text types that have survived from the EModE period (e.g. fiction, trial proceedings, witness depositions, letters, sermons). In plays it is possible to distinguish between different speaker turns most easily (cf. speaker identifications) and they may contain a variety of characters of male and female gender and of different social status. The selection of this text type, however, entails that all the data analysed are constructed, i.e. fictional, in nature, which has to be borne in mind when interpreting the results.

The drama text samples of the *CED* (subperiods 1 and 2) and the *PPCEME* are not tagged for sociopragmatic information, in contrast to the *SPC* files. Therefore, the total number of words representing speakers (and/or addressees) from a particular social group or of a certain gender cannot easily be obtained for these corpora. These figures are, however, needed in order to weight frequencies of occurrence and avoid biased results in an analysis of the gender or social rank distribution of a linguistic item. Consequently, I had to tag the relevant files, marking off each speaker turn with an opening <u> and a closing </u> tag, and to classify the characters of each play according to their gender and social status. Both of these steps (the tagging and the application of appropriate sociopragmatic categories) enabled me to extract the total number of words with which the different social groups and genders are represented in the drama text samples.

When designing the *SPC*, Culpeper and Archer (2007: 9f.; see also Archer & Culpeper 2003: 47ff.; Archer 2005: 112ff.) introduced a six-way categorisation in order to account for the social status of the speakers and addressees in their exclusively dialogic data. They based these categories on concepts like rank, estate or sort discussed by EModE contemporaries (e.g. Harrison, Wilson, King) and used criteria like title, ownership or income to delimit the individual layers from each other.

> Nobility *[status= "0"]: Royalty, and those with particular inherited or conferred 'titles' that allow them to sit in the House of Lords, including the Lord's 'spiritual'. Prototypical examples – Duke, Marquess, Earl, Viscount, Baron, Archbishop, Bishop.*
>
> Gentry *[status= "1"]: Upper Clergy and non-hereditary knights not able to sit in the House of Lords, people entitled to carry arms and/or recognised as having the (legitimate) capacity to govern (Wrightson 1991: 38), and those able to append the title esquire (Esq.) to their name (legitimately). Likely to be of a certain income (e.g. substantially above £2,000 per annum) (see Hunt 1996: 16). Prototypical examples – Knight, Sir, Major General.*
>
> Professional *[status= "2"]: Those involved in skilled tertiary-sector occupations, whose focus is upon 'service' (Corfield 1995: 25), including civil servants,*

> *teachers, army and naval officers and members of the 'learned professions' or, to use Addison's (1711) phrase, the 'three great professions' of Law, Medicine and the Church. Prototypical examples – clergymen, lawyers, medical practitioners, school-teachers, military and naval officers.*
>
> *Other middling groups [status="3"]: Those directly involved in trade and commerce (see Hunt 1996: 19), whose focus is upon production or distribution as opposed to service (see Corfield 1995: 25) and whose income is likely to have been between £50–£2,000 (see Hunt 1996: 15) […] They include manufacturers, wholesalers, retailers, merchants, money-lenders, skilled craftsmen, and financiers. Prototypical examples – merchant, shopkeeper, carpenter, shipbuilder, warehouseman, cloth dealer.*
>
> *Ordinary commoners [status="4"]: Those who laboured on someone else's materials or in someone else's fields, household or manufactory, and whose income is likely to have been less than £50 per annum (see Hunt 1996: 21, 15). Prototypical examples – 'labouring folk', yeomen, poor husbandmen, wage labourers, apprentices to the non-professional occupations.*
>
> *Lowest groups [status= "5"]: Common seamen, servants, cottagers and paupers, the unemployed, common soldiers and vagrants. Prototypical examples – servant, vagrant.*
>
> (Archer & Culpeper 2003: 48f.)

These status distinctions were used in the classification of the characters of the drama text samples of the *CED* (subperiods 1 & 2) and the *PPCEME* according to their social rank.[9] Each character was assigned one of the values 0-5 in order to specify his or her social status. In assigning a particular value to a character in the supplement to the *SPC*, I followed the procedures used by Archer and Culpeper (2003: 53; see also Archer 2005: 119; Culpeper & Archer 2007: 11f.) when implementing their annotation scheme. Thus, I referred to three sources of information: secondary data, textual evidence (e.g. speaker-identification labels, participant comments, authorial/editorial comments, specific terms of address) as well as inferential clues (e.g. networks of interaction, patterns of behaviour) – avoiding linguistic evidence because of the danger of circularity. Nevertheless, for some characters the relevant social category could not be determined when referring to any of these sources and I thus created an additional value 'X' for 'unknown' or

[9] As the 'Drama Corpus' builds on the *SPC* and the newly tagged and annotated *CED* and *PPCEME* files may be regarded as extending this core corpus both quantitatively and as far as the time span covered is concerned, I adopted the status classification of the *SPC* for the entire 'Drama Corpus' – not least for reasons of consistency. Note, however, that Walker (2007: 23ff.), for instance, developed a different classification system for the parameter 'social rank' in her study of *thou* and *you* in EModE dialogues, which is partly based on data drawn from the *CED*.

'problematic'. For example, characters that are only very vaguely referred to (e.g. '1st gossip', 'a man'), without any more detailed information about their social background, fall into this group. Furthermore, figures like 'the devil', 'a ghost' or 'a magician' were also attributed to this class.[10]

When all the drama text samples of the supplement to the *SPC* had been tagged, the total number of words with which each of the social groups is represented in the 'Drama Corpus' was determined. The word counts for each social rank, which I extracted from the newly tagged and the *SPC* drama text files with the help of the MLCT (Multi-Lingual Corpus Toolkit), are given in Table 3.

Social group	number of words
0	14,900
1	121,054
2	13,590
3	28,322
4	15,724
5	28,935
X	16,226
total	238,751

Table 3: 'Drama Corpus' – word count per social group

As mentioned in the introduction, apart from the parameters 'social class' and 'directions of use', my sociopragmatic analysis of *marry* also includes the parameter 'gender', which has been claimed, along with several other non-linguistic factors, to influence discourse marker use (see Brinton 1996: 35; Müller 2005: 40f.; for the Present Day English (PDE) discourse markers *you know* and *like* see e.g. Andersen 2001: 287f.; Dailey-O'Cain 2000: 64ff.; Erman 1992: 227ff.; Holmes 1986: 4ff.; Macaulay 2002: 753ff.; Östman 1981: 71ff.; Romaine & Lange 1991: 255f., 267ff.).

In order to enable the analysis of the gender distribution of *marry*, I had to assign one of the values 'male' or 'female' to the characters of the drama texts in the supplement to the *SPC*. As the gender of a character could, however, not always be unambiguously established, I introduced a third value 'X' for 'unknown' or 'problematic'. For example, when a character is referred to in very general terms as 'a child' or 'a servant' and it cannot be inferred from the context whether this character is male or female, this character was

10 While the compilers of the *SPC* introduced the values 'as' for characters speaking in disguise and 'p' for 'problematic' next to the 'X' value for 'unknown', this distinction was not made in the supplement to the *SPC*, but the generic value 'X' was chosen whenever the social rank of a character could not be unambiguously determined. This includes cases in which a character adopts a disguise and speaks in his or her disguised role.

classed as 'X'. The total number of words with which each group is represented in the 'Drama Corpus' was again determined with the help of the MLCT, and the results are given in Table 4.

gender	number of words
male	163,709
female	72,203
X	2,839
total	238,751

Table 4: 'Drama Corpus' - word count per gender

4. Empirical analysis

For the empirical analysis, the 'Drama Corpus' first had to be searched for all attestations of the discourse marker *marry*. Using the concordance programme of *Oxford WordSmith Tools* (Scott 2004-2006), the relevant tokens were extracted from the corpus, paying particular attention to the spelling variations of the discourse marker (namely *mary*, *marie*, *marye*, *marrie*, *mare*, *mari*, *mayry*, *marrye*, *marra*, dial. *marrey*; *OED*: s.v. *marry*, *int*., 19/08/2008) and excluding any tokens with an identical spelling but a non-pragmatic function (cf. e.g. the female name *Mary*, the verb *marry*). In total, I identified 72 attestations of the discourse marker *marry* in the 'Drama Corpus'. While this token number may appear to be small, it needs to be pointed out that *marry* was generally not very frequent in the EModE period as it was already dropping out of use in the course of the sixteenth and seventeenth centuries. This assumption is supported by the fact that in an analysis of the *CED*, the EModE section of the *Parsed Corpus of Early English Correspondence* (*PCEEC*) and selected text types of the *PPCEME*, i.e. a sample of 3,636,193 words, the discourse marker *marry* appears with a frequency of only 228 tokens and declines steadily from 1500 onwards.

4.1 Social status

The sociopragmatic annotation of the 'Drama Corpus' allowed me to determine the social rank of the characters who use *marry* in each of its 72 attestations, so that its tokens could be grouped according to the social status of the speakers. Table 5 shows the social rank distribution of *marry* in the 'Drama Corpus'. The plain token numbers were weighted to the number of words with which each social group is represented in the 'Drama Corpus' and

the normalized frequencies are given in brackets in Table 5.[11] For six attestations, the social status of the character using the discourse marker could not be determined.

DM *marry* used by	tokens (weighted)
0	0
1	14 (1.16)
2	5 (3.68)
3	16 (5.65)
4	10 (6.36)
5	21 (7.26)
X	6 (3.70)

Table 5: Social rank distribution of *marry* in the 'Drama Corpus'

As shown by the weighted frequencies in Table 5, which are graphically represented in Chart 1, the density of attestation of the discourse marker *marry* increases in inverse proportion to social rank. With regard to the top layers of the social hierarchy, one can observe that the discourse marker is not at all attested in the speech of the nobility (0) in the 'Drama Corpus' and its representation in the next category (1) is rather low. In fact, *marry* is least prominently attested among the gentry compared to the remaining social groups and its density of occurrence among the professionals (2) is already more than three times higher. However, as the weighted frequencies of *marry* in the speech of the ordinary commoners (4) and the lowest groups (5) show, the discourse marker is clearly most prevalent in the speech of the lower social ranks in the 'Drama Corpus'. The data at hand consequently provide empirical support for the hypothesis that *marry* may have been used above all by speakers of lower social status in the EModE period. While the claim that *marry* may have been a lower social rank feature was voiced by Diller with reference to Shakespeare's plays (see Fischer 1998: 43, quoted above), my empirical analysis offers a wider perspective by taking a range of EModE drama texts composed by different authors into account. Nevertheless, the 'Drama Corpus' comprises only a single text type of an exclusively constructed nature, and the hypothesis that the discourse marker *marry* has a tendency to appear primarily among the lower ranks and to occur least frequently at the very top of the social scale should be restricted to this type of data.

11 Note that the *marry* tokens listed for each of the categories are attested in the speech of several different characters, i.e. it is not the case that the total number of attestations in any one category is attributable to the idiosyncratic use of the discourse marker by a single character.

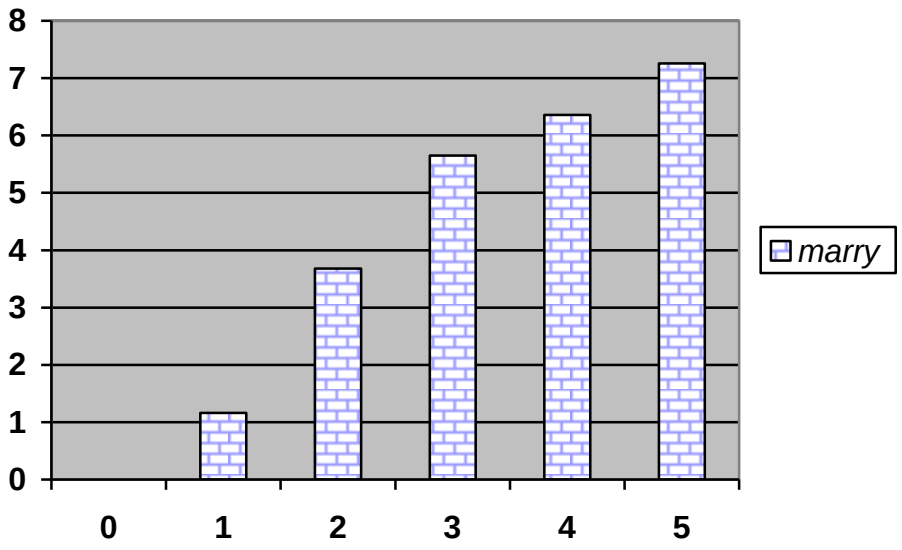

Chart 1: Social rank distribution of *marry* in the 'Drama Corpus'

4.2 Directions of use

While the previous section dealt with the social rank distribution of *marry*, the following analysis focuses on the discourse marker's directions of use. Apart from the social rank of the speaker using the discourse marker, this part of the analysis also takes the addressee and his or her social status into account so that the interactive nature of the discourse marker moves into the foreground.

Directions of use	upward	downward	among equals	X
	33	7	17	15

Table 6: Directions of use of the discourse marker *marry* in the 'Drama Corpus'

As Table 6 shows, the discourse marker *marry* appears with both an upward and a downward direction of use and is also attested among social equals in the 'Drama Corpus'. The last column of Table 6, labelled 'X', indicates that for fifteen of its attestations no direction of use could be determined. Note that the category 'X' also comprises those attestations of the discourse marker *marry* which are not directed to another character in a play, as the speaker using it is talking to him- or herself (e.g. a monologue, an aside).

In total, the discourse marker *marry* shows a direction of use in 57 of its 72 attestations in the 'Drama Corpus'. The majority of these attestations (more than half) have an upward direction of use. Consequently, *marry* primarily occurs in the speech of a socially inferior character when addressing a social superior. As could be expected from its social rank distribution (see

previous section), *marry* is least frequently attested with a downward direction of use. In fact, its use by characters of higher social status talking to their social inferiors is less than half as frequent as its use among social equals.[12] In the following, examples from the 'Drama Corpus' will illustrate each of these directions of use.

Example (1) illustrates the most prominent direction of use of *marry* – the upward direction of use. The example shows an excerpt from Thomas Heywood's *A Pleasant Conceited Comedie, Wherein Is Shewed How A Man May Chuse A Good Wife From A Bad* in which the gentlewoman Mistress Arthur (social group 1) is talking to her servant Pipkin (social group 5). As can be seen in the text extract quoted below, *marry* appears twice in example (1), each time being used by the servant Pipkin when addressing his mistress, i.e. it is attested in the speech of a socially inferior character when addressing his superior. Mistress Arthur is the wife of Young Arthur. While she is very devoted to her husband and tries to please him in whatever she does, he does not appreciate her efforts but wished he had another wife, which in the end makes him attempt to poison her. In example (1), Mistress Arthur is questioning her servant Pipkin, asking him if he has seen her husband, who has been away from home for some time. As can be observed, she has to repeat her question several times because Pipkin keeps giving empty answers. When Mistress Arthur has asked him the same question a third time, Pipkin introduces his answer with the discourse marker *marry*. While *marry* marks the transition from one turn to the next, it also expresses Pipkin's surprise at Mistress Arthur's continuous inquiries, as he apparently feels that his answers are rather straightforward. Furthermore, there appears to be a certain degree of annoyance in Pipkin's use of the discourse marker, which at the beginning of his turn lends additional emphasis to his following words. The second attestation of *marry* in example (1) also appears in response to one of Mistress Arthur's interrogatives, this time inquiring whether her husband was at her father-in-law's house. Pipkin answers this question in the affirmative, with the particle *yes* being immediately followed by *marry*. The discourse marker here

12 Apart from the *SPC* files, the 'Drama Corpus' is only tagged and annotated for the speaker's but not the addressee's social rank. Therefore, the total number of turns which show an upward or a downward direction of use or which are exchanged between social equals is not available. Consequently, the plain token numbers of *marry* could not be normalized with respect to directions of use. While one may hypothesize that the number of turns with an upward and a downward direction of use should be relatively balanced, with regard to the marker's use among social equals, a more sophisticated annotation of the corpus would be needed in order to gain completely unbiased results. This could not be provided for the current study but will be developed in a future project.

shows an intensifying function, rendering the expression of his agreement more vigorous, and could be glossed by PDE *indeed*.

(1) [$ (^Mis. Ar.^) $] Sirra when saw you your Maister?
[$ (^Pip.^) $] Faith Mistris when I last lookt vpon him.
[$ (^Mis. Ar.^) $] And when was that?
[$ (^Pip.^) $] When I beheld him.
[$ (^Mist. Ar.^) $] And when was that?
[$ (^Pip.^) $] **Mary** when he was in my sight, and that was yesterday, since when I saw not my maister, nor lookt on my M. nor beheld my maister, nor had any sight of my M.
[$ (^Mis. Ar.^) $] Was he not at my father in lawes?
[$ (^Pip.^) $] Yes **mary** was he.
[$ (^Mis. Ar.^) $] Didst thou not intreat him to come home?
[$ (^Pip.^) $] How should I mistris, he came not there to day.
[$ (^Mis. Ar.^) $] Didst not thou say he was there?
[$ (^Pip.^) $] True mistris he was there, but I did not tel ye whe~, He hath bin there diuers times, but not of late.
(*CED*: D2CHEYWO, p. E2R, 1602)

Apart from its use in an upward direction, the discourse marker *marry* is attested with the second highest frequency in interactions among social equals, amounting to almost a third of all attestations of *marry* for which a direction of use could be determined. Thus the discourse marker is used with a noticeable frequency by characters who share the same social rank with their addressees. Examples (2) and (3) illustrate this use of *marry* in dialogues between characters belonging to the gentry, i.e. a social group which is situated high on the social scale, and to the lowest groups respectively.

Example (2) shows an excerpt from *The Merry Wiues of Windsor*, which is the only Shakespearean play included in the 'Drama Corpus'. It comprises a dialogue between Justice Shallow, an esquire of genteel status, his cousin Slender, and Sir Hugh Evans, a member of the upper clergy (all social group 1). Justice Shallow and Sir Hugh Evans would like to marry Slender to Anne Page, who apart from having inherited a fortune from her grandfather is also sure to receive a high dowry. When they want to disclose their marriage plans to Slender, Justice Shallow takes his cousin aside and, at first, rather hesitantly touches upon the topic in question (cf. the excerpt quoted below). As can be seen, the discourse marker *marry* is attested within his turn in the phrase "marry this, Coz". It functions as a means through which Shallow (after having used different means for the same purpose before, e.g. the imperative forms "come Coz") directly addresses his cousin and tries to catch his attention, making him listen to what he has to tell him. The discourse marker in collocation with the deictic *this* thus signals that new information is

about to be shared. However, Slender, who is rather indifferent with regard to their conversation, does at first not understand what Justice Shallow and Sir Hugh Evans are trying to tell him, until Evans finally points out rather bluntly that "the question is concerning your marriage", which is confirmed by Shallow and subsequently taken up once more by Sir Hugh. The second attestation of the discourse marker *marry* appears at the beginning of Sir Hugh Evans's second turn, in which he agrees with Justice Shallow and stresses that the point they would like to discuss with Slender is his marriage to Anne Page. In this case, *marry* clearly has an intensifying function, lending increased emphasis to his words with which he confirms once more what has already been said twice. Finally, it should be pointed out that the two attestations of the discourse marker *marry* in example (2) may, in fact, have been consciously chosen by the author and could be regarded as forming part of a wordplay on the forms *marry* and *marriage*.

