Research in Corpus Linguistics

ISSN 2243-4712

Issue 1 2013

Special issue "The use of corpora for language teaching and learning"

Issue Editors: Antonio Vicente Casas-Pedrosa, Jesús Fernández-Domínguez and Alejandro Alcaraz-Sintes

Contents

RiCL Research in Corpus Linguistics

Introduction: the use of corpora for language teaching and learning

Antonio Vicente Casas-Pedrosa[*], Jesús Fernández-Domínguez[**] and Alejandro Alcaraz-Sintes*
Universidad de Jaén[*], Universitat de València[**] / Spain

Abstract – This paper aims at contextualizing and presenting the first volume of the journal *Research in Corpus Linguistics* and is, therefore, divided into two main parts. First of all, it provides an introduction to the field of corpus linguistics and its increasingly relevant role in language teaching and learning. Secondly, it briefly introduces and discusses the six articles of the volume. Stemming from oral presentations delivered at the 4th International Conference on Corpus Linguistics (CILC2012, Jaén, Spain), these articles have a number of features in common. They all make extensive use of corpora and at the same time deal with language teaching and language learning, the underlying assumption being that a genuine and mutually beneficial connection can be established between teaching and research. For this reason, each of them constitutes an illustrative sample of how different corpora can be exploited for different research purposes.

Keywords – corpus linguistics, language description, language learning, language teaching

This first volume of the journal *Research in Corpus Linguistics* is a special issue with six contributions that were originally presented at *CILC2012* (*4th International Conference on Corpus Linguistics*), held at the University of Jaén (Spain) from 22nd to 24th March 2012. Although these papers were presented in five of the nine thematic panels established by *AELINCO* (*Spanish Association for Corpus Linguistics*) for its conferences, they all share a common feature: they explore the use of corpora for the teaching and learning of a language.

Formerly, the main approach to the description or teaching of a language used to be prescriptive, with an emphasis on the language officially accepted as 'grammatical', while 'ungrammatical' (or non-acceptable) productions were condemned. Later developments underlined that many of those linguistic prescriptions were subjective and not based on empirical evidence, which meant a readjustment whereby the study of linguistic phenomena should be tackled from a strictly scientific point of view (Quirk et al. 1985: 14; Nelson 2005). This subsequently involved a different approach to the analysis of language, and more recent trends of language description have adopted this viewpoint. According to the new perspective, the linguist's task is the analysis of genuine language from a descriptive and empirical point of view. It is at this point that the term 'corpus' comes into play.

According to the *Oxford English Dictionary* (s.v. *corpus* 3.b), the first recorded written example of the word *corpus*, understood as "[t]he body of written or spoken material upon which a linguistic analysis is based", dates back to 1956. It appears in the following excerpt by Allen (1956: 128): "The analysis here presented is based on the speech of a single informant (…) and in particular upon a corpus of material, of which a large proportion was narrative, derived from approximately 100 hours of listening". This involves not only corpora as tools of their own, but also their use and application to the creation of other resources.

Until very recently, dictionaries, grammars and textbooks, among others, have incorporated custom-made examples drawn up by lexicographers and scholars for the purpose of illustrating, *a posteriori*, the concepts, definitions and explanations being presented. Of course, these examples were not necessarily unreliable, since they were most often made up by native speakers, who were considered "a living dictionary" (Rundell 2007: vii). These samples of language, however, lacked something that is considered essential today: the support of empirical evidence. They were often

perceived as somehow artificial, since they seemed to have been created on the spot and without a specific communicative context. Nowadays, most works on language description (teaching manuals, grammar books, dictionaries, etc.) underline the fact that they have used a language corpus as the source of examples. For example, Carter and McCarthy's (2006) backcover includes the catch-phrases "real English guarantee" and "real everyday usage", thus highlighting the idea that the description within the work is based on the actual usage – written and spoken – of the language by native speakers from all over the world.

Thanks to the advances of computer science, it is now possible to access resources essential for the linguistic description of language, first and foremost language corpora, which are also playing an increasingly relevant role in language learning and teaching. Despite their relatively short existence, corpora have already become a crucial tool for the analysis of languages, and a dynamic relationship has flourished between corpora and language teaching. The increasing number of articles, books and journals on the topic published every year, on the one hand, and the amount of conferences, symposia and workshops organized worldwide, as well as the creation of academic associations closely related to corpora and their exploitation, all bear witness to the phenomenon (e.g., Granger et al. 2002; Aston et al. 2004; Gavioli 2005; Braun et al. 2006; Hidalgo et al. 2007, among many others).

The applications of corpora for language teaching have been discussed, for example, by Leech (1997), Römer (2008) and McEnery and Xiao (2011), who differentiate between indirect and direct uses. Corpora are being indirectly used, for example, for the design of teaching syllabi with an emphasis on communicative competence, e.g., the *Collins COBUILD English Course* (Willis and Willis 1989; see Hernández this issue) or when representing the frequency of occurrence of language items in grammar and usage handbooks. Other indirect applications are found in Language for Specific Purposes (LSP) corpora, learner corpora and translation corpora, each with different implications for the language classroom. The LSP lesson, as a token, can benefit from the creation of genre-specific corpus-derived glossaries or from concordance data when creating authentic teaching materials. On the other hand, learner and translation corpora are two of the most widely employed upshots of corpus linguistics for language pedagogy, and offer a variety of practical uses, such as learner dictionaries, syllabus design or the creation of teaching materials based on error analysis (see Granger 2002; Meunier 2002; Carrió Pastor and Mestre Mestre this issue; Wierszycka this issue). As regards the direct applications of corpora, scholars have often reported positive results when students are faced with hands-on tools such as online corpora or when they are able to retrieve and discuss concordance lines on a relevant topic (Bernardini 2002; Milojkovic this issue). It seems that this kind of data-driven learning furthers an autonomous and interactive kind of learning between students and language data, while teachers are able to move from the role of information provider to that of facilitator.

Notwithstanding the above benefits, the use of corpora in the classroom is not without difficulties. Once the decision is made to employ a corpus, one of the most frequent dilemmas is whether to exploit one of the many corpora available or to compile a new one, which naturally depends on issues such as the target audience and the availability of appropriate texts. There are numerous general-purpose corpora today, especially for languages like English, French, German, Italian or Spanish, and there exists a vast range of alternatives to choose from. For more specific registers, the option of an *ad hoc* corpus is at hand as well, although such relevant issues as balance, representativeness, design or sampling will have to be seriously considered in the compilation process. It should also be remembered that large size is not always advantageous, and that smaller corpora can prove very effective if accurately selected (Leech 1997: 22; Meunier 2002: 129).

The present collection of articles intends to illustrate some of the aforementioned matters by resorting to the experiences of six scholars in this field. Following this introduction, Hernández delves into the relevance of corpus linguistics for the linguistics curriculum by looking at material from a corpus made up of computer-mediated texts. Based on the experience gathered from a research project at the University of Duisburg-Essen, this article pays attention to the language of 'digital discourse' (Crystal 2010) as found in e-mails, chat rooms, text messages, blogs, forums as well as in a variety of social media services (*Facebook*, *Twitter*, *YouTube*, etc). The article discusses the various specific procedures that these types of texts require before they can be effectively incorporated to a corpus, among them user privacy, a definition of textual units, the tagging of non-standard spellings and errors, and the codification of images or emoticons. The author also explores the possibility of implementing major issues in corpus construction into the academic curriculum in the form of project-based learning. Hernández ends by presenting a variety of new challenges and possible solutions regarding the compilation and processing of this corpus, for example, the fact that the students themselves wrote a corpus manual with a general description of the textual mark-up and processing guidelines.

In their contribution, Carrió Pastor and Mestre Mestre view errors as a key feature of language learning and focus on the identification and classification of errors related to the students' grammar acquisition process and pragmatic competence. In particular, they look at errors detected in writing with the aim of shedding light on the nature of the mechanisms that foreign learners employ in language production. In contrast to the traditional focus on grammatical errors in second language teaching, this work turns to pragmatics, here understood as the difference between the *official* meaning of a word or sentence and the meaning perceived by the hearer derived from what the speaker said (see Archer et al. 2012). From the comparison between grammar and pragmatics, two objectives are pursued here: first, a proposal for tagging grammatical and pragmatic errors according to the competences laid out in the *Common European Framework of Reference for Languages* (*CEFR*; see Council of Europe 2001) and, second, the establishment of a

correspondence between these two types of errors. The basis for this experiment is a corpus of written texts produced by undergraduate students (B1 level) at the Universitat Politècnica de València, in which the assignments were specifically targeted at the development of pragmatic and grammatical competences. The conclusions of the study are that some grammatical and pragmatic errors coincide and that such a correspondence should be taken into account by language teachers in order to help students in their language learning process.