(2) [$ (^Shal.^) $] Come Coz, come Coz, we stay for you: a word with you Coz: **marry** this, Coz: there is as 'twere a tender, a kinde of tender, made a farre-off by Sir (^Hugh^) here: doe you vnderstand me?
[$ (^Slen.^) $] I Sir, you shall finde me reasonable; if it be so, I shall doe that that is reason.
[$ (^Shal.^) $] Nay, but vnderstand me.
[$ (^Slen.^) $] So I doe Sir.
[$ (^Euan.^) $] Giue eare to his motions; (M=r=. (^Slender^)) I will description the matter to you, if you be capacity of it.
[$ (^Slen.^) $] Nay, I will doe as my Cozen (^Shallow^) saies: I pray you pardon me, he's a Iustice of Peace in his Countrie, simple though I stand here.
[$ (^Euan.^) $] But that is not the question: the question is concerning your marriage.
[$ (^Shal.^) $] I, there's the point Sir.
[$ (^Eu.^) $] **Marry** is it: the very point of it, to Mi. (^An Page^).
[$ (^Slen.^) $] Why if it be so; I will marry her vpon any reasonable demands.
(*CED*: D2CSHAKE, p. 40C2, 1602)

Example (3) comprises an extract from the play *Roister Doister* by Nicholas Udall. Contrary to example (2), where the discourse marker appears in an interaction among characters of genteel status, in example (3) *marry* is attested in a dialogue between characters who may be classed among the lowest groups (social group 5) of Archer and Culpeper's social rank model (cf. above). The discourse marker is attested in an interaction between Tibet Talkapace, Dame Christian Custance's maid, and Dobinet Doughtie, Ralph Roister's servant, i.e. between members of the servant class. Dobinet Doughtie was sent to Christian Custance's house by his master to bring her a token and a ring. Taking Dobinet for the servant of Dame Custance's fiancée,

her maids and her servant Tom Truepenie welcome him heartily and finally bid him come inside to meet their mistress. On the pretext that he still has to go on an errand for his master, he declines their offer and asks them to bring the token and ring to their mistress instead of him. The discourse marker *marry* introduces Tibet Talkapace's answer to Dobinet's appeal in which she agrees to take the present to her mistress. While the discourse marker at the beginning of her turn, on the one hand, functions as a device for catching Dobinet's attention before the other servants speak up, on the other hand, it also conveys her joyous assent and renders her affirmative answer more passionate.

(3) Tib Talk . Wyll you now in with vs vnto our mistresse go?
 D. Dough. I haue first for my maister an errand or two . But I haue here from him a taken and a ring , They shall haue moste thanke of hir that first doth it bring .
 Tib Talk . **Mary** that will I .
 Trupen. See and Tibet snatch not now .
 Tib Talk . And why may not I sir , get thanks as well as you ?
 Exeat .
 An. Alyface . Yet get ye not all, we will go with you both . And haue part of your thankes be ye neuer so loth .
 Exeant omnes .
(*PPCEME*: UDALL-E1-P1, p. L689, 1552-53/1534-41)

Finally, example (4) illustrates the downward direction of use of the discourse marker *marry*. It presents an extract from George Chapman's play *An Humerous Dayes Myrth* in which the courtier Catalian joins a group of gentlemen after having spent some time on the tennis court. On entering the stage, Catalian immediately asks a boy to call for a course napkin. When some time later a maid brings the said course napkin and asks who called for it, Catalian addresses her, introducing his turn with the discourse marker *marry*. The discourse marker is thus used in a downward direction by the socially superior character Catalian (social group 1) when talking to the socially inferior maid (social group 5). The exclamation of *marry* at the beginning of Catalian's turn, voiced in answer to the maid's question, has a clear attention-catching function, indicating emphatically to whom she is supposed to bring the course napkin.

(4) [$ (^Enter Catalian sweating.^) $]
[$ (^Ca.^) $] Boy, I prethee call for a course napkin. Good morrow Gentlemen, (^I^) would you had bin at the tenniscourt, you should haue seene me a beat Monsieur (^Besan^) , and I gaue him fifteene and all his faults.
[$ (^Le.^) $] Thou didst more for him, then euer God wil do for thee.

[$ (^Ca.^) $] Iaques, (^I^) prethee fill me a cup of canary, three parts water
[$ (^Le.^) $] You shall haue all water and if it please you.
[$ (^Enter Maide.^) $]
[$ (^Ma.^) $] Who cald for a course napkin?
[$ (^Ca.^) $] **Marry** (^I^) , sweete heart, do you take the paines to bring it your selfe, haue at you by my hosts leaue.
[$ (^Ma.^) $] Away sir, fie for shame.

(*CED*: D1CCHAPM, p. E2R, 1599)

4.3 Gender distribution

Discourse marker studies taking the gender variable into account are comparatively rare (see e.g. Culpeper & Kytö 2000; Holmes 1986; Müller 2005; Östman 1981).[13] Moreover, most of them are based on PDE spoken data; only Culpeper and Kytö (2000) approach the use of discourse markers (and hedges in general) by male and female speakers from a historical perspective.[14]

With regard to the analysis of the gender distribution of the discourse marker *marry*, it has to be noted that the clear majority of the EModE drama samples included in the *SPC*, the *CED* and the *PPCEME* were composed by males. In fact, only one sample included in the 'Drama Corpus' can be attributed to a woman – Mary Manley's *The Lost Lover*. Of the remaining 23 samples, 22 were written by male authors and one by an anonymous author. Thus, even though female characters may appear in these texts, their 'voices' were often created by male writers, and what we are therefore confronted with is mainly a male vision of gender in Early Modern England. Consequently, the present analysis cannot draw far-reaching conclusions about the use of *marry* by males and females in the EModE period but it can illustrate the distribution of the discourse marker in male and female speech in the data sample at hand.

The quantitative distribution of the discourse marker *marry* in male and female speech in the 'Drama Corpus' is summed up in Table 7. Apart from providing the plain token numbers and weighted frequencies with which male and female characters make use of the discourse marker, Table 7 furthermore

13 Following Raumolin-Brunberg (1996: 13; cf. Romaine 1994: 101), I use the term 'gender' in relation to sex differences, "thus emphasizing the socio-cultural dimension of the division of human beings into male and female persons".

14 Other historical studies discussing gender differences focus on different linguistic phenomena like spelling, verb inflections, first-person expressions of epistemic evidentiality, or exclusive adverbs (see e.g. Kytö 1993; Meurman-Solin 2000; Nevalainen 1991, 1996; Palander-Collin 1999).

takes the addressee into account, indicating how frequently *marry* is attested in same-gender as opposed to mixed-gender interactions.

DM marry…	male	female	X
…used by	49 (0.30)	22 (0.30)	1 (0.35)
talking to male	28 (0.17)	16 (0.22)	1 (0.35)
talking to female	16 (0.10) -> 0.23	5 (0.07) -> 0.16	0
talking to X	5 (0.03)	1 (0.01)	0

Table 7: Gender distribution of *marry* in the 'Drama Corpus'

As the first line in Table 7 reveals, the discourse marker *marry* has exactly the same density of occurrence in male and female speech (cf. the weighted frequencies in brackets). The following lines, i.e. lines two to four, provide the frequencies with which *marry* is used by male and female characters (as well as one character of unknown gender) when talking to male or female addressees (or characters whose gender could not be established). Leaving aside the 'X' category, frequencies were here additionally normalized with regard to the total number of words addressed to male recipients on the one hand, and to female ones on the other. Thus the numbers in italics represent double-weighted frequencies, and were calculated on the basis of the hypothetical estimate that the total number of words addressed to speakers should correlate with the number of words spoken by them. This hypothesis was tested and largely confirmed in a pilot study based on selected *SPC* files, which are annotated for both the speakers' and addressees' gender, and which was then extended and regarded as applicable to the entire 'Drama Corpus'. Focusing first on the use of *marry* by male characters, it can be observed that the discourse marker is attested more frequently when the addressee is a female (0.23) rather than a male character (0.17). Likewise, Table 7 shows that the discourse marker is primarily used by female speakers when talking to male interlocutors (0.22), whereas its density of occurrence is lower with regard to female addressees (0.16). *Marry* thus appears with equal frequency in male and female speech, and no gender difference can be observed concerning the speakers using the marker. However, it seems to be associated with addressees of the opposite gender. This observation is supported by the fact that the discourse marker is most prevalent in male-female and female-male interactions, but shows a noticeably reduced density of attestation in same-gender dialogues, being least frequently used by female characters when addressing females.

The last row and column respectively of Table 7 include those token attestations of the discourse marker *marry* for which the gender of the speaker or the addressee could not be unambiguously determined. Concerning addressees, this may be due to the fact that a character is addressing a group

of people of mixed gender or that it is not clear to whom he or she directs a statement. The 'X' category also comprises those attestations which appear in monologues or asides, i.e. when a character is talking to him- or herself and not addressing another character.

5. Conclusion

The study at hand set out to investigate the sociolinguistic potential of the discourse marker *marry* in the EModE period. For this purpose, I designed a supplement to the drama section of the *SPC*, resulting in the 'Drama Corpus', a sociopragmatically annotated corpus of 238,751 words. This corpus was used as the basis for a sociopragmatic analysis of the discourse marker *marry*, which yielded the following results: first, I obtained empirical evidence confirming the hypothesis that *marry* may have been a lower social rank feature in EModE – a hypothesis that had been posited but never empirically tested before. Second, my analysis of the discourse marker's directions of use revealed that *marry* is attested primarily with an upward direction of use, i.e. in the speech of socially inferior characters addressing their social superiors, whereas its density of attestation in dialogues between social equals is in comparison already almost halved. Finally, although I did not observe a difference with regard to its frequency of occurrence in male and female speech when studying the marker's gender distribution, it turned out that *marry* appears predominantly when a character of the opposite gender is addressed in EModE drama, i.e. it is most prevalent in mixed-gender dialogues. These results may only give an indication of the sociolinguistic potential of *marry* in the EModE period due to the small overall size of the 'Drama Corpus' and its restriction to a single text type which is of a constructed, i.e. fictional, kind. Nevertheless, they provide original, empirical insights into the sociolinguistic nature and distribution of the discourse marker *marry*, which will hopefully be supplemented by future research.

References

Andersen, Gisle. 2001. *Pragmatic markers and sociolinguistic variation: a relevance theoretic approach to the language of adolescents*. Amsterdam: Benjamins.
Archer, Dawn; Culpeper, Jonathan. 2003. "Sociopragmatic annotation: new directions and possibilities in historical corpus linguistics". In Wilson, Andrew; Rayson, Paul; McEnery, Tony (eds.). *Corpus linguistics by the Lune*. Frankfurt am Main: Lang, 37-58.
Archer, Dawn. 2005. *Questions and answers in the English courtroom (1640-1760). A sociopragmatic analysis*. Amsterdam: Benjamins.
Brinton, Laurel. 1996. *Pragmatic markers in English. Grammaticalization and discourse functions*. Berlin: de Gruyter.
CED = *A Corpus of English Dialogues 1560-1760*. 2006. Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University).
Corfield, P. J. 1995. *Power and the professions in Britain, 1700–1850*. London: Routledge.
Culpeper, Jonathan; Kytö, Merja. 2000. "Gender voices in the spoken interaction of the past: a pilot study based on Early Modern English trial proceedings". In Kastovsky, Dieter; Mettinger, Arthur (eds.). *The history of English in a social context. A contribution to historical sociolinguistics*. Berlin: de Gruyter, 53-89.
Culpeper, Jonathan; Archer, Dawn. 2007. *Guide to the Sociopragmatic Corpus. A specialised sub-section of A Corpus of English Dialogues 1560-1760*. Distributed with the corpus.
Dailey-O'Cain, Jennifer. 2000. "The sociolinguistic distribution of and attitudes toward focuser like and quotative like". *Journal of Sociolinguistics* 4/1, 60-80.
Erman, Britt. 1992. "Female and male usage of pragmatic expressions in same-sex and mixed-sex interaction". *Language Variation and Change* 4, 217-234.
Fischer, Andreas. 1998. "Marry: from religious invocation to discourse marker". In Borgmeier, Raimund et al. (eds.). *Anglistentag 1997 Giessen. Proceedings*. Trier: WVT, 35-46.
Holmes, Janet. 1986. "Functions of you know in women's and men's speech". *Language in Society* 15, 1-22.
Hunt, M. R. 1996 *The middling sort: commerce, gender, and the family in England, 1680–1780*. Berkeley: University of California Press.
Jucker, Andreas H. 2002. "Discourse markers in Early Modern English". In Watts, Richard; Trudgill, Peter (eds.). *Alternative histories of English*. London: Routledge, 210-230.
Kellner, Leon. 1922. *Shakespeare Wörterbuch*. Leipzig: Tauchnitz.
Kytö, Merja. 1993. "Third-person present singular verb inflection in early British and American English". *Language Variation and Change* 5, 113-139.
Lutzky, Ursula. 2006. "Discourse markers? Well…". *VIEWZ* 15/1, 3-24. Available online at: http://www.univie.ac.at/Anglistik/views0601.pdf
Lutzky, Ursula. Forthcoming. *Discourse markers in Early Modern English*. Unpublished PhD dissertation, Vienna University.
Macaulay, Ronald. 2002. "You know, it depends". *Journal of Pragmatics* 34, 749-767.
Meurman-Solin, Anneli. 2000. "Change from above or from below? Mapping the loci of linguistic change in the history of Scottish English". In Wright, Laura (ed.). *The development of Standard English 1300-1800*. Cambridge: Cambridge University Press, 155-170.

MLCT = *Multi-Lingual Corpus Toolkit*. 2002. Designed by Scott Songlin Piao.
Müller, Simone. 2005. *Discourse markers in native and non-native English discourse*. Amsterdam: Benjamins.
Nevalainen, Terttu. 1991. *BUT, ONLY, JUST. Focusing adverbial change in Modern English 1500-1900*. Helsinki: Société Néophilologique.
Nevalainen, Terttu. 1996. "Gender". In Nevalainen, Terttu; Raumolin-Brunberg, Helena (eds.). *Sociolinguistics and language history. Studies based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi, 77-91.
OED = *The Oxford English Dictionary*. Online edition. Oxford: Oxford University Press.
Östman, Jan-Ola. 1981. *You know. A discourse-functional approach*. Amsterdam: Benjamins.
Palander-Collin, Minna. 1999. *Grammaticalization and social embedding: I THINK and METHINKS in Middle and Early Modern English*. Helsinki: Société Néophilologique.
PCEEC = *Parsed Corpus of Early English Correspondence, text version*. 2006. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin, with additional annotation by Ann Taylor. Helsinki: University of Helsinki and York: University of York. Distributed through the Oxford Text Archive.
PPCEME = Kroch, Anthony; Santorini, Beatrice; Delfs, Lauren. 2004. *Penn-Helsinki Parsed Corpus of Early Modern English*. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/
Raumolin-Brunberg, Helena. 1996. "Historical sociolinguistics". In Nevalainen, Terttu; Raumolin-Brunberg, Helena (eds.). *Sociolinguistics and language history. Studies based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi, 11-37.
Romaine, Suzanne. 1994. *Language in society*. Oxford: Oxford University Press.
Romaine, Suzanne; Lange, Deborah. 1991. "The use of like as a marker of reported speech and thought: a case of grammaticalization in progress". *American Speech* 66/3, 227-279.
Schmidt, Alexander. 1875. *Shakespeare-Lexicon. A complete dictionary of all the English words, phrases and constructions in the works of the poet. Volume II, M-Z*. Berlin: Georg Reimer.
Scott, Mike. *2004-2006. Oxford WordSmith Tools. Version 4.0*. Oxford: Oxford University Press.
SPC = *Sociopragmatic Corpus*. 2007. Annotated under the supervision of Jonathan Culpeper (Lancaster University). A derivate of *A Corpus of English Dialogues 1560-1760*, compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University).
Walker, Terry. 2007. *Thou and you in Early Modern English dialogues. Trials, depositions and drama comedy*. Amsterdam: Benjamins.
Wrightson, Keith. 1991. "Estates, degrees, and sorts: changing perceptions of society in Tudor and Stuart England". In Corfield, P. J. (ed.). *Language, history and class*. Oxford: Basil Blackwell, 30-52.

A world of words: processes of lexical innovation in VOICE

*Marie-Luise Pitzl, Angelika Breiteneder & Theresa Klimpfinger, Vienna**

1. Introduction

Whenever a language is used, it is adapted to suit the particular contexts in which it is used. This holds true also, or maybe especially, for English as a lingua franca (ELF) which is used in a variety of very different contexts, L1 constellations and regions of the world (cf. Seidlhofer, Breiteneder & Pitzl 2006). Numerous empirical studies on ELF in and outside of Europe have already documented how ELF speakers successfully exploit English as a common communicative resource for their own specific needs and purposes. These studies have focused on various aspects of lexicogrammar (e.g., Breiteneder 2005 and forthc., Dewey 2007a, Hülmbauer 2007 and forthc., Ranta 2006, Seidlhofer 2005), phonology (e.g., Jenkins 2000), and pragmatics (e.g., Böhringer 2007, Cogo 2007, Kordon 2006, Lichtkoppler 2007, Pitzl 2005), but also on ELF speakers' communication of their multilingual identities (e.g., Klimpfinger 2007 and forthc. for code-switching) and the online creation of idioms and metaphors (Pitzl forthc., Seidlhofer & Widdowson 2007).

Building on these studies, the present paper focuses specifically on lexical innovations in ELF, an area that has, to our knowledge, so far not been extensively discussed in ELF research. As Schendl points out,

> *[s]peakers constantly have to adapt language to changing communicative needs in a changing environment. Thus new words are coined, old ones get their meanings extended, while on the other hand existing words or meanings constantly fall into disuse. (Schendl 2001: 25)*

*The authors' e-mails for correspondence: marie-luise.pitzl@univie.ac.at, angelika.breiteneder@univie.ac.at, theresa.klimpfinger@univie.ac.at.

Since ELF reveals such adaptation in progress, the study of its lexical innovations would seem to be of particular relevance, especially if, as Ferguson argues, lexis is "more open than other areas to innovations introduced by speakers grappling with new communicative demands" (Ferguson 2006: 173).

The investigation of lexical innovations in this paper draws on the Vienna-Oxford International Corpus of English (VOICE), the first general corpus of naturally-occurring ELF interactions. In the process of transcribing data for this corpus, we regularly came across words which could not be found in the reference dictionary used for compiling VOICE. Although non-codified, these words seemed to be communicatively effective, and so they were specifically tagged as <pvc> (**p**ronunciation **v**ariations and **c**oinages). While some of these tagged words may be regarded as part of specialized terminology in various disciplines, others appear to be new and innovative. All items captured in this so-called <pvc> tag serve as the basis for the analysis. Our initial observations suggested that most of these items were not coined arbitrarily, but seemed to follow certain trends or processes. This paper is an enquiry into what these trends and purposes might be.

Two interrelated questions guide our enquiry: What are the characteristics of the words tagged as <pvc> in VOICE? And what makes them work (effectively)? Section 2 explains the rationale of the <pvc> tag and elaborates on potential caveats in applying and operationalizing the tag. Section 3 presents an analysis of items captured in the <pvc> tag in a subcorpus of VOICE, bringing the newly coined words in relation to attested word-formation processes in English. Section 4 looks at the functional aspect of coining words in spoken interactions and suggests some general functions ELF coinages seem to fulfil.

2. Pronunciation variations and coinages in VOICE

2.1. A subcorpus of VOICE

VOICE, the Vienna-Oxford International Corpus of English, is the first general corpus of naturally-occurring spoken ELF and has been compiled at the University of Vienna by the authors of this paper and the IT assistant Stefan Majewski under the direction of Barbara Seidlhofer. VOICE comprises just over one million words of transcribed speech, which equals close to 120 hours of recordings and approximately 150 different speech events, which are included in their entirety. The interactions recorded are naturally-occurring and interactive, and happen face-to-face. From its early beginnings, VOICE

was designed and compiled in order to provide a "broad empirical base" (Seidlhofer 2001: 149) on which a thorough description of ELF could become possible and will therefore be made available online (cf. http://www.univie.ac.at/voice).