Wierszycka concentrates on learner English as well, in this case by delving into the semantics of phrasal verbs (PVs). In view of the alleged difficulties that non-native speakers of English experience when faced with PVs (Celce-Murcia and Larsen-Freeman 1999), this article starts from the hypothesis that Polish learners of English master a significantly smaller range of PVs than English native speakers, and that their degree of use of the semantic categories of PVs is inversely proportional to the PVs' level of idiomaticity (see Dagut and Laufer 1985). The evidence for this study is drawn from the Polish component of *LINDSEI* (*Louvain International Database of Spoken English Interlanguage*; see Gilquin et al. 2010) and from the *LOCNEC* (*Louvain Corpus of Native English Conversation*; see De Cock 2003), and is framed within Granger's (1996) scheme of Contrastive Interlanguage Analysis. This comparison of PV usage confirms that while native speakers use PVs in a linear manner, considerable underuse is found on the part of Polish learners; in particular, and thanks to a previous analysis of PV compositionality, Wierszycka verifies that idiomatically opaque PVs are especially neglected when used by Polish learners.

Next, Lee examines the use of modal verbs in academic written feedback as hedging expressions (see Salager-Meyer 2011: 35). By using a corpus of around 36,000 words collected at two Humanities departments in the UK, this investigation turns its attention to the language used by tutors when giving feedback to their students. The author sets out from a wordlist of nine modal verbs (*can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*) and provides extensive evidence (frequencies of occurrence) to discuss the functions of modal verbs in the genre of written feedback, among which we find criticism, suggestion, possibility, necessity, certainties, permission and advice. Among other results, this piece of research shows that *could*, *might* and *would* are the most widely used modal verbs of the set, while *shall* is not present given its lower relevance in the context of written feedback. Intermediate positions are occupied by *may*, *must*, *should* and *will*, each with a different level of certainty that directly affects its higher or lower usage. Interestingly, the author shows that the language used by tutors is rather assertive and direct when the feedback concerns an aspect of writing, while it becomes more indirect and tentative when levelling criticism. These findings can be exploited, for instance, for the improvement of feedback-writing practices in teacher training programmes.

Milojkovic's article is a corpus-based approach to Bill Louw's (1993 and subsequent publications) Contextual Prosodic Theory, which reports the experience resulting from the application of stylistics research on the part of second-year students of English. Two research questions are posed here: one, whether text corpora can help infer authorial text, as postulated in Louw's "text reads text" principle, and two, whether this methodology can be effectively applied to the stylistics classroom. Once the foundations are laid through quantitative and qualitative tests, the experiment depicts the process of meaning construal, where the participants, after an absolute minimum of theoretical background, are provided with concordance lines as a means of interpreting a collocation in a given short excerpt. This virtual absence of instruction in principle leads to unbiased acceptance on the part of the majority of the students. The subjects are tested on semantic prosodies, absent collocates and auras of grammatical strings, through tasks that vary in their format. The relevance of this study lies in the fact that, besides supplying positive answer for the aforementioned two research questions, it is the first application of Louw's theory in the classroom, sporadic studies on semantic prosody aside. The article confirms that text does read text for the non-native students of English at the Belgrade English Department regardless of their level of proficiency.

The volume closes with Szabó's inspection of language ideologies in Hungarian school metalanguage. Revolving around an array of theoretical frameworks (Coulter 2005; Laihonen 2008; Aro 2012), this contribution draws on the *Corpus of Hungarian School Metalanguage-Interview Corpus* (*CHSM-IC*), an annotated transcription of spoken metalanguage based on semi-structured research interviews of Hungarian students, in order to investigate interactional routines used in metadiscourses. On the basis of this material, the author compares texts from well-known handbooks with interview data from *CHSM-IC* and then contrasts the participants' narratives with their own communicational experiences. The article also includes a case study on the Hungarian discourse marker *hát* ('so', 'well'), which illustrates the conflict between language ideologies disseminated by the Hungarian school system and the linguistic self-representation in the interviewees' narratives. This analysis reveals that the narratives of both teachers and students carry a negative evaluation of *hát*, and a detailed discussion of the topic concludes that the use of this marker is an important part of everyday communication practice. Accordingly, one conclusion is that metalinguistic utterances (e.g., answers on grammaticality, statements on linguistic accuracy, etc.) and observable, spontaneous (or semi-spontaneous) language use patterns are regularly not in accordance with each other.

REFERENCES

Allen, William S. 1956. Structure and system in the Abaza verbal complex. *Transactions of the Philological Society* 55: 127–176.

Archer, Dawn, Karin Aijmer and Anne Wichmann. 2012. *Pragmatics. An advanced resource book for students*. London: Routledge.

Aro, Mari. 2012. Effects of authority: voicescapes in children's beliefs about the learning of English. *International Journal of Applied Linguistics* 22/3: 331–346.

Aston, Guy, Silvia Bernardini and Dominic Stewart (eds.). 2004. *Corpora and language learners*. Amsterdam: John Benjamins.

Bernardini, Silvia. 2002. Exploring new directions for discovery learning. In Bernhard Kettemann and Georg Marko (eds.), *Teaching and learning by doing corpus analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19–24 July, 2000*. Amsterdam: Rodopi, 165–182.

Braun, Sabine, Kurt Kohn and Joybrato Mukherjee (eds.). 2006. *Corpus technology and language pedagogy*. Frankfurt: Peter Lang.

Carter, Ronald and Michael McCarthy. 2006. *Cambridge grammar of English. A comprehensive guide. Spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

Celce-Murcia, Marianne and Diane Larsen-Freeman. 1999. *The grammar book: an ESL/EFL teacher's course*. Second edition. Boston, MA: Heinle and Heinle.

CHSM-IC = Magyar Iskolai Metanyelvi Korusz – Interjúkorpusz / Corpus of Hungarian School Metalanguage – Interview Corpus. Budapest: Research Institute for Linguistics of the Hungarian Academy of Sciences. Project chair: T. P. Szabó. <http://metashare.nytud.hu/repository/search> (17 August 2013).

Coulter, Jeff. 2005. Language without mind. In Hedwig te Molder and Jonathan Potter (eds.), *Conversation and cognition*. Cambridge: Cambridge University Press, 79–92.

Council of Europe. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Crystal, David. 2010. The changing nature of text: a linguistic perspective. In Wido van Peursen, Ernst D. Thoutenhoofd and Adriaan van der Weel (eds.), *Text comparison and digital creativity*. Leiden: Brill, 229–251.

Dagut, Menachem and Batia Laufer. 1985. Avoidance of phrasal verbs – a case for contrastive analysis. *Studies in Second Language Acquisition* 7/1: 73–79.

De Cock, Sylvie. 2003. *Recurrent sequences of words in native speaker and advanced learner spoken and written English*. PhD thesis. Louvain: Université Catholique de Louvain.

Gavioli, Laura. 2005. *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle, Sylvie De Cock and Sylviane Granger (comp.). 2010. *Louvain International Database of Spoken English Interlanguage* (*LINDSEI*). Louvain-la-Neuve: UCL Presses universitaires de Louvain.

Granger, Sylviane. 1996. From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, Bengt Altenberg and Mats Johansson (eds.), *Languages in contrast. Textbased cross-linguistic studies*. Lund: Lund University Press, 37–51.

Granger, Sylviane. 2002. A bird's-eye view of learner corpus research. In Granger et al. (eds.), 3–33.

Granger, Sylviane, Joseph Hung and Stephanie Petch-Tyson (eds.). 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Hidalgo, Encarnación, Luis Quereda and Juan Santana (eds.). 2007. *Corpora in the foreign language classroom*. Amsterdam: Rodopi.

Laihonen, Petteri. 2008. Language ideologies in interviews: a conversation analysis approach. *Journal of Sociolinguistics* 12/5: 668–693.

Leech, Geoffrey. 1997. Teaching and language corpora: a convergence. In Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds.), *Teaching and language corpora*. London: Longman, 1–23.

Louw, William E. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds.), *Text and technology. In honour of John Sinclair*. Amsterdam: John Benjamins, 157–176.

McEnery, Tony and Richard Xiao. 2011. What corpora can offer in language teaching and learning. In Eli Hinkel (ed.), *Handbook of research in second language teaching and learning*. London: Routledge, 364–380.

Meunier, Fanny. 2002. The pedagogical value of native and learner corpora in EFL grammar teaching. In: Granger et al. (eds.), 119–141.

Nelson, Gerald. 2005. Description and prescription. In Keith Brown (ed.), *Encyclopedia of language and linguistics*. Oxford: Elsevier, 460–465.

Oxford English Dictionary online. 2013. Oxford: Oxford University Press. <http://www.oed.com> (30 July 2013).

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Römer, Ute. 2008. Corpora and language teaching. In Anke Lüdeling and Merja Kytö (eds.), *Corpus linguistics. An international handbook*. Volume 1. Berlin: Mouton de Gruyter, 112–130.
Rundell, Michael (ed.). 2007. *Macmillan English dictionary for advanced learners*. Second edition. Oxford: Macmillan.
Salager-Meyer, Françoise. 2011. Scientific discourse and contrastive linguistics: hedging. *European Science Editing* 37/2: 35–37.
Willis, Dave and Jane Willis. 1989. *Collins COBUILD English course*. London: HarperCollins.

RiCL Research in Corpus Linguistics

A proposal for the tagging of grammatical and pragmatic errors

María Luisa Carrió Pastor and Eva María Mestre Mestre
Universitat Politècnica de València / Spain

Abstract – Errors should be viewed as a key feature of language learning and language use. In this paper, we focus on the identification and classification of errors that are related to students' grammar acquisition and pragmatic competence. Our objectives are, first, to propose the tagging of grammatical errors and pragmatic errors according to the competences of the *Common European Framework of Reference* (*CEFR*) and, second, to determine where there is a correspondence between the different types of error. In order to meet these objectives, we designed a grid to tag the pragmatic errors produced by students with a B1 level of proficiency. It was based on the errors found in a corpus of written texts produced by undergraduate students at the Universitat Politècnica de València. Students wrote specific assignments based on the proposals specified in the *CEFR* for the development of pragmatic and grammatical competences. The texts were corrected and tagged manually by raters, who classified the errors using the grids and considered whether the errors were grammatical or pragmatic. Finally, the conclusions of our study were that some grammatical and pragmatic errors coincided and that this correspondence should be taken into account by language teachers.

Keywords – *CEFR*, grammatical error, pragmatic error, tagging

1. INTRODUCTION

Traditionally, grammatical errors have constituted one of the most important issues for research into language acquisition (Ellis 1994; James 1998; Ellis, Loewen and Erlam 2006). As other aspects, such as mental processes, the importance of context, the use of language, specific uses of language, etc., were progressively incorporated into language studies, researchers noticed that these issues had implications for error analysis (Bardovi-Hartlig and Dörnyei 1998; Hasbún Hasbún 2004; Schaeffer 2005, 2011). Grammatical, pragmatic and cognitive errors were also included in the document written by the Council of Europe (2001) for the design of a framework covering the most important aspects to be taken into account for language learning, teaching and assessment. It was drawn up on the basis of linguistic theories and the progress that had been made by different approaches to the analysis of language. In this study, we argue that there is significant correspondence between grammatical and pragmatic errors and that what they share should be identified in order to avoid the duplication of the teaching of competences and skills whose attainment is deemed necessary for second language acquisition.

The hypothesis of this study is that grammatical and pragmatic errors should be identified and compared in order to improve the identification of important factors in second language acquisition. As a consequence, the main aim of this study was to tag the errors that can be placed into these two categories and show that grammar and pragmatics should be taught using the same learning strategies, since grammar and pragmatic competences could be related to some of the same kind of errors. The objectives set in this study were, on the one hand, to create tags and an analysis grid for the identification of grammatical and pragmatic errors and, on the other hand, to identify where there was a correspondence between the two in order to recognise language processing from the point of view of grammar and pragmatics.

We believe that if speakers of different languages do not understand each other it is not because their languages do not lend themselves to translation, but because they do not share a common linguistic background, this entailing differences in the observation and interpretation of reality. Therefore, the values that words signify are not represented in the same way; that is, understanding another language does not depend on the existence of equivalent structures, but on the equivalence between the concepts emerging from reality and the method of expressing these. Errors exist due to there being elements of language production which learners have not assimilated (Carrió Pastor 2004, 2005; Mestre Mestre 2011; Mestre Mestre and Carrió Pastor 2012). The research model of error analysis began with the work of Corder (1967), which rejected structuralism and based itself on Chomsky's theory about mankind's innate ability to learn a language, which was itself a refutation of Skinner's behaviourism. Later, James (1998: 1) dedicated his attention to the definition, identification and classification of errors, identifying a language error as an "unsuccessful bit of language". For him, "[E]rror Analysis is the process of determining the incidence, nature, causes and consequences of unsuccessful language" (James 1998: 1). He sees ignorance as the cause of errors made by second language users, which he always analyses by comparing the production of L2 speakers to that of L1 speakers, and not to an idealised language. He classifies errors according to their degree of deviance from the norm, and distinguishes four categories of learner ignorance: grammaticality, acceptability, correctness, and strangeness and infelicity. For James (1998: 65), grammaticality is synonymous with well-formedness, and it is context-free. In James' words, "[a]ppeal to grammaticality is an attempt to be objective, to take decisions such as whether some bit of language is erroneous or not out of the orbit of human whim". So, if we can point to a bit of language and say that there are no circumstances where this could ever be said in this way, we are dealing with ungrammaticality.

Errors demonstrate the way in which people are able to navigate the most complex social interactions, even in the face of linguistic and cultural obstacles. Errors detected in writing can provide us with knowledge of production and help us to understand the mechanisms that the speaker of a foreign language employs. What could emerge from second language research is that certain grammatical and pragmatic features cannot be correctly acquired by second language students following the same learning process.

Among the many aspects of second language teaching and learning which have been studied, grammatical errors have been a major focus of attention for many years. This might seem a little outdated nowadays, but it is not so, since grammar is still considered a crucial part of language teaching, with many Canadian immersion studies (Swain 1985; Lightbown 1992; Lyster 1998) showing that comprehension of meaning and content by itself does not necessarily lead to the acquisition of a native-like grammar.

Some authors insist on the importance of grammar, which has been relegated to second or third place of importance in the new communicative approaches. Terrell (1991) explains that grammar is one of the main components of communicative competence, and there is a risk of it being overlooked in the new teaching methodologies. Rutherford and Sharwood Smith (1985) argue that attention to grammar has an influence on the acquisition process. These authors encourage the creation of what they call 'grammatical awareness raising', both inductive and deductive. The importance of this approach is that it highlights the need for students to recognise grammatical structures and, as a consequence, what constitutes an error.

Some authors have also studied the relationship between grammar and pragmatics. Pragmatics is concerned with the difference between the official meaning of a word or sentence and the actual meaning the speaker intends to give it, and, in the end, the meaning perceived by the hearer derived from what the speaker said (Sperber and Wilson 1995; Wilson and Sperber 1998; Verschueren 1999; Rose and Kasper 2001; Wang 2007; Kasper 2010; Archer, Aijmer and Wichmann 2012). A pragmatic approach considers that there is, on the one hand, knowledge of language, which includes the meanings of words and the ways in which they combine, and then, on the other hand, some general pragmatic principles (often called 'common-sense reasoning principles'), which structure the non-encoded meaning. Kasper (2010: 13) explains the relationship between grammar and pragmatics in this way: "not all grammatical features are good candidates for studying the relationship between pragmatics and grammar […] not all aspects of pragmalinguistic knowledge have a grammatical counterpart".

Focusing on pragmatics, Grice (1975) proposed a co-operative principle of language by means of which speakers of a language should make a contribution such as is required, at the stage at which it occurs, for the accepted purpose or direction of the talk exchange in which the speaker is engaged, by observing the following maxims:

- Quality: try to make your contribution one that is true.
- Quantity: make your contribution as informative as necessary, but not more.
- Relevance: do not say what is not relevant.
- Manner: be brief and orderly, avoid obscurity and ambiguity.

In this sense, pragmatic principles are the cognitive principles that enable us to enrich information by reasoning strategies and language learners should follow the maxims in order to produce pragmatically correct discourse. There have been some studies that have focused on the correct production of language from a pragmatic perspective (Kasper 2010; Rose and Kasper 2001; Bardovi-Harlig 1996, 2013). Bardovi-Harlig (1999) first referred to the area of research devoted to the development of the pragmatic system in second language acquisition as 'acquisitional pragmatics', but

more recently Bardovi-Harlig (2013) has renamed it as 'L2 pragmatics'. All these studies demonstrate the importance of pragmatics in the broad field of second language acquisition and the more specific area of error analysis. Consequently, modelling the communication process with knowledge of pragmatics offers us the basis on which to explain what knowing a language means (i.e., what language competence is) and to gain insights or draw conclusions from the errors that learners make. This perspective differs from the view that linguistic ability consists of a body of knowledge independent from the principles that determine the way language is used (language performance). On this issue, Bardovi-Harlig (1996: 21) has stated the following:

> A learner of high grammatical proficiency will not necessarily show concomitant pragmatic competence. We also have found at least at the higher levels of grammatical proficiency that learners show a wide range of pragmatic competence.

Researchers have approached pragmatic and grammatical errors from different perspectives. Németh and Bibok (2010) distribute these approaches into four categories, with the different ways of understanding the relationship between pragmatics and grammar leading to the establishment of distinct groups of theories. The first of these maintain that grammar and pragmatics are not separate from each other: all matters usually studied within the scope of pragmatics are here considered as grammar. Holistic cognitive grammars (Rumelhart and McClelland 1986) or functional grammars (García Velasco and Portero Muñoz 2002) would fall into this category. The second group views pragmatics as a functional perspective, and not an additional component of a theory of language (Mey 1993; Verschueren 1999). For proponents of this view, pragmatics affects all levels of language and concerns any kind of linguistic phenomena which affect and are affected by the linguistic choices communicators make. A third group would include pragmatics as a component of grammar. For instance Levinson's (2000) theory of 'generalised conversational implicatures', which relates syntax to pragmatics, belongs to this group. The fourth group of theories sees pragmatics as being separate from grammar. The theories of Sperber and Wilson (1995) belong here, as they consider pragmatics to be a component of cognition.