The following analysis is based on a subcorpus of VOICE, capturing 250.042 transcribed words, i.e. a quarter of the entire corpus. These data chosen for analysis mirror the overall corpus design of VOICE and its target proportions. Accordingly, the subcorpus includes speech events from the educational, leisure and professional domains, with the professional domain being subdivided into professional organizational, professional business and professional research/science. Table 1 reflects the distribution of the 35 speech events included in the subcorpus across the domains represented in VOICE:

%	Domain	Speech events	Number of words
25%	Educational	8	61.567
10%	Leisure	7	25.068
20%	Professional Business	6	50.244
35%	Professional Organizational	11	87.893
10%	Professional Research/Science	3	25.270

Table 1: Subcorpus of VOICE

The speech events included in the subcorpus were chosen according to their distribution across the five domains found in VOICE, but not on the basis of internal linguistic features. Additionally, the subcorpus was sampled with the aim of including as many different speakers as possible, which means that speech events recording a group of people already found in the subcorpus were generally not chosen if another event could be selected which included new speakers. Furthermore, the subcorpus was selected to include a variety of speech event types and comprises one press conference, two service encounters, two seminar discussions, six working group discussions, five workshop discussions, six meetings, three panels, and ten conversations.

2.2. Defining the <pvc> tag in VOICE

Compiling the first large-scale corpus of spoken ELF naturally involves many challenges, not the least of which is deciding on a standardized, unambiguous and replicable system for transcribing the data. First, we had to decide on how to represent spoken ELF orthographically – whether to stick to British English spelling or American English spelling or go for a different spelling system altogether (cf. Breiteneder et al. 2006 for a detailed discussion of our solution for VOICE). Transcribing spoken ELF, however, also involves questions such

as how to actually deal with words which are not part of contemporary English vocabulary, i.e. how to tackle lexical innovations or pronunciation variations.

Since the early stages of the VOICE project in 2001, the transcription conventions have been developed to account for the specific needs of ELF data. The growing amount of experience in transcribing naturally-occurring ELF has resulted in several revisions and improvements of the VOICE Transcription Conventions over the years. While earlier versions already included a tag for newly coined words and the option to mark unusual pronunciations whenever they affected the interaction, the extensively revised version [2.0] of the *VOICE Transcription Conventions*, published in September 2005, marked the introduction of the <pvc> tag.[1]

This <pvc> tag is defined as capturing "[s]triking variations on the levels of phonology, morphology and lexis as well as 'invented' words" (VOICE Project 2007a: 4) in the Mark-up Conventions, which together with the Spelling Conventions (VOICE Project 2007b) form our Transcription Conventions. Its broad definition already points to the difficulty of establishing precise rules about what 'goes into' this tag and what does not. It also highlights a complex of problems involving, at least, three basic questions: 1. Why would we want to subsume different levels of language under just one tag?, 2. How can we define variation on the levels of morphology and lexis? and 3. What makes variation on the level of phonology 'striking'?

To start with the first of the issues just outlined, it should be stressed that in the early stages of discussion, we, in fact, intended to distinguish between lexical coinages on the one hand and pronunciation variations on the other and sought to capture them in two different tags. But throughout the discussion and also transcription process, it became apparent that a clear and unambiguous distinction and categorization of phonological, lexical and morphological variation was simply not feasible. What to do, for example, with a speaker saying *importancy*? Is it just the word *importance* with the last syllable pronounced with a full vowel, and consequently – although the speaker adds an extra syllable to the existing word – 'only' a pronunciation variation? If asked, would the speaker still write down *importance* and only pronounce it *importancy*? Or would the speaker write down *importancy* and shall we hence conceive of it as a morphologically and orthographically different and thus 'new' word? Both interpretations are, of course, tenable and

[1]The current version of the VOICE Transcription Conventions [2.1] is available at http://www.univie.ac.at/voice/voice.php?page=transcription_general_information (9 October 2008).

because we do not have access to the speakers' self-reports and judgements there is no way of arguing that one interpretation is 'more correct' than the other. Moreover, spoken language, and it is this which we try to represent in our transcripts, is intrinsically and naturally bound up with pronunciation. To choose two different tags for the morphological and the phonological level and thus try to separate two constituent parts of an integrated whole would presumably have led to confusion and even greater imprecision in an already 'messy' area.

So, how did we then define striking variations on the level of morphology and lexis, to start with one of the two layers catered for with the <pvc> tag? Basically, if something varies, there must be something it varies *from*. This means that one needs to know what is 'normal' or 'established' in order to judge what is 'different' or 'new'. Of course, one might choose to make this distinction on the basis of intuition, but it seems prudent to look for a more reliable point of reference on which to base this distinction, i.e. a dictionary or corpus. Dictionaries and corpora provide such a stable reference point in that they record and capture language as it is used at a specific point in time. In so doing they create abstractions, recording a snapshot of the synchronic state of the continuously evolving and changing lexicon of a language. Lexicographers then take upon themselves the additional task of categorizing the lexicon of a language into lemmas and, depending on the size, scope, and purpose of the dictionary, define what is 'normal' usage and what is not.

For the purpose of our project, we chose to rely on the *Oxford Advanced Learner's Dictionary 7th edition* (OALD7) as our main point of reference. It should be noted that the OALD7 is not used as a dictionary, i.e. as an authority on matters of correctness, but as a reference tool or manual to support the transcription process and tagging decisions. A number of different reasons motivated our decision to rely on the OALD7 for these purposes (cf. also Breiteneder et al. 2006: 179-181). First of all, the OALD7 constitutes a comprehensive manual, yet it is not too detailed and, in contrast to the OED, for example, it does not include obsolete words. Published in 2005, the year in which the extensive revision of the VOICE Transcription Conventions took place, the OALD7 could also be considered new and up-to-date. For purely practical reasons, our reference tool had to be accessible not only to ourselves but also to our transcribers, whom we could easily equip with the *OALD7 Compass CD-Rom*. The CD-Rom, additionally, provided a stable point of reference which a continuously updated online version of the dictionary could not have offered.

In deciding which lexical items are tagged as <pvc>, we followed a purely operational definition in accepting all those lexical items listed in the OALD7

as 'existing words'[2]. These 'existing words' include main entries but also subentries and words which are used in examples, definitions or explanations in the OALD7. It was deemed necessary to also include them in our definition of 'existing words' since not all derivatives of a word have an individual entry in the OALD7. The noun *conscientiousness*, for example, only has a subentry within the main entry for the adjective *conscientious* and the hyphenated adjective *real-time*, to mention another example, only occurs in a sample sentence within the entry for the noun *real time*. This is, however, not to say that these 'existing words', which are not tagged as <pvc>, need to be used with one of the meanings codified in the OALD7 or in a syntactically 'correct' position in our data. An 'existing word' may also be used with an entirely new or different meaning or it can be used in another syntactic category.[3] Yet, as long as the word itself can be found in the OALD7 it is not tagged as a <pvc>.

In turn, all those lexical items uttered by the ELF speakers and not found in the OALD7 are considered lexical variations and are consequently tagged as <pvc>. Seeing that there is no rule without exceptions, we left untagged all words which are names of countries, regions, and currencies, as well as proper names and acronyms. Note too, that the <pvc> tag only captures individual lexical items. We do not tag two-, three- and multi-word compounds or hyphenated words if the individual components of the compounds are 'existing words'. This decision not to mark compounds is mainly due to limitations of time as well as practical feasibility, but also corresponds to the fact that the first release of VOICE, i.e. VOICE [1.0] Online, contains no syntactic annotation.

In contrast to the lexical and morphological level, it is not as straightforward to decide and define what a striking variation on the level of phonology is. ELF speech, in particular, is characterized by many 'unusual' pronunciations due to the influence of the multilingual and multicultural settings in which ELF discourse takes place. While we could have followed the model of some conversation analysts who render naturally-occurring spoken language in free spelling, adhering to standard spelling was considered the better alternative. Transcripts in free spelling often look odd, are hard to decipher and contort the speakers' language towards the non-

---

2 Of course, the term 'existing words' is not to suggest that all words which are not in the OALD7 do not exist. This wording merely reflects our operational procedure for tagging PVCs in VOICE: if a word can be found in the OALD7, for the purpose of tagging PVCs, it 'exists' and is not tagged.

3 This is to say that, although conversion with zero derivation occurs in VOICE, it is not systematically tagged and not captured in the <pvc> tag due to practical limitations of time.

standard and by implication unfortunately often towards the inferior. This is also the reason why we decided not to indicate so-called 'minor' pronunciation variations, i.e. those which do not affect syllable number. Conversely, only those phonological variations which affect the number of syllables, either by adding one or more syllables or by reducing the word by at least one syllable, are considered 'striking' and are consequently tagged as <pvc>.

It will be obvious at the end of this section that it is a methodological challenge to define clear-cut and reliable, i.e. replicable, rules for what is tagged as a <pvc> and what is not. In compiling an ELF corpus, operational decisions need to be taken when it comes to the question of defining pronunciation variations and coinages (PVCs). It is this corpus perspective that classifies the <pvc> tag as an excellent starting point but not as the result of a fully-fledged analysis of lexical innovations in ELF. It is the aim of the following section to conduct and present the first systematic analysis of a part of this pool of lexical items tagged as PVC in VOICE.

3. The analysis of PVC forms

3.1. Theoretical background and methodology

The practice of word-formation is nothing new. On the contrary, the extension and reconstruction of the lexicon is an essential property of natural languages. There are basically two ways of introducing new words into a language, namely borrowing and word-formation (Schendl 2001: 25). Borrowing is defined as "the process of introducing a linguistic feature, especially a word or a grammatical feature, from another language or variety" (Schendl 2001: 124). The theory of word-formation deals with the structures which underlie the formation of new words from already existing ones, i.e. the "set of processes by which lexical items are derived from, or related to, other lexical items" (Trask 1995: 305). This set includes several word-building processes, such as the use of affixes, of which suffixes and prefixes are the most common types (cf. Plag 2003: 72-106). Similarly, "the formation of a new lexeme by adjoining two or more lexemes" (Bauer 1988: 33), i.e. compounding, is a major type of building new words and can be found in almost every language. Other ways of building words along the rules of word-formation are conversion, reduplication, modification, and shortening of the base, e.g. via backformation or truncation, and the blending of two existing words into a new one (cf. Bauer 1988, Plag 2003).

The approach we adopted in categorizing the lexical innovations found in the <pvc> tag in the subcorpus of VOICE can be described as both top-down and bottom-up. After reviewing relevant literature in the field of word-formation and morphology, we went through the data with the aim of establishing broad categories according to which we then classified the PVCs. In doing so, we took Plag's (2003) categories as our main point of reference, but adapted them in accordance with other researchers' findings and categories (e.g., Adams 2001, Bauer 1988, Biermeier 2008, Schendl 2001) and the fact that we were working with spoken language. As we are dealing with ELF, a multilingual environment and thus a site of language contact, we also found borrowing a relevant category in our data. The 12 categories we established in our analysis are: suffixation, prefixation, multiple affixation, truncation, borrowing, compounding, analogy, reanalysis, backformation, blending, addition and reduction. It should be emphasized that we are applying the terminology taken from the literature in a rather loose sense, for the two reasons pointed out by Plag:

> *First, adopting a certain type of terminology often means committing oneself to a certain theoretical position […], and second, adopting a particular theory is often unnecessary for the solution of particular empirical problems. (Plag 2003: 179)*

In operationalizing the categories and classifying the PVCs it soon became obvious that, even though we first agreed on studying them as lists of lexical items, there is simply no way of classifying many of the individual PVC occurrences without studying their co-text and context of use (cf. Widdowson 2004: 59-73 for a discussion of the terms 'co-text' and 'context'). The word ***misstand***,[4] for example, might easily be categorized as 'prefixation' with the prefix *mis-* being attached to the base form *stand* in order to indicate negative implications. Looking at the transcript, however, it turns out that *misstand* fits better into the category of reduction, since what S1 refers to is obviously *misunderstand*.

4 To avoid ambiguity all examples taken from the subcorpus of VOICE are indicated in bold print and italics the first time they appear in text.

Extract 1: PBmtg27; S1= German (DE)

1010 S1: THEN a lot of discussions and a lot of this and we had agreed to a certain procedure but then hh (.) after people recognized how the procedures REALLY was there was a lot of discussions as well. (1) also within OUR department cos somebody <pvc> **misstand** {misunderstand} </pvc> something <pvc> wrongly </pvc> (.) so? (2) more or less (.) [first name31] will (.) give this clearly? (.) then later on (.) we will (.) THIS year most likely handle christmas and thirty-first (1) AS per the law. (1)

Moreover, the process of analysis revealed that often different interpretations are possible for one PVC, depending on which word is identified as the 'root' or base word from which the new form is being derived. ***Pronunciate***, to mention another example found in the data, could be interpreted as backformation with *pronunciation* being the base word and the verb *pronunciate* following from the deletion of the suffix *-ion*. The same innovation could also be categorized under suffixation if *pronounce* is regarded as the base form to which the verbal suffix *-ate* is being added. As a third alternative, one could argue that *pronunciate* is a blend and formed via combining the words *pronounce* and *enunciate*. Co- and context would allow for any of the three alternative categorizations, and speakers simply cannot be asked retrospectively what it is that they meant to say at the particular time. But even if only one root word is identified, however, there are often several processes at play and items may be assigned to more than one category.[5]

The present analysis deals with all lexical items tagged as PVCs in our subcorpus, irrespective of whether these items can be found in dictionaries other than the OALD7. The examples presented in the following analysis thus include both lexical innovations which appear to be coined ad hoc as well as technical terms commonly used in discussions of special subject matters (e.g., ***commodification***, ***annihilator***, ***orthonormal*** in mathematics). Consequently, some readers might react with surprise with regard to some words and think: 'Hold on, this is not a new word. This word already exists.' Indeed, our initial idea was to exclude all ESP terminology from our analysis a priori. But we decided against such a procedure during the analysis for two main reasons: Firstly, it needs to be borne in mind that any distinction between general and special (ESP) vocabulary is always to some degree arbitrary and depends on the context in which a word is used. What is 'normal' in one context and for one person may be 'new' in another context and for another person. Secondly, items of special terminology, which have often only been coined recently and

5 With three researchers working on the same data set, we also found that such classifications are, of course, to some degree always subjective.

are, diachronically speaking, young, go back to the same word-formation processes which are also observable in words which are coined ad hoc and are not part of any discipline. After all, whether two mathematicians use ELF in order to discuss a theorem or to chat about the weather, they rely on the same word-formation rules, it seems, in order to achieve mutual understanding. And from the point of view of a 'layperson' who is not part of a particular group of experts, most technical terms actually are coinages which may be more or less semantically transparent and understandable.

3.2. Analysis: processes leading to lexical innovations in ELF

All in all, there are 247 different PVCs (i.e. types not tokens[6]) in our subcorpus of approximately 250.000 words. As specified above, 12 categories turned out to be relevant in describing the kind of processes of innovation captured in the <pvc> tag. Table 2 presents these categories and specifies the number of types of PVCs which were found in each category. The figure given in parentheses indicates how many of the PVCs in the respective category were also assigned to another category (double categorization).

Category	**Number of types (double categorization)**	**Category**	**Number of types (double categorization)**
Suffixation	85 (10)	Backformation	4 (3)
Prefixation	65 (2)	Blends	6 (2)
Multiple affixation	19 (4)	Addition	10 (5)
Borrowing	13 (2)	Reduction	19 (4)
Analogy	24 (4)	Compounding	5 (1)
Reanalysis	7 (2)	Truncations	3 (1)

Table 2: Number of types found in word-formation categories

Except for seven PVCs, all instances found could be assigned to at least one of these categories. The seven remaining PVCs could not be assigned to one of the established categories because we could not determine what these words actually meant or were derivations from or because they were obvious slips of the tongue (e.g. due to swapped syllables). In the following, we will

6 This analysis is primarily concerned with the processes leading to lexical innovations. Considering the type-token-ration would clearly lead to further questions such as whether a coined word is taken up by other speakers in the same speech event. While these questions go beyond the scope of this paper, the authors will address them in subsequent analyses.

focus on the most prominent categories and illustrate our discussion with examples from our data.

** Suffixation, prefixation and multiple affixation*

Generally, the most common and also most frequent way of building new words in any language is by using affixes (cf. Bauer 1988: 19f). This also holds true for our data, where the majority of PVCs is formed by attaching one or more pre- or suffixes to a base word. Broadly defined, an affix is "a bound morpheme that attaches to bases" (Plag 2003: 72),[7] in the case of a suffix at the end of a word in order to form derivatives. Suffixes are classified according to the syntactic category the derived words belong to (e.g., *-ness* derives nouns, *-able* adjectives and *-ize* verbs). This way, nominal, verbal, adjectival and adverbial suffixes can be distinguished (cf. Plag 2003: 86-98), all of which are represented in our data. Some examples of suffixation from the subcorpus of VOICE are ***claustrophobicy***, ***conformal***, ***contentwise***, ***cosmopolitanism***, ***devotedness***, ***forbiddenness***, ***gatheral***, ***imaginate***, ***increasive***, ***increasement***, ***opportunality***, ***preferently***, ***publishist***, ***turkishhood***, and ***workal***.

Some of these words seem to fill what Clark (1994) terms "permanent gaps" in the lexicon.

> *More important are permanent gaps where there is no conventional word with the requisite meaning. Here, speakers frequently coin words just for the occasion, in a particular conversation with a particular addressee. (Clark 1994: 785)*

In the OALD7 there is no word that expresses the idea of 'the state of being forbidden'. It seems only logical that an ELF speaker who feels the need to express exactly this idea forms the noun ***forbiddenness*** by attaching the nominal suffix *-ness* to the adjective *forbidden*.

Taking a closer look at the group of suffixed words in our data, we find that suffixes are not necessarily attached in order to alter the word class of the base form. Rather, in some instances they seem to emphasize the original word class, as is the case of the following example:

Extract 2: PBmtg27; S1=German (DE), S7=German (DE)
1046 S1: you get (.) EACH YEAR (1) an <pvc> **increasement** {increase} </pvc> of your salaries. (.) which is paid by the company. (.)
1047 S7: <soft> yeah </soft> =

[7] For a detailed discussion of this definition and the methodological considerations and implications involved see Plag 2003: 72-86.

In this meeting of a forwarding agency, the PVC ***increasement*** is obviously used as a noun by S1, who talks about the raise of salaries of the employees of the company. In ENL, the orthographic word form *increase* constitutes a verb as well as a noun, in spoken language the two forms are only distinguished via stress. In the example, the nominal suffix *-ment* is attached to the base form, which stresses the nominal word class and distinguishes it from the verb. By this means, clarity and explicitness are increased, which supports Seidlhofer's (2005) initial observation that ELF speakers tend to add elements (e.g., nouns or prepositions) to make the propositional meaning clearer. Another noun in our data that shows a similar pattern is ***supportancy***. Here the base form *support*, which again is used as a verb and noun in ENL, is supplemented by the nominal suffix *-ancy* to stress its function as a noun. Similar processes can also be observed in examples of other word classes. In the cases of ***characteristical*** and ***linguistical*** the adjectival suffix *-al* is being attached to base forms which are already adjectives. These observations indicate that one function of suffixation in ELF is increasing clarity via what we call 'overt word-class marking'.

The second large group of PVCs within the group of affixation are those formed with prefixes, which, in contrast to suffixes, alter and qualify the meaning of a word (cf. Plag 2003: 72). Accordingly, prefixes are classified semantically into locative prefixes, temporal prefixes, prefixes that quantify the meaning of the base word and those that express negation (cf. Plag 2003: 98-101). Generally, one specific prefix is not restricted to being attached to one specific word class and at the same time prefixes do not change the class, i.e. a prefix attached to a noun forms a new noun.