Corpus analysis has been employed in a number of studies of error analysis and error classification, such as those of Granger (2002, 2003a, 2003b); Pérez-Paredes and Cantos-Gómez (2004) and Aguado-Jiménez, Pérez-Paredes and Sánchez (2012) show. More specifically, the tagging of errors has also been a matter of interest for researchers such as Dulay, Burt and Krashen (1982); Dagneaux, Dennes and Granger (1998); Granger (2002, 2003a, 2003b); Díaz Negrillo and Fernández-Domínguez (2006); and Díaz-Negrillo and Valera-Hernández (2010). Error identification has mostly been used for the purpose of establishing which elements of language learning need greater attention in foreign language acquisition and in designing the methodology to employ in language learning (James 1998). In this sense, the categorisation of errors is helpful in that by being able to show which parts of discourse require more attention, it can make a contribution to the key issue of identifying needs in language teaching. Error annotation has become an important aspect to take into account when planning or designing language learning syllabuses, as Díaz-Negrillo and Fernández-Domínguez (2006: 84) explain: "error tagging is indeed inherent to learner corpora and has become a central part of methodology of learner corpus analysis known as computer-aided error analysis".

Dulay, Burt and Krashen (1982) suggest two error taxonomies, one based on linguistic categories and another on the way structures have been altered in the learning process. They establish grammatical, morphological and lexical categories, but they do not consider pragmatic or cognitive aspects when carrying out error tagging. James (1998) combines these two taxonomies into a single, bidimensional taxonomy. Also Dagneaux, Dennes and Granger (1998) identify three levels of descriptive annotation: error domain, error category and word category. More recently, Díaz-Negrillo and Fernández-Domínguez (2006) claim that error analysis should incorporate computer-aided error analysis methodology, and they examine the projects on designing error tagging systems to review error categorisations, dimensions and levels of description. In this paper, our aim is not to provide a new tagging system for errors, but to contrast the tagging of grammatical errors with the tagging of pragmatic errors in order to identify those aspects that overlap and should be considered to be the same error for the purposes of error identification. In the following sections we propose several issues that should be taken into account when tagging grammatical and pragmatic errors. Furthermore, our tagging system also takes into account the descriptors identified by the *Common European Framework of Reference* (*CEFR*) (Council of Europe 2001), with regard to the pragmatic and grammatical competences required for B1 proficiency of English.

2. METHODOLOGY

In the present study, we aimed at identifying errors produced by students at a given level of proficiency (B1) in order to find ways to help them in their language acquisition process. The design of the study was based on the idea that an examination of students' errors can help to identify their level of proficiency and their specific needs in the learning process. To do this, it was thought necessary to provide teachers with guidelines which could help them identify,

classify and categorise errors according to the guiding principles provided by the document written by the Council of Europe (2001), the *Common European Framework of Reference* (*CEFR*). The recommendations shown in the *CEFR* have been updated and improved on several occasions in order to provide definitions of the competences and levels that foreign language learners need to attain in order to speak a language correctly.

In the preliminary stage, two analysis grids were created in order to facilitate the tagging of grammatical and pragmatic errors produced by students with a B1 level of proficiency in English. Lower levels were not considered, as pragmatic competences are difficult to express and such errors are problematic to detect at lower levels. The grid proposed in this paper, based on Mestre Mestre (2011), was elaborated using the proposals and competences specified in the *CEFR*, which supports the use of the communicative approach (Council of Europe 2001: 13):

> Communicative language competence can be considered as comprising several components: linguistic, sociolinguistic and pragmatic. Each of these components is postulated as comprising, in particular, knowledge and skills and know-how.

Thus, the tagging system elaborated for the present study included two separate parts, since the aim was to help identify errors related to pragmatic misconceptions, as well as errors related to grammar. The first part of the grid was based on the descriptors included in the *CEFR* regarding pragmatic competences, which are described by the Council of Europe (2001: 13) as follows:

> *Pragmatic competences* are concerned with the functional use of linguistic resources (production of language functions, speech acts) drawing on scenarios or scripts of interactional exchanges. It also concerns the mastery of discourse, cohesion and coherence, the identification of text types and forms, irony, and parody.

First of all, the guidelines of the *CEFR* (Council of Europe, 2001) were summarised and abbreviated in order to create a simple table which could facilitate the tagging and direct identification of pragmatic errors in language learning, specifically those related to the use of English. For this particular piece of research, the focus was placed on Grice's maxims, described in Section 1 above, and on the various *CEFR* descriptors referring to pragmatic competences. The recommendations and specifications about pragmatics included in the *CEFR* were gathered; Table 1 shows the tags proposed for the tagging of pragmatic errors:

Item	Descriptors		Error	Tag
Rhetorical effectiveness	Quality (Try to make your contribution one that is true)	Try new combinations to get message through	Rhetorical effectiveness Getting the message through	RHQ1 RHQ2
		Explain main points	Main points	DSFocus
		Be precise	Precision in the text	RHP
		Sufficient vocabulary		
	Quantity (Make your contribution as informative as necessary, but not more)	Use circumlocution and paraphrases	Accuracy in communication	RHAC
		Explain in own words		
	Relevance (Do not say what is not relevant)	Be precise and concise	Focus on topic	RHF
	Manner (Be brief and orderly, avoid obscurity and ambiguity)	Confine message to what s/he can say	Adequacy to own language limitations	RHA
		Correct discourse		
		Get feedback: ask for confirmation		

Table 1. Items for error analysis based on the *CEFR* and Grice's maxims

The grid shown in Table 1 describes the competence the student should achieve, i.e., rhetorical effectiveness. The second column from the left contains the specific maxims under observation as described by Grice (1975), i.e., quality, quantity, relevance and manner. The next column specifies what kind of skill was expected from the learner, e.g., "Trying new combinations to manage to get his or her message through" or "Explaining in his or her own words". Then, the specific errors students make in this particular area are identified and the tags that are used to mark them (RHQ1-2, RHP, RHF, etc.) are also established.

Some examples of the resulting annotation can be seen in (1) – (7):

(1) We send you a relation of our hotels around the world that you can choose either. <RHQ1>
(2) Will they <GSS> can interact? <RHQ2>
(3) The best of this trip was to met Italy and her cities especially the town called Luca, in this town the people go to the places with bike and all the town it's whole of little shops very interesting. <DSFocus>
(4) The day of the farewell maiden Kat get drunk and she decided that she wants to got Nick, so she stops at an ATM to take some money to pay Nick. <RHP>
(5) Life is beautiful is a film that relates the life of a Jewish family at the time of the Nazis. The protagonist Guido (Roberto Bernini) and Dora his wife have a child. <RHAC>
(6) Life is beautiful is a film that relates the life of a Jewish family at the time of the Nazis. <RHF> The protagonist Guido (Roberto Bernini) and Dora his wife have a child.
(7) In my opinion, college students today have changed Bologna because the plan requires students to attend classes and must pass the courses in the academic year. <RHA>

After this, our attention turned to the way in which grammar is viewed in the *CEFR* in order to complete the second part of the grid, which was designed to identify and classify grammatical errors that may be paired with pragmatic errors. Traditionally, grammar has been included within the linguistic competences necessary to obtain a given level of proficiency, as we have explained in the previous section, and the *CEFR* adheres to this tradition (Council of Europe 2001: 13):

> *Linguistic competences* include lexical, phonological, syntactical knowledge and skills and other dimensions of language as system, independently of the sociolinguistic value of its variations and the pragmatic functions of its realisations.

The grid for the tagging of grammatical errors was a basic part of the design of this analysis and was drawn up regardless of the texts produced by the students and prior to any correction or assessment of the written production. All the issues included in the grid for the tagging of grammatical errors were those recommended by the Council of Europe (2001) and were used to identify and standardise the criteria for the identification and classification of errors, as shown in Table 2.

Item	Descriptors	Error	Tag
Grammatical competence expected	Grammatical accuracy in familiar contexts (syntactic and lexical errors)	Grammatical errors in simple sentences: formation of words, word order, verb tenses, articles, adverbs, voice, auxiliaries	GSS
	Repertoire of routines and patterns associated with more predictable situations	Wrong patterns (infelicities in reproducing the target language)	GP

Table 2. Items for grammatical competence

In the grammatical grid shown in Table 2, the list is reduced to two broad types of errors: accuracy in familiar contexts related to grammatical competence and pattern reproduction. These two items could lead to errors in simple sentences and infelicities in pattern reproduction (GSS and GP).

Some examples of the resulting annotation can be seen in (8) and (9):

(8) At the heart of this story is the question <GSS>
How anyone learnt the things about life and love? <GSS>
(9) Hopefully accept our apologies <GP>

Two further stages of work on the texts provided by the students took place subsequently: the collection and the processing of the data. For data collection, three issues were taken into account: the level of ability of the students who produced the texts, the text types included in the analysis and the errors which had been produced by the students. The texts were marked and corrected using the grids shown in Tables 1 and 2 based, as explained above, on the particular descriptors and text types proposed by the *CEFR*. The study presented here considered the texts produced by 90 students enrolled on the Tourism degree with a B1 level of proficiency over three academic years, from 2008 to 2011. The level of language proficiency of the students was established by means of placement tests assessing writing, listening, speaking and reading skills, with specific attention to their grammatical and pragmatic competence.

The *CEFR* suggests a series of text types as useful materials in the classroom: newspapers, instruction manuals, leaflets, personal letters, and so on. The corpus was made up of 206 texts based on such materials, consisting of three text types: narrations and summaries, opinions and formal writings. The distribution of the texts was as follows: 68 texts

belonged to the first group, 74 to the second and 64 to the third. The total sample length was 58,092 words. The samples were collected from the written assignments sent in electronic format by the students enrolled on the Tourism degree. The texts were monitored for plagiarism and students were asked to upload the writings onto the platform used for this subject during class time. A greater number of texts was compiled during this period than those stated above, but we only took into accounts the texts of students who had a B1 level for this study, since, apart from trialling the tagging system, we were also interested in determining the errors associated with a particular level of language proficiency, so that the writings produced would be classified according to the level the students demonstrated in the entry exam.

The next stage concerned the processing of the data. The results of the study were analysed and processed. The texts were codified according to the year of production and the text type. Three raters participated in the tagging process. The raters manually corrected the corpus and inserted the tags into the text file (see Tables 1 and 2). The taggers were not native speakers of English, as recommended by Dagneaux, Dennes and Granger (1998: 165). They were Spanish teachers of English with a very good knowledge of English grammar and pragmatics, which was considered essential for the activity of tag assignation. When the tagging was complete, the error-tagged student texts were analysed. The different errors were counted and the results inserted into the proposed grids. The raters observed several coincidences in the tagging of grammatical and pragmatic errors and remarked upon the cases in which this occurred. They also included a categorisation of errors in terms of the source of error (mother tongue interference), but this was rejected because it may introduce subjectivity (Dagneaux, Dennes and Granger 1998: 166 and Díaz-Negrillo and Fernández-Domínguez 2006).

The results obtained for the tagging of grammatical and pragmatic errors were analysed and compared in order to reveal whether there was any possible correlation between the learning of grammar and of pragmatics at a B1 level of proficiency in English. The percentages of the results were calculated in order to observe the discrepancies in the results. No statistical analysis was included in this study as our purpose was to propose a tagging system for pragmatic errors following the competences included in the *CEFR* for B1 level and observe the coincidences with the tagging of grammatical errors. The main aim of this study was to highlight the errors that can be tagged in these two categories and thus demonstrate that they should be addressed by means of the same learning strategies, since grammar and pragmatic competences underlie errors of the same kind.

3. RESULTS

The results extracted after the analysis of the corpus can be observed in Figure 1. We obtained more occurrences due to the simple structure of the sentences written by students. The high number of occurrences may be due to the level of the students involved in this study, who could not construct complex sentences as their competences were not sufficient to do so.

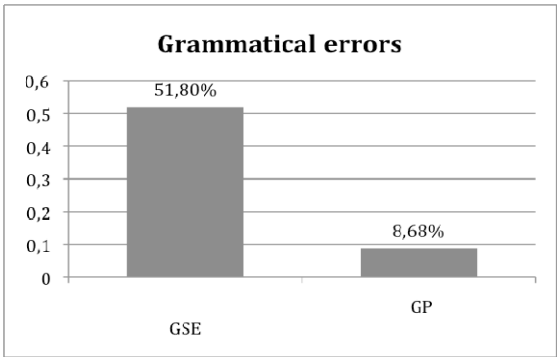


Figure 1. Grammatical errors

The total amount of errors found in this category was 3,444. We noticed that the students tended to prefer simple sentences, as they did not feel competent enough to write complex sentences. The least frequent errors were those related to the use of the wrong patterns.

In order to contrast the tags obtained after the analysis of the texts by the raters, we studied the errors tagged for pragmatic reasons, i.e., the rhetorical effectiveness of students, and the results are shown in Figure 2:

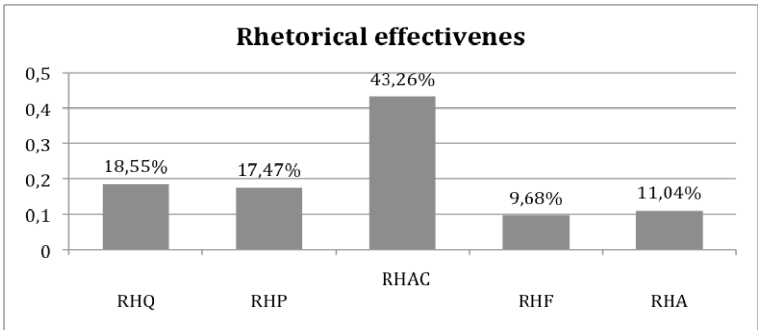


Figure 2. Results of pragmatic errors

The total number of errors tagged in this case was 1,105. The errors due to accuracy were the most frequent, followed by the errors due to rhetorical effectiveness and precision. After the analysis of the results, we considered whether some of the grammatical errors were due to grammatical or pragmatic causes. We also noticed that the grids proposed for the tagging of grammatical and pragmatic errors were useful for raters and all the errors found in the analysis could be placed into the grid. The initial proposal of the raters to identify errors due to the influence of the mother tongue was discarded, as this was considered a cause of errors rather than a possible classification for the tagging system. After contrasting the tagging of the raters, we looked for any coincidence between the two tagging systems, with the raters tagging one error as both a grammatical and a pragmatic error. The coincidence of the grammatical and pragmatic errors analysed can be seen in Table 3:

GRAMMATICAL ERRORS	PRAGMATIC ERRORS
Wrong patterns	Rhetorical effectiveness Accuracy
Grammatical simple sentencing	Accuracy Rhetorical effectiveness Adequacy to own limitations Focus Precision

Table 3. Coincidence between grammatical and pragmatic tagging

We noted that the grammatical errors classified by the raters as the wrong patterns coincided with the pragmatic errors regarding rhetorical effectiveness and accuracy. We also observed that the errors caused by writing simple sentences were classified by the raters as pragmatic errors due to lack of accuracy, rhetorical effectiveness, adequacy to own limitations and focus. Some examples of the coincidence of error tagging can be seen in (10), (11) and (12):

(10) We download the car and to sleeping<GP>, since the following day it was waiting for us our first day of ski in europa's bigest <GP>ski resort. <RHAC>
(11) All the film, the protagonist talks the reasons <GSS>of these answers. These histories are success<GSS> that he lives with his friend when there are children's. <RHQ1>
(12) His pizzeria restaurant will be served >GSS>all types of foods, meats, fish, and as no pizzas, <GSS>that are the speciality of the house. <RHP>

As can be observed in the examples, the raters tagged the errors twice as they considered that, in these cases, the errors were both of a pragmatic and a grammatical nature. The nature of the errors is grammatical most of the times, but when students write a sentence, errors also entail poorness of pragmatic competence. Students should acquire grammatical competence in order to produce rhetorically adequate language.

4. CONCLUSIONS

Initially, a proposal for the tagging of grammatical errors was undertaken in this study and it was observed that some grammatical errors were implicit in pragmatic errors. The elaboration of an analysis grid proved to be a valuable tool as it allowed us to detect the coincidences of the grammatical and pragmatic errors in our corpus. Raters were able to use the grid to facilitate the tagging of errors and to classify them depending on their nature. Furthermore, the methodology we propose for the detection and classification of errors demonstrates that grammatical and pragmatic competences interact intensively in various ways in second language learning. We believe that error tagging is a powerful tool to

determine language proficiency and the stages of language learning and development. As Dagneaux, Dennes and Granger (1998: 173) have said, error tagging "can be used to generate comprehensive lists of specific error types, count and sort them in various ways and view them in their context and alongside instances of non-errors".

In this study, the results obtained enable us to propose some guidelines for the avoidance of errors in written language. First, grammar should be considered as a basic part of communicative competence. Second, grammatical errors in simple sentences and accuracy errors in pragmatics are the issues on which the focus should be placed, as they are linked in effective communication, as we have shown in Table 3.

Grammar and pragmatics should be defined as two separate but not independent components of a theory of language that seeks to model grammatical and pragmatic competences. We propose a tagging classification (shown in Tables 1 and 2) that could be useful to detect and classify pragmatic and grammatical errors, but the results shown in Table 3 should be taken into account. In addition to assuming their close interaction in contexts of language use, it is important to note that certain aspects of grammar and pragmatics are inextricably linked to each other and this may have a significant bearing on learners' ability to achieve communicative competence. Ariel (2008: 1) states: "Any specific instance of language use is neither wholly grammatical nor wholly pragmatic. To pick deixis again, it combines grammatical aspects (there is a grammatically specified difference between *I* and *this*) with pragmatic aspects (pinning down who the speaker is, what object this denotes)". This means that grammar is responsible for what speakers express explicitly and pragmatics explains how speakers infer additional meanings, in this sense, one aspect is embedded in the other.

Communication does not simply consist of packing thoughts or ideas into the form of words so that the reader can unpack and understand them. In order to reach out to readers and grasp their attention, writers need to link their text with whatever background information readers may possess. The most fundamental task of a pragmatic theory is to explain how the intended context is recognised, that is, how the reader is able to work out which of all the assumptions available to his cognitive system at any given time is the set that he/she is intended to use in processing the utterance.