The two most common prefixes in our data are ***non-*** and ***re-***, leading to the formation of 19 out of 65 types of prefixation, such as ***non-confidence***, ***non-formal***, ***non-graduate*** and ***non-transparent*** or ***re-enrol***, ***re-read***, ***re-send*** and ***re-orient***. The emphasis in these words is clearly placed on the semantic level, as these two prefixes present a general, but straightforward and economical, way of expressing the idea of reversal and repetition respectively. This tendency of economy of expression can also be seen in the following example, where a speaker uses the prefix ***pre-*** to express the idea of 'prior to; before'.

Extract 3: POwgd14; S1=Swedish

971 S1: developed. er in each case (.) no? hh (.) but i think er: (.) if you talk about er interdisciplinary er er joint er programs that SOME part (.) er that wou- could be very interesting wo- would be hh (.) very interesting if it was er:m er developed as new. as a sort of an intersection of of er (.) the idea what you can contribute from different sides and make some part perhaps it's (.) the most er sort of specialized (.) <pvc> **pre-thesis** </pvc> (.) a course <@> so to say </@> that could be more integrated and new. (2) i <fast> you understand what i mean (.) no? </fast><3> er er </3> oh i think yeah. er (.)

972 SX: <3> mhm </3>

In this working group discussion on joint degree programmes in Europe, S1 talks about interdisciplinary courses. Instead of elaborating on the concept of a compulsory paper that has to be written in a certain course preceding the actual thesis – a rather complicated matter even if it is put down in writing – the speaker expresses the concept in a more economical way via coining the word ***pre-thesis***.

Within the category prefixation, we found a number of prefixes to belong to the ambivalent group of 'neoclassical elements' which involve lexemes of Latin or Greek origin (cf. Plag 2003: 155-159), such as ***anti-***, ***bio-***, ***geo-***, ***hyper-***, ***meta-***, and ***mono-*** from our data. These word forms are ambivalent in the sense that they may combine with other combining forms, which is something affixes do not do, but is a characteristic of compounds. This is the reason why some researchers classify neoclassical elements as compounds (cf., e.g., Bauer 1988, Plag 2003). Since compounds are generally not captured in the <pvc> tag (see section 2) and some of these neoclassical word forms are listed in the OALD7 as prefixes, we decided to discuss them in the group of affixes.[8] Among these we find the following words: ***anti-dumping***, ***anti-marketing***, ***biopolitics***, ***geostrategical***, ***hyper-complex***, ***meta-perspective***, and ***monodisciplinarity***.

The third group of PVCs formed via affixation are derived by attaching more than one affix, i.e. they show multiple affixation (cf. Plag 2003: 38f). Either a pre- as well as a suffix are attached to one word, as in ***overdebted*** and ***pseudo-conformal***, or suffixes or prefixes occur in sequences. Following this second pattern, ***regionization*** seems to be coined with two suffixes. A verb is derived from the noun *region* via adding the verbal suffix *-ize* (→ *regionize*) from which in turn the speaker derives a noun via adding the nominal suffix *-ation*. Further examples along the same line are ***urbanistic*** and ***re-emplace.***

8 Indeed, Kastovsky (forthc.) even argues for completely abandoning the category of combining forms in this context. Other characteristics inherent to neoclassical elements that point, e.g., towards affixation, are sufficient to analyze the resulting formations.

In the latter case, one might also hazard the guess that ***re-emplace*** is an example of prefixation (and not multiple affixation) with the prefix *re-* being attached to the base word *emplace*. According to the OALD7, the verb *emplace* and the corresponding noun *emplacement* are technical terms used in the field of weaponry. A look at the co-text of its occurrence, however, suggests that this interpretation does not fit the context.

Extract 4: POwgd375; S3=French, S7=Dutch, S9=English (GB)

228 S3: and i <3> i know </3> that i know in two years the person who will <pvc> (**re-emplace**) </pvc> me will not know ANYTHING about it <4> AGAIN </4> (.)
229 S9: <3> (that's true) </3>
230 S7: <4> right </4>
231 S3: and (.) we start the whole story again (.) and i:n order to be: yeah to to be aware of (.) what is going on you have to be <un> x </un> most professional and be in your office every day and for ten years

S3 talks about the problems involved when taking over a new position in an organization. ***Re-emplace*** (line 228) thus obviously conveys the idea of 'substituting somebody' or 'taking over somebody's position'. This suggests that the word is derived from the base word *place* via attaching two prefixes, i.e. multiple affixation.

Considering the speaker's first language another interpretation is possible. S3's first language is French, which makes it seem likely that she borrows the word ***re-emplace*** from French, where F *remplacer* means 'to replace, to substitute'. This shows that in some cases even explicit categorization rules can still be ambiguous, and it is not easy (and in some cases not possible) to assign a PVC to one category only.

** Borrowing*

As has been indicated in the last example, a further category we could establish is that of borrowing, that is introducing a word from another language into English, which is, generally speaking, the result of language contact. When looking back in the history of the English language, it can be seen that there is a long tradition of borrowing: a large number of words found their way into the English lexicon due to contact with other languages (cf. Schendl 2001: 56). ELF involves speakers from a variety of linguacultural backgrounds and thus finds itself in a situation of language contact which is unique. As 'multi-competent' individuals (cf. Cook 2002), ELF speakers play their part in the multilingual context of ELF conversations and influence of their first and other languages, such as code-switching and cross-linguistic influence, is well documented to form an intrinsic part of ELF (cf. Hülmbauer 2007 and forthc., Klimpfinger 2007 and forthc.). Thus, borrowing appears to

be a completely natural and indeed expected phenomenon in ELF, with 13 examples of PVCs illustrating this.

Particularly when analyzing instances of borrowing, considering the co-text of the PVC in more detail is of prime importance for subsequent analyses. Taking a look at the PVC ***decreet*** might suggest that the speaker omitted a consonant while probably aiming at *discreet*. The context, however, shows that the speaker, when talking about politics, means to say *decree* and borrows from his first language Dutch, where the English word *decree* translates as *decreet* in Dutch.

Extract 5: LEcon227; S1=Dutch (BE), S2=Danish

65 S1: seven prime ministers =
66 S2: = but (.) how much power do they have? as (.)
67 S1: quite a lot =
68 S2: = they must have different (.) tasks
69 S1: yeah (.) but quite a lot actually it's very much (.) so (.) erm (.) er de- a <pvc> **decreet** {decree} </pvc> (.) has the same power as a law (.)
70 S2: yeah (.)

Similarly, the German speaker featured in the next example borrows from his first language, when he talks about a party and describes the whole, to his mind rather surreal, situation as follows:

Extract 6: LEcon573, S1=German (DE), S2=Italian

70 S1: = with champagne glasses standing there not SPEAKing because there was hh (.) VERY loud techno music <ono> ə tʃə tʃ tʃ də tʃə </ono>
71 S2: @ @@@@@@ <9> @@ </9>
72 S1: <9> and you </9> couldn't communicate? (.) @@
73 S2: so they were just standing a<7>round in front of </7>
74 S1: <7> no JUST STANDING </7> there yeah
75 S2: hh <@> people</@> (.) uhu: (.)
76 S1: it was like a surreal <pvc> **inscenation** </pvc> or something

A first glance at the PVC ***inscenation*** (line 76) without co-text might indicate that S1 means to say *scene*, which, however, does not quite fit the context. The idea the speaker presumably had in mind when trying to give a close picture of the situation at the party, was what is expressed by *Inszenierung* in German. This seems to have influenced the borrowing in the first place. The similarity of form and meaning (cf. Hülmbauer 2007: 26f) of the three words, ***inscenation***, *Inszenierung* and *scene* stimulates a similar picture in the mind of the speaker, which further enhances the formation of the PVC.

** Analogy, reanalysis, backformation and blends*

In the course of analyzing and categorizing the PVCs in our subcorpus, it has also become apparent that many of the words that can be found in one of the categories just discussed could also be analyzed in terms of analogy, i.e. "[a] process by which a form or a pattern becomes more similar to another (usually more regular) one, e.g. mouses for mice, in analogy with regular plural *-(e)s*" (Schendl 2001: 123). ***Unformal***, to mention one of many examples, can be analyzed as a combination of *formal* and the prefix *un-* , with the prefix *un-* indicating the reversal of a state. Similarly, *unformal* could be analyzed as being used and created in analogy with a large set of morphologically related words such as *unable* or *unhappy*, all combining with *un-* and all sharing important aspects of meaning. These interpretations, in fact, represent two sides of the same coin and refer to the basic distinction between "word-based and morpheme-based morphology" (Plag 2003: 179). The first interpretation is in line with the idea of morpheme-based morphology in that "[i]n this model of morphology, morphological rules combine morphemes to form words much the same way as syntactic rules combine words to form sentences" (Plag 2003: 180). The conceptual framework of word-based morphology, on the other hand, focuses on the "relationship between morphologically related words" (Plag 2003: 184), as illustrated in the second interpretation of *unformal* above.

Following this idea of word-based morphology, analogy "can be modelled as a proportional relation between words" (Plag 2003: 37). Accordingly, ***thinked***, ***catched***, ***drived***, ***feeled***, ***losed***, ***putted***, ***selled***, ***sended***, ***splitted***, ***teached*** or ***thrusted***, all of which are examples from our data, are formed in analogy with other regular past tense forms such as *walked*, *rushed* and *sounded*. Similarly, ***advices***, ***ambivalences***, ***fundings***, ***informations*** and ***knowledges***, to mention yet some further PVCs, are established in analogy with regular plural nouns. To put it into proportional relations *apple* : *apples* = *advice* : *advices* or *apple : apples = information* : *informations*. While in these examples the words themselves are not newly coined, the regularized forms derived via inflections and plural marking are nevertheless innovations and diverge from their codified form. As highlighted by Schendl (2001) in the definition of analogy quoted above, analogical processes also tend to establish more regular words. The PVCs just listed clearly illustrate this process of regularization by analogy and thus confirm a general tendency that has also been discussed with reference to other levels of ELF speech (cf. Breiteneder 2005: 12ff).

Related to analogy but less frequent in our data is the process of reanalysis. Adams defines reanalysis in the following way:

> *When a complex word whose structure is perceived in a certain way is compared to other words to which it can be seen as somehow similar, it may be reanalysed, and perceived as having a different structure, thus paving the way for an abductive change. (Adams 2001: 133)*

Two examples of reanalysis we find in the subcorpus of VOICE are the forms ***medias*** and ***criterias***. On the one hand, the use of an *-s* suffix represents a regularization in plural marking where we normally find an irregular plural form which is derived from Latin. But on the other hand, it is noteworthy that the *-s* suffix is attached to a form that is already in the plural and not, as would also be possible, to the singular form (leading to the equally possible and regularized forms *mediums* and *criterions*). The irregular Latinized plural ending in *-a* thus appears to be reanalyzed in both cases and becomes the alleged base form in the singular, which is in turn marked with an *-s* suffix to indicate plural. A similar process seems to happen in the word ***displayses***. The plural form *displays* appears to be reanalyzed as singular and the *-es* suffix (regularly used for singular nouns ending in *-s*) is attached to mark the plural.

The three other instances of reanalysis we find in the subcorpus also seem to follow a common pattern in that the past tense forms are reanalyzed as present tense stems and inflected (***cames***) or used to create derivatives (***spokers***, ***spoking***). Like with other examples before, one needs to look at the immediate co-text of *cames* in the interaction to determine that it is an example of reanalysis:

Extract 7: PBmtg27, S1=German (DE)
249 S1: so e:r also for the rest if er [first name18] [last name18] <pvc> **cames** </pvc> to you with [org10] rates you refuse to talk to him (.)

The form *cames* in this context seems to indicate present tense with reference to the future and is thus an instance of reanalysis. If the form were intended to refer to the past, it would make more sense to categorize it as an instance of analogy where a 3rd person *-s* suffix is attached to a past tense form in analogy to present tense 3rd person *-s* marking.

Another related process is backformation and involves the shortening of the base word by "deleting a suffix (or supposed suffix)" (Plag 2003: 37), such as in the much quoted case of the verb *edit* which is derived from *editor*. Whereas some researchers (e.g., Adams 2001) see backformation as a subcategory of reanalysis, since the new word is reanalyzed as the base form of the word, others (e.g., Plag 2003) emphasize the proportional analogy to

other word pairs: *editor : edit = actor : act*. Two examples of backformation in our data show the same underlying processes, i.e. they are obviously perceived as analogous to word pairs such as *evaluation : evaluate*. These PVCs are ***devaluated*** as from *devaluation* and ***examinates*** as from *examination*.

One further category established in the analysis of our PVCs is that of blends, a group of words subsumed under the general heading 'derivations without affixation' by Plag (2003). While the literature offers a variety of explanations for this class of complex words, definitions generally converge on that "blends are words that combine two (rarely three or more) words into one, deleting material from one or both of the source words" (Plag 2003: 122). Following this definition, we could identify the following PVCs as blends: ***econometric*** (economic + metric), ***flexicurity*** (flexibility + security), ***ranglish*** (Russian + English), and ***webmail*** (web, email). *Webmail*, to start with the last, can be found several times in VOICE and actually turns out to be a rather common word at least among technically versed people. The other blends identified, however, are used in highly specialized settings and are, particularly in the case of *flexicurity*, discussed as specialist topics. The likelihood is thus that these blends have not been coined ad hoc by the ELF speakers but were known to them, being experts in their respective fields, before.

* *Addition and reduction*

The last groups of PVCs in the subcorpus are strongly influenced by the fact that VOICE is a corpus of spoken language where lexis and morphology are influenced by aspects of pronunciation. It was stated in section 2 that we do not tag as PVCs pronunciation variations which do not affect the number of syllables of a word. By implication it follows that we tag as PVCs only those pronunciation variations which lead to a new word as they diverge from an existing word by at least one syllable being added or left out. The resulting two categories are 'addition' with words like ***creativitly***, ***adevertisement***, and ***innovatiations*** and 'reduction' exemplified by words like ***manufacters***, ***contination***, *and* ***diversication***. Of course with regard to these two categories the 'root' or base word plays a particularly important role as the addition or reduction of syllables can only be judged on the basis of the root word.

In these two categories, we find many examples which suggest double categorization, as we cannot be certain whether they are the result of a 'mere' pronunciation variation, so to speak, or are due to other word formation processes discussed above, an observation which reinforces our decision to capture pronunciation variations and coinages in one tag (cf. section 2). Thus,

controversity can be categorized as an addition (the additional syllable *it* is inserted into the word *controversy*) and as suffixation (the nominal *-y* suffix in *controversy* is replaced with another nominal suffix, namely *-ity*). The same double categorization, i.e. addition and suffixation, is also true for the words ***opportunality*** and ***pragmatistic***. Similarly, the word ***fragmentated*** can be regarded as an addition (*fragmented* with the inserted syllable *-at-*) or as multiple affixation with a verb being derived from the noun *fragment* via adding the verbal suffix *-ate* (→ *fragmentate*) and then being marked for past tense with *-(e)d*.

A couple of instances of PVCs which can be linked to matters of pronunciation are the result of swapped syllables, consonants or vowels: ***comptetiviness***, ***sotteck*** (for socket), ***prerequities*** and ***unsiternity*** (presumably for uncertainty). Given that the speech captured in VOICE is processed online and in a linear way, the likelihood is that these are slips of the tongue and a direct result of speech processing and production constraints.

4. Discussion of findings: functional motivations in ELF

The analysis of lexical items captured in the <pvc> tag clearly reveals that the boundaries between already existent and new, between special vocabulary and so-called 'normal' words are not simple and clear-cut and may vary between different contexts and speaker constellations. This is the case in accepted and codified ENL varieties and it is also the case in ELF. It is, however, particularly these 'grey' areas between normal and special, between existent and new that bear testimony to the vibrant nature of language in use and the ongoing linguistic change.

Thus whether a particular word is considered 'new' is a question of context and of point of reference. Of course, if you check other dictionaries – whether specialized or historical – the likelihood is that you will eventually find some of these newly coined words. Chances are probably even higher that you will find a new word on the World Wide Web, if you type it into a search engine like *Google*, for example. Naturally, the World Wide Web as a domain open to all users is considerably less regularized and normative than dictionaries. In contrast to dictionaries, it is constantly in flux, changes by the second, and it is also a place (or space) where different uses and usages of English – as a lingua franca as well as a native, second and foreign language – increasingly merge and, in many cases, become indistinguishable if authorship is not explicitly acknowledged.

In light of this, one might say that the PVC is not really novel then if it can be found somewhere else, i.e. if others have coined and used it before. Yet,

with the exception of specialized terminology, the PVCs in VOICE are presumably coined online and ad hoc and in this sense novel. As has been illustrated in section 3, most of the lexical innovations captured in the <pvc> tag are not erratic, irrational or unmotivated but follow well attested word-formation processes and in this respect represent a continuation of the long-standing history in the natural development of languages. Considering the general regularity of the processes observed as well as their 'motivatedness' (to fill a 'permanent lexical gap' ourselves), it is not surprising that words like ***increasement*** or ***devotedness*** are not only coined by ELF speakers in VOICE but also by others following the same natural route. In fact, it would be surprising if no one else followed this line and thus arrived at the same coinage.

It is at least partly because of this underlying structure and naturalness, we would suggest, that the PVCs in VOICE seem to work. As Cornbleet and Carter (2001: 64) point out, "inventiveness can only communicate if it's understandable" and creating new words along the lines of well established processes seems to be conducive to understanding. But what is more is that such inventions normally happen for a reason in an interaction, in ENL as that is what Carter refers to, but also in ELF:

> *[...] speakers clearly find the interaction sufficiently supportive and co-productive to allow the invention not only to be accepted but also to be seen to be necessary and motivated. (Carter 2004: 98)*

Indeed, this is the case with the PVCs we discussed in section 3. There is a strong sense that the speakers in VOICE do find interactions supportive and co-productive and the newly coined words necessary and motivated. The coining of new words thus appears to be not only a feature of ELF which is "non-disturbing", a categorization made by Björkman (2008), but which is effective and also functional. The analysis of PVCs in the subcorpus of VOICE clearly reveals that the new words serve a particular purpose on a pragmatic-functional level and that their surface forms are related to underlying functions. Establishing this connection between surface forms and underlying functions is indeed one of the key objects of ELF research:

> *So the crucial challenge has been to move from the surface description of particular features, however interesting they may be in themselves, to an explanation of the underlying significance of the forms, to ask what work they do, what functions they are symptomatic of. (Seidlhofer forthc. b)*

The underlying functional motivations that we identified include:

- Increasing clarity
- Economy of expression

- Regularization
- Filling lexical gaps

The first of these functions is illustrated by PVCs such as ***increasement***. In adding a nominal suffix to a word that is already a noun, the speaker avoids potential ambiguity (*increase* could be used as noun and verb) and reinforces the word class. The motivation for such overt word-class marking is **increasing clarity**, a function which has also been attested by other ELF researchers as a general tendency in ELF. Seidlhofer (2005) mentions increasing clarity in relation to ELF speakers adding prepositions (e.g. *discuss about*) or nouns (e.g. *how long time*), and points out that this increasing of clarity is accompanied by adding redundancy. The nominal suffix in the coinage ***increasement*** is, in fact, redundant. Similarly, Dewey (2007b) mentions "explicitness and clarity of proposition" as one of the underlying processes that stimulate innovation in ELF. Dewey identifies this function primarily with regard to repetition, synonymy and rephrasing, features which are prominent in his data and have also been shown to occur frequently in other ELF interactions (cf. Lichtkoppler 2007). Repetition, rephrasing, adding redundancy and 'overt word-class marking' are thus all ELF characteristics prompted by the general functional motivation of increasing clarity. This, in turn, emphasizes the cooperative and listener-oriented nature of ELF talk well documented in other ELF studies (e.g., Jenkins 2000 and Cogo 2007 on accommodation, Pitzl 2005 on joint negotiation work, and Kordon 2006 on phatic communion).

The second functional motivation for lexical innovations in ELF we term **economy of expression**. It is exemplified by words such as ***pre-thesis*** which are coined in order to express concepts or ideas in a concise way rather than using many words and producing long explanations. Whereas the first function, increasing clarity, may (sometimes) be accompanied by adding redundancy, this second function rather points towards reducing redundancy.[9] Although these two trends might seem to be in contradiction to each other at first glance, this contradiction is only superficial. While both trends operate in ELF, they do so at different times in an interaction. Redundancy may be added when this is perceived useful for the sake of explicitness and clarity, and it may be exploited when this seems to enhance efficiency.

Moreover, the observed tendency towards economy of expression also goes in line with a strong emphasis on the semantic properties inherent in linguistic elements. Again, this finding is corroborated by other ELF

9 See Breiteneder (2005 and forthc.) on exploiting redundancy with regard to the 3rd person *-s*.

researchers, like Seidlhofer, who points out that ELF speakers draw on "what is semantically encoded in the grammar and lexis of the language" (Seidlhofer forthc. a). The trend towards compositionality and reliance on semantic properties in ELF is also pointed out by Pitzl (forthc.) with regard to the use of idioms and metaphors. Breiteneder (forthc.) documents the tendency to focus on semantic meaning in the ELF speakers' usage of the third person *-s*. The present study thus again highlights that this focus on semantic properties does not only pertain to words but also operates with regard to morphological elements such as affixes. These are employed and combined with words or a word base to express more complex ideas in an explicit but also economical way.

The third of the functional motivations that we were able to identify with regard to lexical innovations in ELF is that of **regularization**. Both in the context of reanalysis and the context of analogy we noted several coinages which seem to be motivated by this general tendency towards regularization. ***Medias***, ***criterias***, ***unformal***, ***thinked***, ***teached***, ***advices*** and ***knowledges***, to mention just a couple of examples of the PVCs discussed under these headings, can all be seen as the result of a process of regularization which, in fact, affects 'irregular forms' or what Trudgill (1999: 125) would term "grammatical idiosyncracies of Standard English". By using ***teached*** instead of *taught* or ***unformal*** instead of *informal* the ELF speakers create a more regular and probably also less ambiguous system: while the affix *in-* refers both to "not; without" as well as to "in, into" (Quinion 2008) and is therefore ambiguous, the affix *un-* , as employed in ***unformal***, seems to be more straightforward in only referring to the negative. Once again, our findings confirm discussions of the process of regularization in other ELF studies, such as Breiteneder (forthc.) and Dewey (2007a), which highlight the importance of the language contact situation, intrinsic to ELF discourse, in this process of regularization and point out that some of the processes observed in the present paper are entirely to be expected considering the multilingual nature of ELF situations.

The fourth function of **filling lexical gaps** is clearly the one most relevant only to the level of lexis and includes the filling of both momentary as well as permanent gaps (cf. Clark 1994). Coinages that are used to fill momentary lexical gaps "can help to get us out of a communicative jam", to use Crystal's words (1998: 31): "When a word is on the tip of the tongue, and despite our best efforts we cannot recall it, an invented word can get our meaning across". These coinages that fill momentary gaps are, however, "typically repaired – replaced by the correct word – as soon as the speaker can do so" (Clark 1994: 785). ***Examinates***, as discussed in section 3 under the heading of

backformation, nicely illustrates this function of filling momentary lexical gaps in so far as S11 first coins the word ***examinates*** but immediately replaces the invented word by the codified item *examines*.

Extract 8: POwgd14, S11=Danish
468 S11: <pvc> **examinates** {examines} </pvc> them (.) examines them (.) and sort of conclude perhaps (.) er

Carter (2004: 98) refers to this kind of coinages as "survival words" which speakers invent "as a kind of survival mechanism to ensure that the conversation continues to flow". Coining new words is part and parcel of any spoken language interaction and skilful communicative 'survival' via coinages is thus also found in ELF. In addition to filling momentary gaps, some PVCs discussed in section 3 are also prompted by 'permanent gaps', which are permanent in so far as there is as yet no codified word available to express a certain idea or concept. In section 3, we highlighted that ***forbiddenness***, for example, is coined to express the notion of 'the state of being forbidden'. The ELF speaker thus expands the language and fills a 'permanent gap' by coining a new word for a particular occasion.

While some of the coinages presented in our discussion might only be used to fulfil the particular purpose of the particular situation they were coined in and might be forgotten afterwards, there are others which are taken up by co-speakers and might over time become established as a 'regular' and 'normal' word. As pointed out above, the internal regularity of the structure of most of the items captured in our <pvc> tag might be conducive to the 'survival' of some of our PVCs – on the one hand, because they are understandable and on the other hand, because there might be other (ELF) speakers coining the same word along the same lines. In any case one should accept coinages "for what they are", namely "the inevitable consequence of the transplantation of English to new communicative settings and its appropriation by new users" (Ferguson 2006: 173). And while we do not suggest that the examples we discussed should end up in dictionaries, we also cannot rule out that one or two of them might actually develop in this direction.

References

Adams, Valerie. 2001. *Complex words in English*. Harlow: Longman.

Bauer, Laurie. 1988. *Introducing linguistic morphology*. Edinburgh: Edinburgh University Press.

Biermeier, Thomas. 2008. *Word-formation in New Englishes. A corpus-based analysis*. Berlin: LIT Verlag.

Björkman, Beyza. 2008. "'So where are we?' Spoken lingua franca English at a technical university in Sweden". *English Today* 24/2, 35-41.

Böhringer, Heike. 2007. *The sound of silence: silent and filled pauses in English as a Lingua Franca business interaction*. Unpublished MA thesis, University of Vienna.

Breiteneder, Angelika. 2005. "The naturalness of English as a European lingua franca: the case of the 'third person -s'". *VIEWZ* 14/2, 3-26. http://www.univie.ac.at/Anglistik/Views0502ALL.pdf (16 December 2008).

Breiteneder, Angelika. forthc. "English as a Lingua Franca in Europe: an empirical perspective". In Berns, Margie; Seidlhofer, Barbara (eds.). *Symposium 'Perspectives on Lingua Franca' in World Englishes* 28(2).

Breiteneder, Angelika; Pitzl, Marie-Luise; Majewski, Stefan; Klimpfinger, Theresa. 2006. "VOICE recording – methodological challenges in the compilation of a corpus of spoken ELF". *Nordic Journal of English Studies* 5/2, 161-188. http://hdl.handle.net/2077/3153 (16 December 2008).

Carter, Ronald. 2004. *Language and creativity. The art of common talk*. Milton Park: Routledge.

Clark, E. V. 1994. "Creativity in language use". In Asher, R. E. (ed.). *The encyclopaedia of language and linguistics (Vol. 2)*. Oxford: Pergamon Press, 784-785.

Cogo, Alessia. 2007. *Intercultural communication in English as a Lingua Franca: a case study*. Unpublished PhD thesis, King's College London.

Cook, Vivian. 2002. "Background to the L2 user". In Cook, Vivian (ed.). *Portraits of the L2 user*. Clevedon: Multilingual Matters, 1-31.

Cornbleet, Sandra; Carter, Ronald. 2001. *The language of speech and writing*. London: Routledge.

Crystal, David. 1998. *Language play*. London: Penguin.

Dewey, Martin. 2007a. *English as a Lingua Franca: an empirical study of innovation in lexis and grammar*. Unpublished PhD thesis, King's College London.

Dewey, Martin. 2007b. "English as a Lingua Franca and globalization: an interconnected perspective". *International Journal of Applied Linguistics* 17/3, 332-354.

Ferguson, Gibson. 2006. *Language planning in education*. Edinburgh: Edinburgh University Press.

Hülmbauer, Cornelia. 2007. "'You moved, aren't?' – The relationship between lexicogrammatical correctness and communicative effectiveness in English as a Lingua Franca". *VIEWZ* 16/2, 3-35. http://www.univie.ac.at/Anglistik/Views_0702.pdf (16 December 2008).

Hülmbauer, Cornelia. forthc. "'It's so horrible... I mean, I love it!' – The shifting relationship of correctness and effectiveness in ELF communication". In Mauranen, Anna;

Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Press.

Jenkins, Jennifer. 2000. *The phonology of English as an international language*. Oxford: Oxford University Press.

Kastovsky, Dieter. forthc. "English word-formation, combining forms and neo-classical compounds: a reassessment". In *Proceedings of the 18th International World Congress of Linguists, Seoul, July 21 - 26, 2008*. Seoul.

Klimpfinger, Theresa. 2007. "'Mind you sometimes you have to mix' – the role of code-switching in English as a Lingua Franca". *VIEWZ* 16/2, 36-61. http://www.univie.ac.at/Anglistik/Views_0702.pdf (16 December 2008).

Klimpfinger, Theresa. forthc. "'She's mixing the two languages together' – forms and functions of code-switching in ELF". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Press.

Kordon, Kathrin. 2006. "'You are very good' — establishing rapport in English as a Lingua Franca: the case of agreement tokens". *VIEWZ* 15/2, 58-82. http://www.univie.ac.at/Anglistik/views06_2.pdf (16 December 2008).

Lichtkoppler, Julia. 2007. "'Male. Male.' - 'Male?' - 'The sex is male.' The role of repetition in English as a Lingua Franca conversations". *VIEWZ* 16/1, 39-65. http://www.univie.ac.at/Anglistik/views_0701.PDF (16 December 2008).

OALD7 = *Oxford Advanced Learner's Dictionary of Current English*. 2005. 7th edition. Oxford: Oxford University Press.

Pitzl, Marie-Luise. 2005. "Non-understanding in English as a Lingua Franca: examples from a business context". *VIEWZ* 14/2, 50-71. http://www.univie.ac.at/Anglistik/Views0502mlp.pdf (16 December 2008).

Pitzl, Marie-Luise. forthc. "'We should not wake up any dogs': idiom and metaphor in ELF". In Mauranen, Anna; Ranta, Elina (eds.). *English as a Lingua Franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Press.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.

Quinion, Michael. 2008. *Affixes: the building blocks of English*. http://www.affixes.org/index.html (13 November 2008). Based on Quinion, Michael. 2002. *Ologies and Isms: word beginnings and endings.* Oxford: Oxford University Press.

Ranta, Elina. 2006. "The 'attractive' progressive – why use the -ing form in English as a Lingua Franca?". *Nordic Journal of English Studies* 5/2, 95-116. http://gupea.ub.gu.se/dspace/handle/2077/3150 (16 December 2008).

Schendl, Herbert. 2001. *Historical linguistics*. Oxford: Oxford University Press.

Seidlhofer, Barbara. 2001. "Closing a conceptual gap: the case for a description of English as a Lingua Franca". *International Journal of Applied Linguistics* 11, 133-158.

Seidlhofer, Barbara. 2005. "English as a Lingua Franca". In *Oxford advanced learner's dictionary of current English*. 7th edition. Oxford: Oxford University Press, R 92.

Seidlhofer, Barbara. forthc. a. "Accommodation and the Idiom Principle in English as a Lingua Franca". *Journal of Intercultural Pragmatics*.

Seidlhofer, Barbara. forthc. b. "Common ground and different realities: World Englishes and English as a Lingua Franca". In Berns, Margie; Seidlhofer, Barbara (eds.). *Symposium 'Perspectives on Lingua Franca' in World Englishes* 28(2).

Seidlhofer, Barbara; Breiteneder, Angelika; Pitzl, Marie-Luise. 2006. "English as a Lingua Franca in Europe". *Annual Review of Applied Linguistics* 26, 1-34.

Seidlhofer, Barbara; Widdowson, H. G. 2007. "Idiomatic variation and change in English. The idiom principle and its realizations". In Smit, Ute; Dollinger, Stefan; Hüttner, Julia; Kaltenböck, Gunther; Lutzky, Ursula (eds.). *Tracing English through time. Explorations in language variation*. (Festschrift for Herbert Schendl, Austrian Studies in English vol. 95). Wien: Braumüller, 359-374.

Trask, R. L. 1995. *A dictionary of grammatical terms in linguistics*. London: Routledge.

Trudgill, Peter. 1999. "Standard English: what it isn't". In Bex, Tony; Watts, R. J. (eds.). *Standard English. The widening debate*. London: Routledge, 177-128.

VOICE Project website: http://www.univie.ac.at/voice (13 November 2008).

VOICE Project. 2007a. "Mark-up conventions". VOICE Transcription Conventions [2.1]. http://www.univie.ac.at/voice/documents/VOICE_markup_conventions_v2-1.pdf (16 December 2008).

VOICE Project. 2007b. "Spelling conventions". VOICE Transcription Conventions [2.1]. http://www.univie.ac.at/voice/documents/VOICE_spelling_conventions_v2-1.pdf (16 December 2008).

VOICE. forthc. *The Vienna-Oxford International Corpus of English (version 1.0 online)*. Director: Barbara Seidlhofer; Researchers: Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski, Marie-Luise Pitzl.

Widdowson, H. G. 2004. *Text, context, pretext. Critical issues in discourse analysis*. Oxford: Blackwell.

Learner Corpora of English and German: What is their status quo and where are they headed?

*Barbara Schiftner, Vienna**

> We should enrich the research community with the expertise we have gained from previous projects and should encourage one another not to jump on the bandwagon of corpus-based research without sufficient knowledge of corpus building. (Tono 2003: 806)

1. Introduction

Learner corpora have become an increasingly prominent tool in areas of applied linguistics which deal with language learning and language teaching. When carefully constructed, they can serve as valuable tools for gaining new insights into the way foreign language learners use a language at various stages of proficiency, or for accounting for provisional hypotheses with examples from genuine learner language. The fact that learner corpora have gained more ground in the research community is reflected by the growing number of projects that are involved with the compilation of learner corpora as well as in the steady growth and refinement of existing learner corpora.

As the title states, this article is concerned with German and English learner corpora. This selection is due to my strong interest in both English and German language teaching and learning. In the research that I conducted into learner corpora of both of these languages, I found that learner corpus linguistics is not very prominent in German linguistics, and that available

* The author's email for correspondence: barbara.schiftner@univie.ac.at.

resources in the field thus differ considerably between English and German (cf. the list of learner corpora in the appendix).

Possibly due to the rapid development of the field, the documentation available on individual projects in both areas is scattered and often scarce. Thus, in an attempt to shed light on the development of English and German learner corpus projects, the objective of this paper is threefold:

- to present an overview of current developments in English and German learner corpus compilation
- to point out problematic issues regarding learner corpus design and accessibility
- to reflect on current developments and indicate ways forward for both English and German learner corpus compilation

In this discussion, which is based on a study I conducted in 2007,[1] I will refer to both English and German learner corpora; more precisely to corpora of the written production by learners of English and German. While differences between English and German written learner corpora will certainly be addressed, the main focus is not on a comparison of the two fields, but rather on the discussion of different approaches to the design and compilation of learner corpora. In order to illustrate the development of the field, I will point out approaches with innovative potential and address neglected areas which call for more attention in the future.[2]

2. Defining Learner Corpora

Before turning to the survey of English and German learner corpora, let us specify exactly what we are referring to by the term *learner corpus*. Sylviane Granger, project director of the *International Corpus of Learner English* (*ICLE*), defines computer learner corpora as

> *[...] electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance. (Granger 2002: 7)*

1 This study was part of my M.A. thesis (Schiftner 2007), which was written at the Department of English at the University of Vienna under the supervision of Prof. Dr. Barbara Seidlhofer.

2 Even though this study is based solely on English and German learner corpora, some of the suggestions made might well be relevant for learner corpora compiled for other target languages. Due to the selection the study is based on, however, no claims can be made in this regard.

This definition nicely frames what a learner corpus essentially is, or should be: a structured, well documented collection of texts produced by learners of a language. Learner corpora are often compiled for a particular purpose, which is reflected in the text type collected, the level of education of the respective learners, and other relevant variables, as well as the way in which the texts are archived. These aspects will be discussed in more detail in section 3.

Granger's definition, however, also raises two rather critical issues. One of them is the question of what is meant by foreign or second language textual data in this context. The differentiation between a learner and a user of a language can certainly be controversial, as for example in instances where a language is learnt and used as an official second language or learnt as a foreign language and used as a lingua franca. Thus, the delimitation of learner corpora from native or lingua franca corpora is not always clear cut.[3] Generally, learner corpora are compilations of texts produced in an educational setting. Thus, the term *learner* in this context refers to the institutional setting rather than to more general characteristics of a language learner.

In view of the discussion of authenticity in native speaker corpora (cf. for example Widdowson 2000, 2003; Kaltenböck & Mehlmauer 2005), the notion of *authentic* foreign or second language data presented in Granger's definition also needs to be addressed. In the case of learner language the notion of authenticity is especially problematic. Whether the language a learner produces when prompted to do so by a task can be called *authentic* even at the time of production largely depends on whether or not the learner can appropriate the task to his or her own reality. The extent to which a task is appropriated by a learner, i.e. the extent to which the task is authenticated, may vary, leading to varying degrees of authenticity in the process of text production.[4] Following Widdowson's differentiation of the terminology, corpora generally do not include *authentic discourse*, but *genuine texts*, since the communicative context in which the language is produced does not travel with the text (cf. Widdowson 1980: 165-166). While learner corpora usually comprise detailed information on the provenance of the learner texts (cf. section 3.4), the texts are nonetheless stripped of their context and can consequently not be referred to as *authentic*, but rather as *genuine* learner texts.

---

3 Cf. for example Nesselhauf's take on more or less typical learner corpora (2004: 128).

4 For more extensive discussions of the notion of authenticity in language teaching and learning, see for example Breen (1985), Taylor (1994), and Widdowson (1990, 2003). For a brief comment on the naturalness of learner texts cf. also Nesselhauf (2004: 127-128).

For the purpose of this study, I have thus defined learner corpora as collections of genuine learner texts which are stored electronically. The texts are encoded in a homogeneous way and linked to information about the circumstances of the text production and about the learners who produced them. The representativeness of a learner corpus is limited to a particular group of learners and largely depends on the research purpose the corpus is intended for.

2.1. Potential and Limitations of Learner Corpora

This article is primarily concerned with a survey of completed learner corpora and ongoing projects in the field. Nonetheless, to justify such a survey, a brief reference to the use and possible misuse of these corpora seems expedient. Since many others have discussed the potential and limitations in using learner corpora before, I will refer the interested reader to the respective publications.

As Sylviane Granger points out, learner corpus research provides a "focus on performance (rather than competence), description (rather than universals) and quantitative as well as qualitative analysis" (1998: 3). Learner corpora are thus a powerful resource for the analysis of learner language, especially in that they do not simply reduce learner performance to possible errors and misuse, but provide for the possibility to describe learner language in its own right. Due to their systematic design, learner corpora allow for analyses of learner language with respect to various factors recorded in the corpus, such as the learners' L1, their knowledge of other foreign languages or the task setting (cf. Nesselhauf 2004; Granger 1998, 2009).