While grammar is responsible for what we express explicitly, pragmatics explains how we infer additional meanings. The problem is that it is not always a trivial matter to decide which of the meanings conveyed is explicit (grammatical) and which is implicit (pragmatic). The study of pragmatics and grammar should enable a methodology to be constructed whereby the two can be distinguished. Grammar and pragmatics are combined in natural discourse and, as a consequence, pragmatic uses become grammatical in time.

Nevertheless, we are conscious that further work and a degree of specification are necessary in order to examine pragmatic and grammatical issues. In future studies, our aim is to propose further tagging systems for errors depending on the level of students' language proficiency and to design a taxonomy of errors classified by level. Furthermore, an examination of the correspondence between different error classifications in second language acquisition, such as that between grammatical errors, lexical errors, pragmatic errors and cognitive errors, could be of interest.

REFERENCES

Aguado-Jiménez, Pilar, Pascual Pérez-Paredes and Purificación Sánchez. 2012. Exploring the use of multidimensional analysis of learner language to promote register awareness. *System* 40: 90–103.

Archer, Dawn, Karin Aijmer and Anne Wichmann. 2012. *Pragmatics. An advanced resource book for students.* Oxon: Routledge.

Ariel, Mira 2008. *Pragmatics and grammar*. Cambridge: Cambridge University Press.

Bardovi-Harlig, Kathleen. 1996. Pragmatics and language teaching: bringing pragmatics and pedagogy together. In Lawrence F. Bouton (ed.), *Pragmatics and language learning*. Urbana, IL: University of Illinois at Urbana-Champaign, 21–39.

Bardovi-Harlig, Kathleen. 1999. The interlanguage of interlanguage pragmatics: a research agenda for acquisitional pragmatics. *Language Learning* 49: 677–713.

Bardovi-Harlig, Kathleen 2013. Developing L2 pragmatics. *Language Learning* 63: 68–86.

Bardovi-Harlig, Kathleen and Zoltán Dörnyei. 1998. Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning. *TESOL Quarterly* 32/2: 233–262.

Carrió Pastor, María Luisa. 2004. Las implicaciones de los errores léxicos en los artículos en inglés científico-técnico. *RAEL: Revista Electrónica de Lingüística Aplicada* 3: 21–40.

Carrió Pastor, María Luisa. 2005. *Contrastive analysis of scientific-technical discourse: common writing errors and variations in the use of English as a non-native language.* Ann Arbor, MI: University of Michigan.

Corder, Stephen Pit. 1967. The significance of learner's errors. *IRAL* 5/1-4: 161–170.

Council of Europe. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Dagneaux, Estelle, Sharon Dennes and Sylviane Granger. 1998. Computer-aided error analysis. *System* 26: 163–174.

Díaz-Negrillo, Ana and Jesús Fernández-Domínguez. 2006. Error tagging systems for learner corpora. *RESLA: Revista Española de Lingüística Aplicada* 19: 83–102.

Díaz-Negrillo, Ana and Salvador Valera-Hernández. 2010. A learner corpus-based study on error associations. *Procedia – Social and Behavioral Sciences* 13: 72–82.

Dulay, Heidi C., Marina K. Burt and Stephen D. Krashen. 1982. *Language two.* Oxford: Oxford University Press.

Ellis, R. 1994. *The study of second language acquisition.* Oxford: Oxford University Press.

Ellis, Rod, Shawn Loewen and Rosemary Erlam. 2006. Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition* 28/2: 339–368.

García Velasco, Daniel and Carmen Portero Muñoz. 2002. Understood objects in functional grammar. *Working Papers in Functional Grammar* 76: 1–24.

Granger, Sylviane. 2002. A bird's eye view of learner corpus research. In Sylviane Granger, Joshep Hung and Stepahnie Petch-Tyson (eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins, 3–33.

Granger, Sylviane. 2003a. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* 37/3: 538–546.

Granger, Sylviane. 2003b. Error-tagged learner corpora and CALL: a promising synergy. *CALICO Journal* 20/3: 465–480.

Grice, H. Paul 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and semantics. Volume 3: speech acts*. New York: Academic Press, 41–58.

Hasbún Hasbún, Leyla. 2004. Linguistic and pragmatic competence: development issues. *Filología y Lingüística* 30/1: 263–278.

James, Carl. 1998. *Errors in language learning and use.* London: Longman.

Kasper, Gabriele. 2010. Four perspectives on L2 pragmatic development. (Revised version of a plenary given at the annual conference of the American Association of Applied Linguistics (AAAL), Vancouver, March 2000). <http://www.nflrc.hawaii.edu/networks/NW19/NW19.pdf> (12/06/2013).

Levinson, Stephen C. 2000. *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Lightbown, Patsy M. 1992. Can they do it themselves? A comprehension-based ESL course for young children. In Robert Courchêne, Jennifer St John, Christiane Thérien and John Glidden (eds.), *Comprehension-based second language teaching*. Ottawa: University of Ottawa Press, 353–370.

Lyster, Roy. 1998. Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classroom. *Language Learning* 48: 183–218.

Mestre Mestre, Eva María. 2011. *Error in the learning and teaching of English as a second language at higher education level.* PhD. Valencia: Universitat Politècnica de València.

Mestre Mestre, Eva María and María Luisa Carrió Pastor. 2012. A pragmatic analysis of errors in University students' writings in English. *English for Specific Purposes World* 35/12.

Mey, Jacob L. 1993. *Pragmatics. An introduction*. Blackwell: Oxford.

Németh, T. Enikő and Károly Bibok. 2010. Interaction between grammar and pragmatics: the case of implicit arguments, implicit predicates and co-composition in Hungarian. *Journal of Pragmatics* 42/2: 501–524.

Pérez-Paredes Pascual and Pascual Cantos-Gómez. 2004. Some lessons students learn: self-discovery and corpora. In Guy Aston, Silvia Bernardini and Dominic Steward (eds.), *Corpora and language learners*. Amsterdam: John Benjamins, 245–258.

Rose, Kenneth R. and Gabriele Kasper. 2001. *Pragmatics in language teaching.* Cambridge: Cambridge University Press.

Rumelhart, David E. and James L. McClelland. 1986. PDP models and general issues in cognitive science. In David E. Rumelhart, James L. McClelland and the PDP Research Group (eds.), *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press, 110–146.

Rutherford, William and Michael Sharwood Smith. 1985. Consciousness-raising and universal grammar. *Applied Linguistics* 6/3: 274–282.

Schaeffer, Jeannette. 2005. Pragmatic and grammatical properties of subjects in children with specific language impairment. *UCLA Working Papers in Linguistics* 13: 87–134. <http://www.linguistics.ucla. edu/faciliti/wpl/issues/wpl13/papers/Jeannette.pdf> (12/06/2013).

Schaeffer, Jeannette. 2011. Grammar and pragmatics in specific language impairment and autism spectrum disorder. <http://www.uva.nl/en/about-the-uva/organisation/staff-members/content/s/c/j.c.schaeffer/ j.c.schaeffer.html> (12/06/2013).

Sperber, Dan and Deirdre Wilson. 1995. *Relevance: communication and cognition*. Oxford: Blackwell.

Swain, Merrill. 1985. Communicative competence: some roles of comprehensible input and comprehensible output in its development. In Susan M. Gass and Carolyn G. Madden (eds.), *Input & second language acquisition*. Rowley, MA: Newbury House, 235–253.

Terrell, Tracy David. 1991. The role of grammar instruction in a communicative approach. *Modern Language Journal* 75/1: 52–63.
Verschueren, Jef. 1999. *Understanding pragmatics*. New York: Arnold.
Wang, Man-liang. 2007. Pragmatic errors in English learners' letter writing. *Sino-US English Teaching* 4/2: 39–43.
Wilson, Deirdre and Dan Sperber. 1998. Pragmatics and time. In Robyn Carston and Seiji Uchida (eds.), *Relevance theory: applications and implications*. Amsterdam: John Benjamins, 169–186.

RiCL Research in Corpus Linguistics

New media, new challenges: exploring the frontiers of corpus linguistics in the linguistics curriculum

Nuria Hernández[1]
Universität Duisburg-Essen / Germany

Abstract – This paper introduces a new corpus of computer-mediated communication which is currently being compiled at the University of Duisburg-Essen. Based on the experience from this project, the paper also discusses the possibility of implementing major issues in corpus construction into the academic curriculum of young linguists in the form of project-based learning. A variety of new challenges and possible solutions regarding the compilation and processing of new media language are presented.

Keywords – blogs, CMC, DMC, emoticons, Facebook, image boards, new media, project-based learning, SMS, Twitter, YouTube

1. Introduction

In recent years, the study of language variation and change has extended to include a new and thrilling area of research: the language used in digital media. Research into 'digital discourse', 'computer-mediated communication', 'Internet language', 'Netspeak' and 'Textspeak' (Crystal 2004, 2006, 2010; see Herring 2007 for a more fine-grained classification scheme) is concerned with newer forms of correspondence, such as e-mails, chat, SMS or blogs, and more recently with the wide array of social media facilitating a fast and global exchange of user-generated content. While some people are concerned that the new technology may have a negative impact on language use, others argue the reverse, saying that the new media have encouraged a dramatic expansion in the creative capacity of language (for example Crystal 2006: 275).