Clearly, learner corpora provide a valuable resource and significant input for the analysis of learner language. Nonetheless, learner corpus research is not a panacea but a technique, and as such also has limitations. Leech (1998) provides a comprehensive overview of these, including

- the tediousness of corpus collection,
- the overrepresentation of written production in corpora,
- sampling and the issue of representativeness of a corpus,
- the need for annotation if anything but a specific orthographic representation of a word is to be analysed,
- the focus on the analysis of language produced by a certain group of learners rather than by individuals, and
- the potentially prescriptive use of native speaker reference corpora. (cf. Leech 1998: xxi-xix)

While some of these limitations relate to learner corpus analysis, which is not the focus of this study, issues such as sampling, mode of production, annotation, and the process of corpus compilation are relevant in designing a corpus. These aspects relating to corpus design are of particular significance in the following discussion of learner corpus projects.

3. Learner Corpora of English and German – The Status Quo

Learner corpora have been surveyed before, as for example by Norma Pravec (2002), who conducted an extensive study of English learner corpora, and Yukio Tono (2003), who summarizes what he calls "major learner corpus projects" of English. No such survey exists for German learner corpora, however.

Not surprisingly, English learner corpora have grown both in number and in size since the surveys mentioned above were conducted. For the present study, I have updated my 2007 survey of German and English learner corpora (cf. Schiftner 2007), which not only differs from the abovementioned studies in that it includes a survey of German learner corpora, but also in that it comprises a considerably larger number of English learner corpora. A list of the 26 English and five German learner corpora collated in this updated survey, on which the following sections of this article are based, can be found in the appendix.[5]

The information I was able to gather on the individual corpora varies considerably both in extent and type. Thus, for some corpora technical and methodological information could be compiled, while for others the information that could be assembled is rather limited. The very fact that documentation greatly varies illustrates the difficulty in gaining a comprehensive overview of the state of affairs in learner corpus research.

3.1. Amount of data

In the 26 corpora of learner English described here, we find a range from 16,500 words in the *Learner Journals* Corpus to 30,000,000 words in the *Hong Kong University of Science and Technology (HKUST)* Learner Corpus. The variation in the five corpora of German learner language ranges from 30

[5] I would like to keep this survey updated and as comprehensive as possible. Should you be aware of a learner corpus of English or German that I do not mention, please contact me at barbara.schiftner@univie.ac.at.

texts in *LeKo* (short for *Lernerkorpus*) to 1,200,000 words in the *Telecollaborative Learner Corpus of English and German* (*TELEKORP*).[6]

These differences in size might well be accredited to the very different purposes for which the corpora were collected. *LeKo*, for example, was collected in a seminar at HU Berlin with the aim of finding a system of classification for learner errors. For this specific purpose, the small set of 30 texts may well have been adequate. At the same time it is certainly an atypical learner corpus, which has been compiled for a single purpose. For collections such as the *Cambridge Learner Corpus*, which is compiled for a broad analysis of learner language at different levels of proficiency, focusing on learners with various L1 backgrounds, a large collection is indispensable.

Apart from the fact that the research purpose certainly influences the data collection, the defining features of a learner corpus should nevertheless be met. Even when using a very specific definition of learner corpora such as Sylviane Granger's above, it is debatable whether *LeKo* can be called a corpus. If the criterion of representativeness in McEnery & Wilson's more general definition of corpora (2001: 29) is taken into consideration, however, *LeKo* does certainly not qualify as a corpus (cf. also Leech 1998: xix on the issue of representativeness of learner corpora).

Besides the kind of purpose pursued, a rather pragmatic reason for the differences in size are the resources available for the compilation of a corpus. With the exception of the *HKUST* corpus compiled in Hong Kong (30,000,000 words), non-commercial learner corpora normally do not reach the size of commercial corpora such as the *Longman Learner Corpus* (10,000,000 words) or the *Cambridge Learner Corpus* (25,000,000 words), but have a word count of 1 million words or less.

3.2. Type of data and text type

As mentioned in the introduction, this study is concerned solely with corpora of written learner language. The majority of learner corpora comprised in the survey are collections of cross-sectional data. More precisely, only about 25% of the 26 corpora of written learner English explicitly include longitudinal data, i.e. texts written by the same learners over a longer period of time. Among these longitudinal corpora, two (*LANCAWE*, *TELEKORP*) include texts collected over a period of between one and three months, while for two others (*SILS*, *USE*) a collection period of several semesters could be ascertained. For both the *Longitudinal Database of Learner English*

6 This amount of data in Telekorp does, however, include both English and German learner language.

(*LONGDALE*) and the *Database of English Learner Texts* (*DELT*) the data collection is planned for several semesters or even years; both were launched in 2008. Interestingly, all three large-scale German learner corpora are at least partially longitudinal.

The text type collected is not always clearly stated in the corpus descriptions. Using a broad definition of the terms, one can detect a majority of argumentative and expository essays. However, other text types are also collected, as for example business correspondence in the *Learner Corpus of English for Business Communication* or summaries in the German *Fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache*[7] (*FALKO*). As regards the task setting, some corpora, such as the *Cambridge Learner Corpus*, *HKUST*, or the *Québec Learner Corpus* comprise texts written as exams in a timed setting, while most corpora seem to comprise untimed written production.

The (average) length of the texts, which is certainly closely connected to the proficiency level and age of the learners (cf. section 3.3), varies considerably between 20 words in the *TELEC Secondary Learner Corpus* (*TSLC*) and 1000 words in the *Uppsala Student English Corpus* (*USE*), while most commonly texts seem to be between 300 and 500 words in length.

As regards the mode of production, *TELEKORP* is certainly exceptional in that it comprises texts produced as computer-mediated communication (CMC), which is, though written, inherently different from conventional written production (cf. David Crystal's discussion of computer-mediated language or *Netspeak* in Crystal 2001).

3.3. The Learners

Probably due to the fact that learner corpus projects are usually based at universities, most of them compile written productions of learners in tertiary education. Exceptions are the *Chinese Learner English Corpus* (*CLEC*), the *Santiago University Learner of English Corpus* (*SULEC*), the *TELEC Secondary Learner Corpus* (*TSLC*), and the *Japanese EFL Learner Corpus* (*JEFLL*), the latter two being the only corpora that also comprise data produced at the lower secondary level.[8]

---

[7] Translated into English this means Error-Annotated Learner Corpus of German as a Foreign Language.

[8] To my knowledge, a project for the compilation of primary and lower secondary international EFL data has recently been launched by Yukio Tono (Tokyo University of Foreign Studies, Japan). Since no further information could be gathered regarding this project, it could not be included in the survey (cf. http://lexicon.tufs.ac.jp/icciwiki/).

The representation of certain first languages in a learner corpus is usually influenced by the location of a learner corpus project. Corpora such as the *International Corpus of Learner English* (*ICLE*), as well as corpora compiled in the countries where the target language is spoken, such as the *Montclair Electronic Language Database* (*MELD*), an English learner corpus in the United States, or *FALKO*, a German learner corpus compiled in Germany, are exceptional in that they include texts written by learners with various first languages. Out of 26 corpora of learner English, 11 are compiled in Asia, 11 in Europe, 3 in North America (1 in Québec, 2 in the USA), and 1 in South America. Interestingly, corpora of learner German in this study are compiled in Europe and the Americas; none could be found in Asian countries. The biggest collections seem to exist for Japanese and Chinese learners of English, while the most common L1 in German learner corpora seems to be English.

3.4. Organization of the Data

Information on how the texts and the task-related and learner-related additional information are organized and stored is especially difficult to obtain. What is clear is that corpora make use of very different individual systems for organizing the data. While the texts are usually saved as ASCII or Unicode text files, the way the additional information is stored ranges from simply employing identification numbers that can be manually associated with the background information to employing relational database systems.

Closely related to the subject of corpus organization is that of the collection of background information. With *background information*, I refer to task-related and learner-related details (cf. Granger 1998, 2002 for a discussion of these details) that can be used to create subcorpora that are homogeneous in terms of certain variables. The amount and detail of the background information collected might well be due to the very different purposes for which the corpora are collected. Especially those large scale projects that compile learner corpora in view of reusability for various research questions should aim at a description of setting, task and learner that is as comprehensive and comprehensible as possible.

One is tempted to think that the technical structure of a corpus reflects the elaborateness of the background information collected. Examples from the study, however, clearly show that this is not necessarily true. Both *ICLE* and *USE*, for example, follow very rigorous descriptions of topics and text types

and make use of very detailed learner profiles.[9] While the *ICLE* data is stored in a relational database, background information in *USE* is documented in an Excel sheet separately from the text files. The relevant files can be identified by querying the Excel sheet and can then be selected from the corpus. Needless to say, it is convenient if subcorpora are retrievable from a database according to a combination of certain variables. Even more important for the traceability of the provenance of the data, however, is the design of the corpus compilation and the information recorded and available, no matter in which format it is stored.

What is most important is that the texts and the additional information are stored in a file format that is compatible with available text analysis software or programs that may be written by the researchers themselves. As Barnbrook stated in 1996, most available

> *text exploration software [...] and most of the programs that you might write yourself assume that the text is in the approximately standardised ASCII (American Standard Code for Information Interchange) format, sometimes referred to as ANSI standard. (Barnbrook 1996: 36)*

Technical possibilities have of course developed since 1996, and text exploration tools such as Wordsmith Tools[10] or AntConc[11] can now handle texts in both plain text and XML format. In the realm of learner corpus linguistics, plain text format is still the most common method of text storage. It is also what (fairly simple) text exploration tools seem to handle best. Since texts saved in the ASCII format do not include information about the texts' structure, such as font size, paragraph indentation, etc., such textual features, insofar as they are relevant for the intended research, have to be encoded using a markup scheme (cf. e.g. Granger 1998: 12; McEnery et al. 2006: 22-28). For most corpora in this study, no information regarding the markup used in the texts could be ascertained. Exceptions are the *SILS* (*School of International Liberal Studies at Waseda*) *Corpus* and *DELT*, both of which employ XML tags to mark features such as quotations and paragraphs.

In his "basic principles" for corpus design, Sinclair argues against the practice of adding such tags into the running text, and claims that any additional information should be stored separately from the texts in a stand-

---

9 The learner profiles used for the ICLE (http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/LPROFILE.htm) and the USE corpus (http://www.engelska.uu.se/use.html#anchor15) are available online.

10 WordSmith Tools, available for purchase at http://www.lexically.net/wordsmith/.

11 AntConc, freely available at http://www.antlab.sci.waseda.ac.jp/antconc_index.html.

off format and "merged when required in applications" (Sinclair 2005: 7). This, however, is certainly not standard in learner corpus compilation. Common practice seems to be to store the plain text as well as multiple versions of the same text interspersed with different annotations.

3.5. Linguistic Annotation

A feature of corpora closely related to markup is annotation. Leech describes corpus annotation as "the practice of adding interpretive linguistic information to a corpus (2005: 21). Linguistic annotation can, for example, include part-of-speech, syntactic, semantic, discourse or error analysis. What is crucial with respect to annotation is that – as defined by Leech above – all linguistic annotation, be it done automatically or manually, is the encoding of a linguistic analysis and as such implies some interpretation of the data (cf. McEnery et al. 2006: 29-45; Leech 2005; Meunier 1998). Raw learner texts, i.e. learner texts that are not annotated, allow only for very limited analyses, not allowing for searches of particular parts of speech, syntactic constituents, etc. Thus, depending on the respective research question, different types of linguistic annotation may be useful and needed (cf. limitations mentioned in section 2.1).

The extent of annotation in the observed learner corpora varies considerably. While some of the corpora are not annotated at all, others include part-of-speech tags, lemma tags, or even error tags. That said, it needs to be added that there is frequently no information on the programs or tagsets used for the annotation available. From the information that could be gathered, however, it can be inferred that the different projects use various programs that utilize differing tagsets.

What has to be considered in this respect is that the different programs not only tag learner data with varying reliability, but also that the format of the tags used differs from one program to the other. Moreover, the respective tagsets vary in size, i.e. they can be smaller and include more generic tags, or they can include a larger number of tags and define individual tokens very specifically. Taking an example from the different tagsets for CLAWS, a part-of-speech tagger frequently used to tag English learner corpora: the C7 tagset, consisting of altogether around 160 tags, includes 31 different verb-tags and differentiates between the use of the base form as a finite or an infinite verb, as opposed to the C5 tagset, which consists of only 60 tags, including 25 verb-tags, but does not include this differentiation. The C8 tagset further differentiates between the lexical and auxiliary use of *be*, *do* and *have* (cf. *CLAWS part-of-speech tagger for English*, http://ucrel.lancs.ac.uk/claws/).

In terms of the annotation that is added to the corpora manually, error annotation seems especially inconsistent. For the annotation of *MELD*, for example, error reconstruction is used, meaning that errors are reconstructed to obtain acceptable sentences (for a detailed description of the process cf. Fitzpatrick & Seegmiller 2004: 4-9). For other projects individual error coding systems were devised. To my knowledge, the only system for which both an editor and a manual including the details of the annotation scheme are available is the one employed for the annotation of the *ICLE*, namely the *UCLEE* (UCL Error Editor), which can be purchased from the Centre for English Corpus Linguistics at the Université catholique de Louvain (cf. Dagneaux et al. 2005).[12]

The annotation of most learner corpora, as for example for the *ICLE*, is done using "flat" or "inline" annotation schemes, i.e. the annotation tags are inserted in the running text and saved in the same file (cf. section 3.4 above). This can be done either in table format, which means that the annotation is directly attached to the tokens, or in a tree model, where SGML/XML tags are inserted at the beginning and the end of the token or sequence of tokens to be annotated (cf. Lüdeling et al. 2005). Anke Lüdeling, coordinator of the German corpus *FALKO*, argues that this annotation scheme does not provide for various levels of annotation, making an efficient search that incorporates various levels of annotation impossible. Especially with regard to error annotation and conflicting interpretations of the same error, Lüdeling favors a stand-off structure of annotation, which means that the annotations are saved in SGML/XML files separately from the texts, making it possible to apply various levels of annotation to the same data (cf. Lüdeling et al. 2005).

McEnery et al. refer to these two different kinds of annotation as "embedded" vs. "standalone" annotation (2006: 44). They point out several advantages of a standalone annotation structure, including that it

- *allows multiple overlapping hierarchies;*
- *allows for alternative annotation schemes to be applied to the same data (e.g. different POS tagsets);*
- *enables new annotation levels to be added without causing problems for existing levels of annotation or search tools;*
- *allows annotation at one level to be changed without affecting other levels. (ibid.)*

[12] For more detailed information on error annotation confer for example Díaz-Negrillo & Fernández-Domíngues (2006); Milton & Chowdhury (1994); Nicholls (2003).

Apart from all these advantages, however, what has to be considered is that common text retrieval or corpus exploration tools such as WordSmith Tools do not support standalone (or stand-off) annotation, but can only be used with texts that employ embedded (or inline) annotation (cf. McEnery et al. 2006: 44). Apart from the technical expertise needed to add annotation in a stand-off format, the fact that a number of well-established and fairly easy-to-use programs cannot be used on data with stand-off annotation is most likely the reason why apart from *FALKO* none of the other projects in this study employ this kind of annotation format.

3.6. Availability and Accessibility

Unfortunately for researchers who might be interested in existing learner corpora either to use them for their own research or to replicate studies that have been undertaken on them, not even half of the learner corpora described are publicly available (cf. Nesselhauf 2004: 133 on the problem of accessibility). Out of 26 English learner corpora in this study, 13 are freely available to other researchers. While in four out of 13 cases, the full texts compiled can be downloaded or bought on CD-ROM, 9 of the available corpora can only be accessed through an online search interface. The only German learner corpus openly available is *FALKO*, which can also be searched online.[13]

Corpora that are available for online-search only, but not as full texts, cannot be fed into text analysis software and thus cannot be used for the creation of wordlists, keyword-lists, the direct comparison with other corpora etc. Hence, the usability of corpora that are available in this form is restricted.

Some learner corpora that are not openly available can be accessed on request; quite a few corpora, however, are not available to a wider public at all, but merely to the department or institution where they are compiled. Findings that are based on data that is not available to other researchers can hardly be approved or argued against. The *Cambridge Learner Corpus*, for example, is only available to linguists working for Cambridge University Press, which implies that those studies that impact Cambridge teaching materials cannot be replicated by independent researchers.

[13] Note that some of the corpora have just been launched or are still under construction and might well be made available at a later stage (e.g. *SILS, SULEC, LONGDALE, DELT*).

4. Implications for the Future Development of Learner Corpora of German and English

The survey has shown that the existing corpora of both English and German learner language differ in many respects. In the following, I will further discuss this heterogeneity and point out possibilities for future development and potential improvement.

4.1. Suggestions for the Type of Data Compiled

With regard to learner corpus design and compilation, aspects to be considered include the type of data collected, the size of the corpus, and the organization of the texts in a database. In all these specifics, we can find considerable differences between the corpora described in this study.

Interestingly, almost all of the corpora in this study, which are taken to constitute a representative cross-section of available learner corpora, are made up of texts produced by students at the tertiary level. Only four of the corpora (all of learner English) include texts from learners at the secondary level. This reflects the observation made by Barlow (2005: 357) that

> *[m]ost of the existing learner corpora are based on the writing of fairly advanced language learners. In order to play a central role in understanding SLA a wider range of learner corpora, including spoken learner corpora, will have to be created.*

Since the progression of competence and performance in a foreign language might vary considerably between different age groups, this seems an aspect important to consider in studies of second language acquisition with the aid of learner corpora.[14]

As mentioned in section 3.2 above, longitudinal learner corpora are still rather scarcely represented. Even though some projects compile texts in a longitudinal or quasi-longitudinal manner, there is certainly a backlog in comparable longitudinal data that covers not only a longer period of time but also younger and less proficient learners. In Europe, English is taught from a very early age – a fact that is certainly not reflected in the available learner corpora. With the exception of *SULEC*, none of the corpora that include texts written at the secondary level are European projects.

---

[14] Besides their relevance for SLA research (as claimed by Barlow 2005 & Myles 2008), spoken learner corpora are not the primary subject of this paper. SULEC, however, contains both spoken and written data.

In the long run, the compilation of longitudinal or quasi-longitudinal data would certainly be a welcome addition to the field. One way of meeting this necessity would be to start a comprehensive project that incorporates texts by learners at different levels of proficiency and from different age groups. Since such a project would involve several different institutions, organizing a joint collection with a large number of contributors might quickly become unmanageable and certainly requires substantial central funding. Another possibility would be to establish ways of cooperation between independent, possibly already existing, projects. Such a networked approach, however, is clearly only possible on the basis of certain shared standards for collection and encoding that have not yet been established.[15]

4.2. Towards Standardization: Markup and Linguistic Annotation

As Sylviane Granger points out in her definition quoted in section 2, learner corpora are supposed to be encoded in a standardized way. As a close look at different learner corpus projects will reveal and the sections above illustrate, it is difficult to make out what this standard is or should be. Undoubtedly, learner corpora are compiled according to multiple different "standards" that fit the needs of the respective projects.

Proper markup is one characteristic that differentiates arbitrary collections of texts from corpora and its importance should not be underestimated. The survey revealed that from the available documentation, the exact variables encoded in learner corpora, and the way those variables are saved, are not always clear. Undoubtedly, making this information available is a necessary step towards improving compatibility between individual projects.

As discussed in section 3.4 above, a certain standard of markup can be very useful, not only for the encoding of the texts with contextual information, but also for the analysis of certain features of the texts such as paragraphing or the use of quotations in academic writing. In order to make mutual readability of such markup possible not only for linguistic research, but in all fields involving the encoding of text, standards for the markup of texts have been developed. One institution which provides guidelines for standardized markup is the *Text Encoding Initiative* (*TEI*) (cf. http://www.tei-c.org/). The *TEI* guidelines also serve as a basis for the *Corpus Encoding Standard* (*CES*) (cf. http://www.cs.vassar.edu/CES/). These standards,

[15] Cf. section 4.3 on the need for more collaboration.

however, have not been drawn up specifically for learner corpora and have to my knowledge not been used in any of the learner corpus projects included in my survey.

The contextual information compiled for learner corpora depends very much on the individual projects, making it difficult to draw up a common standard. In order to ensure the usefulness of the data to other researchers, guidelines for a minimal standard of header information could nonetheless be developed. These could possibly be based on the *TEI* guidelines, but should be tailor-made for learner corpus projects. Even if different variables are recorded in each project, such guidelines would ensure that the same markup is used to encode the information. The guidelines should be open to the addition of new codes, which would have to be recorded in the user documentation. Thus a minimum of description and comparability could be ensured, even if texts are saved in different database programs which include more or less detailed information about each learner and text.

As long as no standards or guidelines are available, it seems a good solution to save markup in a format which can be manipulated relatively easily to match other standards if so desired, as is the case with XML encoding. It is imperative that an account of the conventions used is included in the corpus documentation. However, my survey of learner corpora has made clear that a detailed account of the markup used is rarely available.

The definition of standards or guidelines for linguistic annotation is inherently more difficult to achieve. As discussed in section 3.5, various kinds of linguistic annotation can be added to the texts, the most common types of annotation in learner corpora being lemmatization, part-of-speech tagging and error tagging. As the survey shows, however, several different annotation schemes and programs are used. Needless to say, texts tagged with different programs and/or tagsets cannot easily be compared, let alone grouped together in the same corpus. Therefore, the question of annotation – just like the question of general markup – is obviously an important one when cooperation between different projects comes into play. Even in considering the comparability of different studies, the linguistic annotation of the respective learner corpora may be a critical factor.

A rather controversial type of linguistic annotation is error annotation. As is the case for all types of linguistic annotation, the classification of errors largely depends on the underlying theory. While error identification as such is already a tricky business, some error annotation schemes also include rather interpretative tags classifying the source of an error, such as tags indicating that an error is based on L1 influence (cf. Dagneaux et al. 2005: 10). Clearly, it is close to impossible to devise an error tagset that is acceptable and equally

useful to all researchers working with learner corpora. Nevertheless, it might be worth considering how a set of standard codes for error categories could be devised that can be adapted to different theories and projects (cf. Tono 2003: 801-802).

Another critical issue in error annotation is that, apart from the incompatibility of different annotation schemes, "[t]here are often cases where there is insufficient evidence to assign one unambiguous interpretation of an error. Thus [...] tagging schemes which allow for alternative possibilities in terms of target forms" (Tono 2003: 804) need to be developed. One example for such a multi-layer error annotation scheme is the annotation of *FALKO* (cf. Lüdeling et al. 2005, and section 3.5 in this paper). While this certainly bears considerable advantages compared to the more common inline annotation models, it poses a serious problem to those projects that do not have computer experts at their disposal. Apart from that, the utilization of complex multi-layer structures that are not available for the majority of corpus projects does not necessarily aid the compatibility of learner corpora.

As Tono suggested in 2003, there is certainly a need for making the various existing tools available to other researchers in order to "facilitate the standardization of corpus annotation in the future." (2003: 804) This is not only true for the tools used, but also for the respective tagsets and tagging manuals. Sad to say, there does not seem to have been much movement towards standardization since.

4.3. Accessibility of Information and Data and the Need for a Networked Environment

As has already become clear in the preceding sections, the need for multiple collaboration cannot be denied: between different corpus projects, in a networked environment within the learner corpus community, but also beyond it. The *ICLE* is clearly an outstanding learner corpus in this respect since it is based on the collaboration of researchers who collect corpus data all around the world. This collaboration is, however, the exception rather than the rule. One of the areas in which a more networked approach is required is corpus collection and the availability of corpora to other researchers. As Fitzpatrick and Seegmiller rightly state,

> *[a] corpus is a large investment in time, money and equipment and the lack of access to corpus data diminishes the advantages that these collections provide. (Fitzpatrick & Seegmiller 2004: 2)*

Surely, no one will deny the fact that the collection of a corpus is a laborious task. Once so much effort has been put into creating a corpus, it seems a

waste not to share it with the greater research community. This possibility can, however, only be ensured if ethical issues are considered during the process of data collection, i.e. if the learners are asked for their permission to have their texts used for research purposes, and if this permission is documented.

A related problem is that of the disclosure of programs and methods used in studies. This is not only important for other researchers to challenge or support a study, but can in fact be essential for researchers that are new to the field of learner corpus research. If information on procedures is not communicated, it is lost to those who are working in the same field, and instead of building on the expertise of others, the same efforts of finding out about methodologies for collecting the data as well as tools and procedures for the analysis are repeated over and over again. Making the corpus data and the know-how available to the linguistic community could greatly aid the collection of new learner corpora as well as collaboration between individual projects.

Availability of information and resources is, however, only one side of the coin. A truly networked environment between linguists working with learner corpora implies regular communication between the organizers of individual projects, which could lead to a more collaborative approach. Data collected for one purpose might still be useful for a range of other analyses, and should thus be compiled in a way that makes it more generally useful (cf. Pravec 2002: 108, Nesselhauf 2004: 127). In order to facilitate collaboration and provide for the reusability of learner corpora, which are created at enormous expense in cost and time, it therefore seems advisable to draw up some guidelines that ensure a comparable and reproducible standard (cf. section 4.2, and Tono 2003: 804).

What needs to be stressed at this point is that a stronger network is not only necessary between individual learner corpus projects, but also between those disciplines that make up learner corpus research, i.e. corpus linguistics, linguistic theory, second language acquisition and foreign language teaching. (cf. Granger 2009 on the inherent interdisciplinarity of learner corpus research). Especially when it comes to the collaboration and mutual transfer of information between learner corpus research and foreign language teaching, there appears to be great potential for the improvement of a networked and collaborative environment. The fact that, thus far, the impact of research into computer learner corpora on language teaching has been rather limited might well be a result of too little collaboration between researchers and teachers.

What I plead for is an effort for more collaboration and a better organization of the exchange of information and data. Initiatives such as the Corpora List (cf. http://gandalf.aksis.uib.no/corpora/) provide a network where corpus-related questions can be discussed with other researchers, and books like McEnery et al. (2006) or Wynne (2005) certainly provide a valuable resource for those who are starting to work on a corpus project. There is, however, a need for information exchange and guidelines specifically tailored to learner corpus research. Even though articles such as those by Pravec (2002) and Tono (2003) are an extremely valuable effort in organizing information on various learner corpora, it becomes clear from their surveys as well as from the survey in this study that in the field of learner corpus research, there is very little common ground in the way the data is collected and organized (cf. Pravec 2002: 108).

It is of course an unrealistic wish that all projects should one day use the same technical equipment and methodologies. However, in order to facilitate the launch of new learner corpora and collaboration among existing projects, and to provide for the possibility of a more networked environment, basic "building blocks" integral to this very specific field of corpus linguistics should be documented and this documentation be made available.[16]

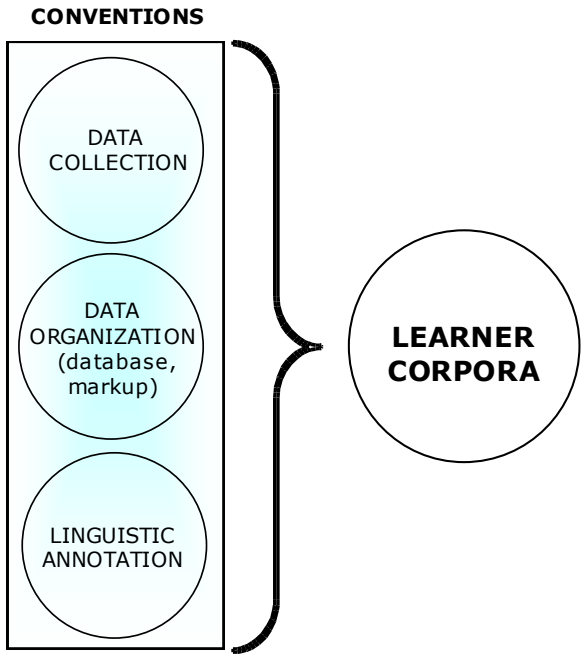

Figure 1: Areas in which guidelines could be drawn up to aid learner corpus collection and facilitate future collaboration

16 As mentioned in section 4.2, the guidelines drawn up by the *Text Encoding Initiative* (*TEI*) can serve as the basis for these conventions.

As illustrated in figure 1, these "building blocks" include procedures for data collection, concepts and standards for data organization, i.e. information about database programs and markup conventions, and guidelines for linguistic annotation. While such a "manual" for learner corpus creation seems an essential prerequisite for future collaboration and more transparency in this field, it is also something that can hardly be drawn up without a collaborative effort of those already working with learner corpora today.

4.4. The issue of German Learner Corpora

Even though most of the aspects discussed thus far apply to both English and German learner corpora alike, it is obvious that the current situation is not at all the same for the two languages. This implies that different conclusions can be drawn regarding their future prospects. Since English learner corpora have, due to their larger number, been given a more prominent position in this paper thus far, this section will focus solely on implications the survey revealed for German learner corpora.

In German linguistics, the collection of learner corpora is a very recent development and has therefore not yet advanced to the same extent as corpora of learner English have. Whether this field of research will grow as rapidly in German linguistics as it has in English linguistics remains to be seen. The compilation of such resources certainly bears great potential for research into German language teaching and learning. German, however, has a very different status from English, which will certainly have an impact on the progress in this fairly new field of research. The use of English as a global lingua franca obviously sets it apart from other *foreign languages* such as German. A large learner community provides both for a setting of widespread research interests and commercial demand for new findings and enhanced materials.

It is obvious from the survey that corpora of learner German are lagging behind corpora of learner English in both number and size. While no large and comprehensive collection of learner German has been compiled, the potential of German learner corpora seems to be the innovative approaches they encompass. Among the corpora discussed, *TELEKORP* is significantly different from other learner corpora due to the type of data it includes; *FALKO* stands out from other learner corpora because of its approach to annotation (cf. section 3.5); and the German part of the *MLC* constitutes a component of a project which facilitates the comparability of texts in different foreign languages produced by learners with the same L1 background as a new research perspective (cf. Tagnin 2006).

Even though these novel approaches have to be valued in their own right, and they certainly provide models of new approaches for other learner corpus projects, they do not lessen the need for larger corpora of learner German. As is apparent from the research that has been conducted with English learner corpora, large collections of texts, such as the *ICLE*, the *Cambridge Learner Corpus*, or the *Longman Learners' Corpus* are rich resources for comprehensive studies of learner language. While smaller learner corpora can reveal insights into features of the interlanguage of a very specific group of learners, projects like the work on learner dictionaries done with the *CLC*, *LLC*, and *ICLE* (cf. Gillard & Gadsby 1998; Rundell & Granger 2007) can only be realized on the basis of large learner corpora. To my knowledge, there are no German learner dictionaries equivalent to English learner dictionaries like the *Macmillan English Dictionary for Advanced Learners* (Rundell 2007), the *Longman Language Activator* (Summers 2006), or the *Cambridge Advanced Learner's Dictionary* (Woodford 2003), and standard dictionaries of German for foreign language learners do not include specific notes on common errors. To devise such materials for German as a foreign language, which would surely be a great advantage for language learners, broader collections of learner language appear to be an indispensable resource.

In this respect, the question of quality vs. quantity certainly comes into play. Taking *FALKO* as an example, it is clear that the focus on the multi-layer annotation scheme considerably slows the compilation process and decreases the capability for processing new material. It does, however, increase the possibilities for future analysis of the corpus. Then again, as we have seen, complex structures like this do not allow for analysis with common corpus exploration tools and are probably difficult to use in collaboration with other projects, unless they use the same system. Consequently, the decisions regarding a certain corpus structure and annotation scheme, as well as intended size of a learner corpus, need to take into account aspects such as

- the research envisaged (possibly based on a needs analysis),
- the amount of linguistic annotation needed,
- possibilities for collaboration, and
- the reusability of the data.

Even (comparatively) simple part-of-speech tagging can open up a wide range of research perspectives, and the possibilities for both the collaboration with other projects and the reusability of the data increase with simpler systems for data organization. It seems, therefore, that it might be a desirable step for German learner corpus research to start collecting learner data in a less complex system in order to arrive at a database that provides for a broader analysis of features of German learner language. This, of course,

would not exclude the integration of this data into a more complex database with a multi-layer annotation scheme in the future. That is to say that, without derogating the potential of projects such as *FALKO*, a less specialized database of learner language would be a valuable addition to the existing projects.

The fact that there are fewer German learner corpora, and also fewer German L1 corpora, than English L1 and L2 corpora, also implies that there are fewer automatic taggers for German available and that those available have not been used on the same amount and variety of data. Thus, German learner corpus research cannot yet draw on as many reports and evaluations reflecting the experience of other projects with certain programs, which obviously makes planning a new project all the more laborious. To conclude, even though the limited number of German learner corpus projects means that there have been very few possibilities for collaboration up to now, a networked environment seems all the more important in order to organize and join the limited resources of German learner corpus research.

5. Concluding remarks

The status quo of corpora of learner English and of learner German is certainly very different and cannot simply be compared in disregard of the political and social status of the two languages. This is especially true for quantitative aspects. In terms of qualitative aspects and methodological issues, however, I would argue that considering the developments in the compilation of learner corpora for different target languages, i.e. in this case English and German learner corpora, can be a fruitful enterprise.

The challenges faced by German and English learner corpus research certainly differ considerably in many respects. While in the case of German, there is certainly room for more projects and for a larger and more comprehensive collection of German learner language, in the case of English, the diversity of existing corpora calls for an effort towards standardization and mutual compatibility. One might argue that, due to the diverse rationales for their collection, learner corpora are in principle incompatible. However, the accessibility of detailed documentation and resources in a networked environment could aid the comparability of individual projects.

As regards collaboration, the corpus projects based at the Centre for English Corpus Linguistics at the Catholic University of Louvain, i.e. the *ICLE* and the recently launched project *LONGDALE* are certainly exceptional in that they bring together several local compilations in a centralized standard form. As with most learner corpora, however, they compile data at the level of

tertiary education. As the present study revealed, not only lower level but also longitudinal data is still underrepresented in learner corpus projects (cf. section 4.1).

While I have argued and strongly believe that cooperation and standardization would certainly be worth aiming at, this is undoubtedly often difficult to realize. Every new corpus project faces the problem of having to find a system of storing the data in a way that fits the respective needs and resources. In the dilemma between complexity and realizability, knowledge of other projects, of the data they comprise and the systems they employ, is indispensable. The provision of detailed information in an accessible manner would be a crucial first step towards the networked environment envisaged. Speaking as someone who is herself involved in a recently launched 'solo project' (namely *DELT*), I would thus like to modify Yukio Tono's assertion quoted in the beginning and claim that apart from sharing knowledge on corpus building we should support a networked approach by making up-to-date information on individual projects easily available and encourage one another not to jump on the bandwagon of corpus-based research without sufficient knowledge of organization and content of existing corpora and ongoing projects.

Networking seems feasible especially with regard to technical aspects, which could free up resources for other issues concerning collaborative projects or corpus analysis. Without some movement towards a collaborative environment, English learner corpus research is headed towards a situation of unconnected proliferation of individual projects and increasing incompatibility. Should more projects be launched, German learner corpus research would most likely face the same problem. This leaves me to assert that a networked environment should be an imperative in all learner corpus projects, irrespective of the target language.

Appendix

The listings of learner corpora of English and German provided are based on a study I conducted in 2007. However, where information was available on the websites of the respective corpora, information such as corpus size was updated, and eight corpora that were not included in the 2007 study were added. References to the sources from which the information about the corpora was gathered are provided in the column *references*.

Since some of the corpora are part of ongoing projects, there is a possibility that recent changes in certain particulars are not rendered in the table. As already pointed out in section 3, though aiming at a comprehensive record of the status quo, the list does certainly not include all corpora of learner English and German. I would thus be grateful for any advice regarding the information given or regarding corpora that are not mentioned here.

I. Learner Corpora of English

NAME	L1	data type	size/ annotation	access	references	*
CEJL (Corpus of English by Japanese Learners)	Japanese	cross-sectional, tertiary level	?	documentation and texts available at: http://www.eng.ritsumei.ac.jp/asao/lcorpus/	- *Asao Kojiro. Learner Corpus Data*: http://www.eng.ritsumei.ac.jp/asao/lcorpus/ - *Corpus of English by Japanese Learners:* http://www.eng.ritsumei.ac.jp/asao/lcorpus_prev/	A
CLC (Cambridge Learner Corpus)	100 different L1s	cross-sectional, exam scripts	30,000,000 words, error tagged	not publicly available (available only to authors working for CUP and for staff at Cambridge ESOL)	- *Cambridge International Corpus. Cambridge Learner Corpus*: http://www.cambridge.org/elt/corpus/learner_corpus2.htm - Pravec 2002	E
CLEC (Chinese Learner English Corpus)	Chinese	cross-sectional, upper secondary & tertiary level	1,000,000 words, error tagged	available to users in the Department of English at HKPU, http://langbank.engl.polyu.edu.hk/engine.aspx?Submit=Search&lang=1&corpus=7 (password protected)	- *The PolyU English Department Language Bank. Chinese Learner English Corpus*: http://langbank.engl.polyu.edu.hk/indexl.html	A
DELT (Database of English Learner Texts)	mostly German	longitu-dinal	under construction (pilot phase launched 2007)	-	- *Centre for English Language Teaching (University of Vienna)*: http://www.univie.ac.at/FDZ-Englisch/projects.html	E
HKUST Corpus (Hong Kong University of Science and Technology)	mostly Cantonese	cross-sectional, tertiary level, school-leaving exams (1 m words)	30,000,000 words; 200,000 words POS tagged, partly error tagged	large parts have been made available to researchers upon request	- Milton 1998 - Milton 2001 - Milton & Chowdhury 1994 - Pravec 2002 - correspondence with John Milton	A
ICLE (Internatio-nal Corpus of Learner English)	various L1s	cross-sectional;	2,500,000 words (in 2002); French, German and Spanish subcorpus are currently being error tagged	available for purchase on CD-ROM (2002) (2nd POS tagged edition to be published in 2008)	- Granger et al. 2002 - *Centre for English Corpus Linguistics. International Corpus of Learner English*: http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm - Pravec 2002 - correspondence with Sylviane Granger	E

JEFLL Corpus (Japanese EFL Learner)	Japanese	quasi-longitu-dinal, lower/ upper secondary level	700,000 words, POS-tagged, partly error tagged	online access at http://jefll.corpuscobo.net/ (currently Japanese interface; English interface will be available in 2008)	- *The JEFLL Corpus Project*: http://jefll.corpuscobo.net/ - Pravec 2002	A
JPU Corpus (Janus Pannonius University)	Hungarian	longitu-dinal ?	300,000 words	concordance search: http://www.lextutor.ca/concordancers/concord_e.html, thematic search: http://joeandco.blogspot.com/	- Horváth 1999 - Pravec 2002 - correspondence with József Horváth	E
LANCAWE (Lancaster Corpus of Academic Written English)	various L1s	longitu-dinal (4-8 weeks), tertiary level (pre-sessional & under-graduate courses)	under construction	freely available at http://www.ling.lancs.ac.uk/groups/slarg/lancawe/data/index.htm (not all the material has been made available yet)	- *LANCAWE. Lancaster Corpus of Academic Written English*: http://www.ling.lancs.ac.uk/groups/slarg/lancawe/	E
Learner Corpus of English for Business Communi-cation	Cantonese ?	cross-sectional, tertiary level (different types of business correspon-dence)	~ 117,500 words	available online at http://langbank.engl.polyu.edu.hk/engine.aspx?Submit=Search&lang=1&corpus=15	- *The PolyU English Department Language Bank. Chinese Learner English Corpus*: http://langbank.engl.polyu.edu.hk/indexl.html	A
Learner Corpus of Essays and Reports	Cantonese ?	cross-sectional, tertiary level (essays, project reports)	188,000 words	available online at http://langbank.engl.polyu.edu.hk/engine.aspx?Submit=Search&lang=1&corpus=16	- *The PolyU English Department Language Bank. Chinese Learner English Corpus*: http://langbank.engl.polyu.edu.hk/indexl.html	A
Learner Journals	Cantonese ?	cross-sectional, tertiary level	16,500 words	available online at http://langbank.engl.polyu.edu.hk/engine.aspx?Submit=Search&lang=1&corpus=17	- *The PolyU English Department Language Bank. Chinese Learner English Corpus*: http://langbank.engl.polyu.edu.hk/indexl.html	A
LLC (Longman Learner Corpus)	various L1s	cross-sectional, various levels	10,000,000, partly error tagged	available for research?	- *Longman Corpus Network. The Longman Learners' Corpus*: http://www.pearsonlongman.com/dictionaries/corpus/learners.html - Pravec 2002	E