The corpus presented in this paper provides an empirical basis upon which these assumptions can be tested. Despite the fact that linguistic research in computer-mediated communication (CMC) is growing at a fast pace, corpus-linguistic studies in the field often cite project-related corpora which are not readily accessible to the linguistic community (cf. Beißwenger and Storrer 2008). Examples would be the *CoSy corpus* (Yates 2001) or the *Swiss German Webchat Corpus* (Siebenhaar 2006). Databases for general use, on the other hand, include corpora as varied as, for instance, the *Dortmunder Chat-Korpus* (http://www.chatkorpus.uni-dortmund.de) or the *Enron Email Dataset* (http://www-

[1] I would like to thank all of my students who contributed to the first version of the DMC during the academic winter term 2011–2012 and thereafter. This project would not have been possible without their scrutinising questions, enthusiasm and their valuable input and ideas. For the 'Blogs' component: Linda Bleyer, Mark Elpers, Dominik Nebel, Yvonne Willuhn and Frauke Witt. For 'Image Boards': Sean Sams. For 'SMS German': Axel Bund, Seda Kiraç and Sebastian Krebs. For 'SMS English': Kate Roberts, Fiona Seward and Seán Upton. For 'Twitter': Catherina Hofmann, Melike Inan and Sabine Lange. For 'Facebook': Josua Ehmann, Lena Raue, Tina Terlinden and Nadine Vangenhassend. For 'Youtube': Julia Daitche, Shaun Hughes, Julian Kloss, Maria Laura Salerno and Ganna Strashnenko.

2.cs.cmu.edu/~enron). At present, however, no multi-genre corpus is available which could be used for research purposes as well as for the teaching of corpus-linguistic methods.

The need for a detailed linguistic investigation of current developments in CMC and the shortage of available data make a strong argument for a new corpus. This led to the project *Digital Media Corpus* (*DMC*), which was first presented at CILC2012, under the supervision of the current author. At the beginning of the project it was decided to realise this task within the framework of a graduate linguistics seminar in order to give students the chance to explore the world of corpus linguistics from a different angle, using a problem-oriented, project-based approach (Wrigley 1998; Stoller 2002). Viewed in this light, the lack of available data is an opportunity for exploring new frontiers.

The current paper reports on experience drawn from the *DMC* project which could prove useful for future experiments in the field, including the multiple challenges faced in the processing of different CMC genres, or 'socio-technical modes' (cf. Herring 2002; 'genre' and 'mode' will be used interchangeably in this paper). At present, the *DMC* comprises over 104,000 words from weblogs (blogs), image boards, SMS, *Twitter*, *Facebook* and *YouTube*, in English and German. The components differ in size and each component looks slightly different due to compositional differences between the source texts (e.g., 'text only' vs. 'text + pictures'; more details in section 3). Nevertheless, the preliminary version presented in this paper is a first milestone in working towards a consistently formatted database that will be made freely available for linguistic studies on CMC.

2. THE PROJECT

The *DMC* project began in the winter term of 2011, with an advanced seminar called 'Language in the New Media' for students in their third or fourth year of studies in English Literature and Linguistics (teacher education programme and bachelor's degree). The ultimate reason for including such a project in the linguistics part of the curriculum was to explore a different way of teaching corpus linguistics. An additional appeal of digital modes such as blogs, SMS, *Facebook* or *Twitter*, was that they are frequently used by the students and teachers themselves, often on a daily basis, thus adding a valuable emic perspective to their investigation. From a linguistic point of view, the spontaneity and the high level of emotivity in these modes promise a highly idiomatic and less self-monitored use of language which is difficult to elicit by other means.

While introductions to corpus linguistics tend to focus on the discussion and analysis of already existing databases, the aim of this seminar was to confront students more directly with the problems usually faced by corpus linguists themselves. Starting from scratch, the compilation of a completely new corpus provided ample opportunity for active discussion and decision-making. Unlike in previous seminars, the analysis of linguistic features was not realised in class, but was outsourced to subsequent term paper projects in order to allocate more time to the acquisition of corpus-compilation skills and data awareness. The results described in the following sections will therefore mainly refer to data compilation and processing; the advantages of the present approach for students analysing CMC language, and its comparability with other didactic approaches, will be touched on in section 5.

The overall time frame of the seminar consisted of weekly 90-minute classes in one of the university's computer pools, over a period of fourteen weeks. All participants had basic computer proficiency, but no previous experience in data collection or corpus compilation. After a general introduction to corpus linguistics, the theoretical issues relating to the changing nature of text (Ferrara, Brunner and Whittemore 1991; Crystal 2010), the different technical and social factors influencing CMC language (Yus 2011), as well as different CMC resources (websites, journals, dictionaries) in weeks 1 through 3, the project was divided into the following steps and goals, each under the guidance of the lecturer as the project supervisor.

- Planning and organisation (weeks 4 and 5)
  The first step consisted in the specification of the overall task and goals, along the lines of an "ill defined task with a well-defined outcome" (Capraro and Slough 2009), and the assignment of collaborative workgroups with up to 5 students per team, each group focusing on one CMC genre chosen by the students themselves; the only selection restrictions were copyright and privacy concerns (e.g., informed consent in the case of SMS); the result was a general corpus structure with 6 individual components; the individual teams decided how to proceed with collecting the respective data.
- Cooperation and creation (weeks 6 through 8)
  Raw data were collected by the different teams between the sessions, as an on-going home assignment, followed by partially supervised data processing in class (transcription and tagging); each seminar session started with an open discussion of issues relating to the textual markup and the tagging of special symbols and icons found in the different modes; step by step, a common tag list was generated for the entire corpus; a common text header format with text and user variables for all components was devised; the corpus was given a name.
- Control and reflection (weeks 9 and 10)

This stage consisted in mutual proofreading, feedback and correction of text files across the teams, in class and outside of class; the strategies chosen by each team were revised by another team, in some cases leading to major changes in the markup.

- End product to share (weeks 11 through 14)
  The project concluded with the collective writing of the corpus manual; the introductory part was written by the lecturer, and subchapters about the individual components were written by the student teams, once more in and outside of class; at the end of the seminar, students were offered the opportunity to further explore their data in a term paper focusing on the language used in the corpus, and to continue being involved in the corpus project.

In order to achieve the goals set out at the beginning, a variety of challenges had to be addressed, due in part to the fact that the corpus was being compiled from scratch in a self-motivating approach, and due in part to common issues in empirical linguistics, such as the protection of the authors' privacy.

The greatest challenge lay in formatting texts from different CMC genres. Although the up-and-coming research area of computer-mediated communication has attracted growing attention over the last few decades, linguistic publications and information on corpus design are still scarce (e.g., the electronic journal *Language@Internet*). For some modes, such as image boards, no corpora or linguistic studies existed at all, which put the respective students in the role of linguistic pioneers – in some cases enlivening their enthusiasm, in others deflating their confidence. Even well-researched modes, such as SMS or blogs, posed some open challenges. How should we tag special symbols and emoticons in a consistent, machine-readable format? What should we do with the many colloquial expressions, non-standard abbreviations and creative uses of language found in these new media? How, for example, would one tag a mixed-code expression such as *4tel 4 4* (German *viertel vor vier*; see section 4.7)? Should references to pictures and other websites be included in the transcripts? And, last but not least, how can user variables such as age, sex and origin be retrieved in media used by a largely anonymous global community? These are only some of the questions that had to be addressed. Before we look at possible solutions in more detail, a brief description of the corpus itself is in order.

3. THE *DMC* CORPUS

3.1. General structure

This section gives a brief introduction to the overall structure of the corpus and some special properties of its components. At present, the *DMC* contains approximately 104,200 words from 216 transcripts in 6 individual components: 'Blogs', 'Image boards', 'SMS' (English and German), '*Twitter*', '*Facebook* posts' and '*YouTube* comments'. For the time being, e-mails were not included because of the difficulties that this medium presents for the definition of 'text', due to partial text deletion, framing and intercalation of responses (cf. Crystal 2011). However, the multi-genre design of the corpus would allow a later inclusion, which could also make an intriguing topic for a future installment of the course.

The word counts of the individual components seen in Table 1 differ considerably, due to characteristic differences between genres (for example, short text messages vs. long text passages in 'Blogs').

COMPONENT	TEXT ID	TEXT FILES	WORD COUNT[2]
Blogs	BLG	3	30,800
Image boards	IMB	12	7,300
SMS, German	TXT…G	69	5,000
SMS, English	TXT…E	73	2,900
Twitter	TWT	7	43,800
Facebook posts	FBP	25	1,500
YouTube comments	YTC	27	12,000

Table 1. *DMC* components and word counts (June 2012)

Differences between the components also become visible in the directory structure. Figures 1 and 2, for instance, show the directory structures of the 'Blogs' and '*Twitter*' components. In 'Blogs', the folder for each blog contains a text file (tagged transcript), as well as the different pictures from the original website (also compare Figure 4 below), whereas '*Twitter*' contains text files only.