LONG-DALE (Longitu-dinal Database of Learner English)	Various L1s	longitu-dinal	under construction (launched 2008)	-	- *The LONGDALE Project. Longitudinal Database of Learner English*: http://cecl.fltr.ucl.ac.be/LONGDALE.html	E
MELD (Montclair Electronic Language Database)	various L1s (second language context)	cross-sectional, advanced level of proficiency, ESL context	~ 100,000 words, ~ 50% POS-tagged & error tagged	~ 25% of the data available online at http://chss.montclair.edu/linguistics/MELD/	- Fitzpatrick and Seegmiller 2004 - *The Montclair Electronic Language Database*: http://www.chss.montclair.edu/linguistics/MELD/ - Pravec 2002 - correspondence with Eileen M. Fitzpatrick	NA
MLC (Multi-lingual Learner corpus)	Portuguese (Brazilian)	cross-sectional / longitu-dinal?	under construction	http://www.jr.icmc.usp.br/~comet/ (currently password protected; there are plans to make the data publicly available)	- Tagnin 2006 - correspondence with Stella Tagnin and Guilherme Fromm	SA
PICLE (Polish sub-corpus of the ICLE)	Polish	cross-sectional, tertiary level	330,000 words	available online at http://ifa.amu.edu.pl/~ifaconc/main.php	- *The PICLE Corpus Homepage*: http://www.staff.amu.edu.pl/~przemka/picle.html	E
PLE (PELCRA corpus of learner English)	Polish	quasi-longitu-dinal?, tertiary level (exams)	500,000 words; POS-tagged with CLAWS	not available?	- Lénko-Szymanska 2004 - Pravec 2002	E
PLEC (PolyU Learner English Corpus)	Cantonese ?	cross-sectional, tertiary level (exams)	1,000,000 words	available online at http://langbank.engl.polyu.edu.hk/engine.aspx?Submit=Search&lang=1&corpus=39	- *The PolyU English Department Language Bank. Chinese Learner English Corpus*: http://langbank.engl.polyu.edu.hk/indexl.html	A
QUÉBEC LEARNER CORPUS	French (Canadian)	cross-sectional; tertiary level; (placement essays)	250,000 words; no linguistic annotation	concordance search: http://www.lextutor.ca/concordancers/concord_e.html	- Cobb 2003 - correspondence with Tom Cobb	NA
SILS Learner Corpus of English (School of Internatio-nal Liberal Studies at Waseda	mostly Japanese	longitu-dinal	in 2007: 3,180,000 words, (1,650,000 words of first drafts, the rest second drafts)	plans to make the corpus available	- Muehleisen 2006 - *The SILS Learner Corpus of English*: http://www.f.waseda.jp/vicky/learner/index.html - correspondence with Victoria Muehleisen and Steve Chen	A

University)						
SULEC (Santiago University Learner of English Corpus)	Spanish	Cross-sectional; secondary and tertiary level (spoken & written data!)	in 2007: 450,000 words (written and spoken; aim: 1,000,000 words)	available on request (?) at http://sulec.cesga.es/ (password protected)	- *The Santiago University Learner of English Corpus (SULEC)*: http://www.usc.es/ia303/SULEC/SULeC.htm - correspondence with Ignacio M. Palacios Martinez	E
TELE-KORP (Telecollaborative Learner Corpus of English and German)	English, German (bilingual corpus)	longitudinal, computer mediated communication	~ 1,500,000 words (both German and English)	not available	- Belz; Vyatkina 2008 - *Telekorp: The Telecollaborative Learner Corpus of English and German*: http://www.personal.psu.edu/faculty/j/a/jab63/Telekorp.html (checked in 2007; no longer available)	NA
TLCE (Taiwanese Learner Corpus of English)	Mandarin / Taiwanese ?	cross-sectional; tertiary level	730,000 words; POS-tagged, lemmatized	not available ?	Hsue-Hueh Shih 2000	A
TSLC (TELEC secondary learner corpus)	Cantonese	cross-sectional; secondary level	2,200,000 words	available to Hongkong Teachers and TELEC researchers	- Allan 2002 - *TeleNex. A Resource for English Teachers in Honk Kong Schools*: http://www.telenex.hku.hk/telec/pmain/opening.htm - Pravec 2002	A
USE Corpus (Uppsala Student English Corpus)	Swedish	longitudinal, tertiary level	1,211,265 words	available from the Oxford Text Archive http://ota.oucs.ox.ac.uk/headers/2457.xml	- Westergren Axelsson 2000 - *Uppsala Student English Corpus (USE)*: http://www.engelska.uu.se/use.html - correspondence with Ylva Berglund Prytz - Pravec 2002	E

II. Learner Corpora of German

NAME	L1	data type	size/ annotation	access	references	
FALKO (Fehler-annotiertes Lerner-korpus des Deutschen als Fremd-sprache)	various L1s	cross-sectional; tertiary level; 3 subcorpora (1 subcorpus longitu-dinal)	core corpus 2005: ~ 36,000 tokens; POS-tagged, syntactic structure and errors tagged in 2 subcorpora;	available online at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko	- *Falko. Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache*: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko - Lüdeling at al 2005 - Siemen et al 2006 - correspondence with Anke Lüdeling	E
LEKO (Lerner-korpus)	?	cross-sectional; tertiary level	30 texts, annotation: lemma, POS, morpho-syntax, error	available for members of the Humboldt University from https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/institutkorpora/	- *LEKO Lernerkorpus. Handbuch*: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/lehre/alt/ws-2004/hs-phaenomene/pdf/LekoHandbuch.pdf	E
MLC (Multi-lingual Learner corpus)	Portuguese (Brazilian)	cross-sectional / longitu-dinal?	under construction	http://www.jr.icmc.usp.br/~comet/ (currently password protected; there are plans to make the data publicly available)	- Tagnin 2006 - correspondence with Stella Tagnin and Guilherme Fromm	SA
TELEKO RP (Telecolla-borative Learner Corpus of English and German)	English, German (bilingual corpus)	longitu-dinal, computer mediated communi-cation	~ 1,500,000 words (both German and English)	not available	- Belz; Vyatkina 2008 - Telekorp: The Telecollaborative Learner Corpus of English and German: http://www.personal.psu.edu/faculty/j/a/jab63/Telekorp.html (checked in 2007; no longer available)	NA
C-LEG (Weinber-ger, Lancaster)	English	quasi-longitu-dinal?	27635 words (95 texts), error tagged	not available	- Lüdeling at al 2005 (- Weinberger 2002 quoted in Lüdeling et al 2005)	E

*** location of the corpus project**

A = Asia NA = North America

E = Europe SA = South America

References

Allan, Quentin Grant. 2002. "The TELEC secondary learner corpus. A resource for teacher development". In Granger, Sylviane; Hung, Joseph; Petch-Tyson, Stephanie (eds.). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam, Philadelphia: Benjamins, 195-212.

Barlow, Michael. 2005. "Computer-based analyses of learner language". In Ellis, Rod; Barkhuizen, Gary. *Analysing learner language*. Oxford [et al.]: Oxford University Press, 335-359.

Barnbrook, Geoff. 1996. *Language and computers. A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Belz, Julie A.; Vyatkina, Nina. 2008. "The pedagogical mediation of a developmental learner corpus for classroom-based language instruction". *Language Learning & Technology* 12/3, 33-52. Available at http://llt.msu.edu/vol12num3/belzvyatkina.pdf (14 Nov. 2008).

Breen, Michael P. 1985. "Authenticity in the language classroom". *Applied Linguistics* 6, 60-70.

Cobb, Tom. 2003. "Analyzing late interlanguage with learner corpora: Québec replications of three European studies". *The Canadian Modern Language Review/La Revue canadienne des langgues vivantes* 59/3, 393-423. Available at http://www.lextutor.ca/cv/pdf/learner_corpus.pdf (14 Nov. 2008).

Crystal, David. 2001. *Language and the internet*. Cambridge: Cambridge University Press.

Dagneaux, Estelle [et al.]. 2005. *Error tagging manual version 1.2*. Centre for English Corpus Linguistics, Université Catholique de Louvain.

Díaz-Negrillo, Ana; Fernández-Domíngues, Jesús. 2006. "Error tagging systems for learner corpora". *RESLA* 19 (2006), 83-102.

Fitzpatrick, E.; Seegmiller, M.S. 2004. "The Montclair electronic language database project". In Connor, U.; Upton, T.A. (eds.). *Applied corpus linguistics: a multidimensional perspective*. Amsterdam: Rodopi, 223-237. Available at http://www.chss.montclair.edu/linguistics/MELD/rodopipaper.pdf (14 Nov. 2008).

Gillard, Patrick; Gadsby, Adam. 1998. "Using a learners' corpus in compiling ELT dictionaries". In Granger, Sylviane (ed.). *Learner English on computer*. London, New York: Longman, 159-171.

Granger, Sylviane. 1998. "The computer learner corpus: a versatile new source of data for SLA research". In Granger, Sylviane (ed.). *Learner English on computer*. London, New York: Longman, 3-18.

Granger, Sylviane. 2002. "A Bird's-eye view of learner corpus research". In Granger , Sylviane; Hung, Joseph; Petch-Tyson, Stephanie (eds.). 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam, Philadelphia: Benjamins, 3-33.

Granger, Sylviane. 2009 (forthcoming). "The contribution of learner corpora to second language acquisition and foreign language teaching: a critical evaluation". Aijmer, K. (ed.). *Corpora and Language Teaching*. Benjamins.

Granger, Sylviane; Dagneux, Estelle; Meunier, Fanny (eds.). 2002. *International Corpus of Learner English*. (CD-ROM & Handbook). Presses Universitaires de Louvain.

Horváth, József. 1999. *Advanced writing in English as a foreign language. A corpus-based study of processes and products*. (PhD Dissertation. Defended on May 12, 2000) Online edition. Pécs, Hungary. Available at http://www.geocities.com/writing_site/thesis/ (14 Nov. 2008).

Hsue-Hueh Shih, Rebecca. 2000. "Compiling Taiwanese learner corpus of English." *Computational Linguistics and Chinese Language Processing* 5/2, 89-102. Available at http://www.aclclp.org.tw/clclp/v5n2/v5n2a4.pdf (14 Nov 2008).

Kaltenböck, Gunther; Mehlmauer-Larcher, Barbara. 2005. "Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching". *ReCALL* 17/1, 65-84.

Leech, Geoffrey. 1998. "Preface". In Granger, Sylviane (ed.). *Learner English on computer*. London, New York: Longman, xiv-xx.

Leech, Geoffrey. 2005. "Adding linguistic annotation". In Wynne, Martin (ed.). *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books, 1-16. Available at http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm (14 Nov. 2008).

Lénko-Szymanska, Agnieszka. 2004. "Demonstratives as anaphora markers in advanced learners' English". In Aston, Guy; Bernardini, Silvia; Stewart, Dominic (eds.). *Corpora and language learners*. (Studies in Corpus Linguistics 17). Amsterdam, Philadelphia: Benjamins, 89-107.

Lüdeling, Anke [et al.]. 2005. "Multi-level error annotation in learner corpora". In *Proceedings of Corpus Linguistics 2005*, Birmingham. Available at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/pdf/FALKO-CL2005.pdf (14 Nov. 2008).

McEnery, Tony; Wilson, Andrew. 2001. *Corpus linguistics. An introduction*. (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, Tony; Xiao, Richard; Tono, Yukio. 2006. *Corpus-based language studies. An advanced resource book*. London: Routledge.

Meunier, Fanny. 1998. "Computer tools for the analysis of learner corpora". In Granger, Sylviane (ed.). *Learner English on computer*. London, New York: Longman, 19-37.

Milton, John. 1998. "Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment". In Granger, Sylviane (ed.). *Learner English on computer*. London, New York: Longman, 186-198.

Milton, John. 2001. *Elements of a written interlanguage: a computational and corpus-based study of institutional influences on the acquisition of English by Hong Kong Chinese students*. (Language Centre Research Reports 2). Hong Kong: Language Centre, HKUST. Available at http://hdl.handle.net/1783.1/1055 (14 Nov. 2008).

Milton, John; Chowdhury, Nandini. 1994. "Tagging the interlanguage of Chinese learners of English". *Proceedings joint seminar on corpus linguistics and lexicology, Guangzhou and Hong Kong, 19-22 June, 1993, Language Centre, HKUST, Hong Kong, 1994*, 127-143. Available at http://repository.ust.hk/dspace/bitstream/1783.1/1087/1/entertext03.pdf (14 Nov. 2008).

Muehleisen, Victoria. 2006. "Introducing the SILS Learners' Corpus: a tool for writing curriculum development". *Waseda Global Forum* 3, 119-125.

Myles, Florence. 2008. "Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues". In Ortega, Lourdes; Byrnes, Heidi. *The longitudinal study of advanced L2 capacities*. New York: Routledge, 58-72.

Nesselhauf, Nadja. 2004. "Learner corpora and their potential for language teaching". In Sinclair, John McHardy (ed.). *How to use corpora in language teaching*. (Studies in Corpus Linguistics 12). Amsterdam, Philadelphia: Benjamins, 125-152.

Nicholls, Diane. 2003. "The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT". *Proceedings of the Corpus Linguistics 2003 Conference, Lancaster University (UK), 28 - 31 March 2003*. Available at http://ucrel.lancs.ac.uk/publications/CL2003/papers/nicholls.pdf (14 Nov. 2008)

Pravec, N. A.. 2002. "Survey of learner corpora". *ICAME Journal* 26, 81-114. http://icame.uib.no/ij26/pravec.pdf (14 Nov. 2008).

Rundell, Michael (ed.). 2007. *Macmillan English dictionary for advanced learners*. 2nd edition. Oxford: Macmillan.

Rundell, Michael; Granger, Sylviane. 2007. "From corpora to confidence". *ENGLISH TEACHING professional* 50, 15-18.

Siemen, Peter; Lüdeling, Anke; Müller, Frank Henrik. 2006. "FALKO – Ein fehlerannotiertes Lernerkorpus des Deutschen". *Proceedings of Konvens 2006, Konstanz*. Available at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/pdf/SiemenLuedelingMueller-Konvens06.pdf (14 Nov. 2008).

Schiftner, Barbara. 2007. *Learner corpora of English and German – the status qQuo and future prospects*. Unpublished MA Thesis, University of Vienna.

Sinclair, John. 2005. "Corpus and text – basic principles". In Wynne, Martin (ed.). *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books, 1-16. Available at http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm (14 Nov. 2008).

Summers, Della [ed.]. 2006. *Longman language activator*. (2nd ed.). Harlow: Longman.

Tagnin, Stella E. O. 2006. "A multilingual learner corpus in Brazil". In Wilson, Andrew; Archer, Dawn; Rayson, Paul (eds.). *Corpus linguistics around the world*. Amsterdam: Rodopi, 195-202. Available at http://www.fflch.usp.br/dlm/comet/artigos/A%20multilingual%20learner%20corpus%20in%20Brazil.pdf (14 Nov. 2008).

Taylor, David. 1994. "Inauthentic authenticity or authentic inauthenticity?". *Teaching English as a Second or Foreign Language* 1/2. http://www-writing.berkeley.edu/TESL-EJ/ej02/a.1.html (14 Nov. 2008).

Tono, Yukio. 2003. "Learner corpora: design, development, and applications". In Archer, D.; Rayson P.; Wilson A.; McEnery T. (eds.). *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003)*. Technical Papers 16. Lancaster University: University Centre for Computer Corpus Research on Language, 800-809.

Weinberger, Ursula. 2002. *Error analysis with computer learner corpora. A corpus-based study of errors in the written German of British university students*. MA thesis, Lancaster University.

Westergren Axelsson, Margareta. 2000. "USE – the Uppsala Student English Corpus: an instrument for needs analysis". *ICAME Journal* 24, 155-157. Available at http://icame.uib.no/ij24/use.pdf (14 Nov. 2008).

Widdowson, H.G. 1980. *Explorations in applied linguistics*. (2nd impr.). Oxford: Oxford University Press.

Widdowson, H.G. 1990. *Aspects of language teaching*. Oxford, New York: Oxford University Press.

Widdowson, H.G.. 2000. "On the limitations of linguistics applied". *Applied Linguistics* 21/1, 3-25.

Widdowson, H.G. 2003. *Defining issues in English language teaching*. Oxford, New York: Oxford University Press.

Woodford, Kate (ed.). 2003. *Cambridge advanced learner's dictionary*. Cambridge: Cambridge University Press.

Wynne, Martin (ed.). *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books, 1-16. Available at http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm (14 Nov. 2008).

How to contact us:

c/o
Institut für Anglistik & Amerikanistik der Universität Wien
Universitätscampus AAKH, Spitalgasse 2, Hof 8
A – 1090 Vienna; Austria

fax	**(intern.) 43 1 4277 9424**
eMail	**theresa.illes@univie.ac.at**
	marie-luise.pitzl@univie.ac.at
W3	http://www.univie.ac.at/Anglistik/ang_new/online_papers/views.html
	(all issues available online)

IMPRESSUM:
EIGENTÜMER, HERAUSGEBER & VERLEGER: VIEWS, C/O INSTITUT FÜR ANGLISTIK & AMERIKANISTIK DER UNIVERSITÄT WIEN, UNIVERSITÄTSCAMPUS AAKH, SPITALGASSE 2, A - 1090 WIEN, AUSTRIA. **FÜR DEN INHALT VERANTWORTLICH**: THERESA-SUSANNA ILLES, MARIE-LUISE PITZL **WEBMASTER**: STEPHEN FERGUSON **REDAKTION**: HEIKE BÖHRINGER, ANGELIKA BREITENEDER, CHRISTIANE DALTON-PUFFER, CORNELIA HÜLMBAUER, JULIA HÜTTNER, THERESA-SUSANNA ILLES, BRYAN JENNER, GUNTHER KALTENBÖCK, THERESA KLIMPFINGER, KATHRIN KORDON, URSULA LUTZKY, BARBARA MEHLMAUER-LARCHER, MARIE-LUISE PITZL, ANGELIKA RIEDER-BÜNEMANN, NIKOLAUS RITT, HERBERT SCHENDL, BARBARA SCHIFTNER, BARBARA SEIDLHOFER, UTE SMIT, BARBARA SOUKUP, JOHANN UNGER, H.G. WIDDOWSON. ALLE: C/O INSTITUT FÜR ANGLISTIK & AMERIKANISTIK DER UNIVERSITÄT WIEN, UNIVERSITÄTSCAMPUS AAKH, SPITALGASSE 2, A - 1090 WIEN. **HERSTELLUNG:** VIEWS