[2] Approximate word count, excluding text headers and tags.

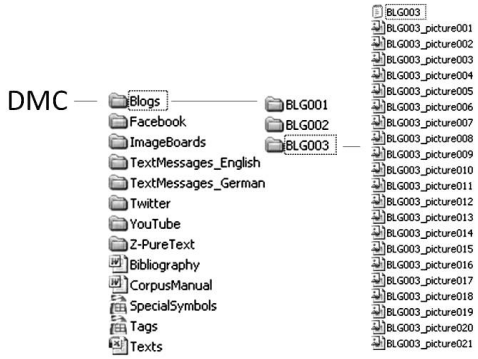


Figure 1. Directory structure, 'Blogs' component

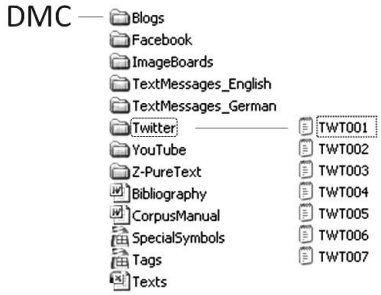


Figure 2. Directory structure, 'Twitter' component

During the collection process (November 2011 through January 2012), all data extracted for the different components were transformed into plain text files, and special symbols, icons and emoticons were marked with tags as seen in the examples below. Each transcript in the corpus was given a file name composed of a component ID (BLG for 'Blogs', IMB for 'Image boards', etc.), followed by a three-digit running number (BLG001, BLG002...), and each transcript was preceded with a text header containing the basic text and user variables.

3.2. The different components

3.2.1. Blogs

Weblogs, or blogs, are personal online journals of individual users or small groups which have enjoyed great popularity since the late 1990s. Unlike synchronous modes where all participants are online at the same time (e.g., Internet Relay Chat), blogs are less 'conversational' and, therefore, often perceived as closer to the written end of the written-spoken continuum (cf. Peterson 2011).

The current version of our corpus contains three different blogs with three different topics. Since the primary focus in the collection process was on language data, the topics were not a decisive factor; the students simply chose blogs they were familiar with.

Each blog transcript starts with a header containing the file name, the blog URL, the user's name, the language used, the posting time, the user's age and sex, and the general topic of the blog. Because of the many pictures occurring in the blog posts, and because of the fact that users frequently refer to the pictures in the text, it was decided that each blog be given its own folder containing the transcript as well as the corresponding picture files (compare Figures 1 and 4).

Blog	URL	Users	Age	Word count (approx. tokens)
BLG001	delicatehummingbird.blogspot.com	female	26	10,000
BLG002	gofugyourself.com	female, female	unknown, unknown	12,900
BLG003	dooce.com	female	36	7,900

Table 2. Blogs in the *DMC* (June 2012)

3.2.2. Facebook *posts*

Launched in 2004, *Facebook* has become the most popular social network worldwide. According to information provided on *Facebook*'s website, over 650 million people are said to be currently using the network on a daily basis (*Facebook* 2013a). Its mission is "to give people the power to share and make the world more open and connected" (*Facebook* 2013b). *Facebook* users may upload pictures, share links and videos and connect with friends all over the world. All users can comment on any content added by their friends, a special feature being the option to signal approval of another user's comment or content by giving it a 'thumbs up'.

The data collected for the *DMC* consists of comments which the students themselves, as *Facebook* users, had previously posted in reply to other users' status reports. Each transcript presents one 'conversation,' starting with a status update by one user and the subsequent posts responding to this update (see example (3)). Threads which contained links or pictures were not included. Since status reports basically describe what is on the user's mind, some posts can be confusing or do not seem to make much sense to someone who is not immediately involved in the exchange. The reader of a *Facebook* post does not necessarily know the context of the respective entry and commenters are in no way obliged to explain themselves.

The current version of the *DMC* contains 24 *Facebook* transcripts in German, but other languages, including English, could be added at any time.

3.2.3. Image boards

Image boards are a kind of bulletin board system, much like a public chat room, where users can create threads on different topics. Originally invented in Japan, image boards have been copied in other countries, especially in the United States. The most famous image board at present is *4chan*, which stars among the top 900 most visited websites with up to 450,000 postings per day. The main language in image boards is English, but any user may start a thread in another language.

The hallmark of this medium is its total anonymity. All image board users are anonymous, to the extent that even nicknames are avoided, and anybody can read any uploaded post. Instead of official registration, image boards use tripcodes which contain no user details. In addition, the threads are extremely short-lived and often deleted after one or two hours, making them the least persistent contributions with the, assumedly, least meta-linguistic awareness in the corpus (cf. Herring 2007: 15). By saving the data, our project breaches this policy to some extent, but anonymity remains guaranteed in the transcripts.[3]

Currently, the 'Image boards' component of the *DMC* contains 12 text files with over 7,300 words. In this mode, too, posts are often accompanied by pictures which comment on the written text in some way. In fact, discussions are highly graphic-centric, often initiated by posted images which can have follow-up pictures posted as responses. Researchers should note that these threads are possibly incomplete, since posts can be deleted after the image limit has been reached and extremely long threads were only partially extracted.

3.2.4. SMS

For SMS, as for most of the other modes described in this paper, no linguistic corpus was publicly available when the project started. So far, this component contains messages in English and German, with the addition of further languages being planned. The total word count currently amounts to almost 5,000 for German, and 2,900 for English (excluding text headers and tags). A first example of the brief messages sent between (mobile) phones and other devices is shown in Figure 3, followed by further examples below.

SMS are usually short, and individual exchanges do not go on for very long. Together with *Facebook* posts, these data are the most difficult to obtain, since they are generally perceived as more personal than other CMC modes.

---

[3] Complaints against the use of these data should be directed to the author; they will be taken seriously.

<Simone> Hey Anja, ja war leider in der uni.. Bin total fertig! Ich hoffe zumindest, dass ihr gestern alle spass hattet! [reg=xxx] kisses [\reg]

Figure 2. Original SMS on mobile phone screen, plus transcript (*DMC*, TXT009G)

3.2.5. Twitter

The social networking service *Twitter* was created in 2006 as a medium for keeping in touch with both friends and the general public. *Twitter* enables its users to send and read text-based posts of up to 140 characters, known as *tweets*. The character limit was imposed to interface easily with text messaging services. In the last few years, this medium has been increasingly used by celebrities who enjoy regular contact with their fans and supporters, including singers, actors and politicians. This is also reflected in the *DMC* '*Twitter*' component, which contains original data from seven different *Twitter* accounts of various singers, such as Katy Perry's and Bruno Mars's. Each file in this component contains the tweets of the account owner and comments by various commentators (answers to the original tweets).

In order to use *Twitter*, one has to set up an account, including a username (usually a nickname) and a profile picture. *Twitter* is hence slightly less anonymous than the above-mentioned image boards and even age and gender are occasionally provided in the commentator profiles.

Collecting data for this medium was relatively easy—a fact which is reflected in the highest word count of 43,800; see Table 1.

3.2.6. YouTube *comments*

YouTube, a video-sharing website created in 2005, is the first address for many Internet users looking for free videos and music, including those who also want to share their thoughts and impressions with a larger community. *YouTube* language has been severely criticised as "[j]uvenile, aggressive, misspelled, sexist, homophobic" (Owen and Wright 2009), but so far such assumptions have not been tested on any empirical grounds. Corpora like the *DMC* can help close this gap.

In the corpus, the audiovisual material itself is not included, the focus being on the concurrent user comments. Assuming that comments on different topics might differ linguistically, the student team decided to include a range of topics in order to give a more balanced picture of *YouTube* language. At present, the '*YouTube*' component contains 27 different files with 6 different topics selected from the large variety discussed online: music, education, comedy, babies, politics and news stories. A first example of a '*YouTube*' file is shown in (4).

4. CHALLENGES AND RESULTS

4.1. User privacy

The first challenge that the students were confronted with during data collection concerned the users' privacy. The protection of user (speaker/author) privacy is a well-known issue in empirical linguistics, concerning especially those genres where the users themselves decide how much private details they give out and with whom they want to share their thoughts.

Two components in our corpus are especially affected by this issue: 'SMS' and '*Facebook* posts'. In these modes, most of the data was contributed by the team members themselves, i.e., their own text messages and posts from their own *Facebook* accounts, in agreement with the respective co-users. Despite the fact that the project was conducted in the Department of Anglophone Studies, this procedure resulted in both an English and a German SMS subcomponent, and predominantly German *Facebook* posts (which will hopefully be extended to English in the future).

As an additional protective measure in both components, the names of users who were not part of the research teams were made anonymous, and some messages or fragments of text which were considered to contain very personal information were deleted. Other user variables were kept, as seen in Table 3.

The privacy issue does not only concern the usernames. In any online genre there are users who prefer not to disclose their personal details, which makes the user variables less reliable than in other types of linguistic data. Especially the 'age' variable should always be taken with a grain of salt. It is virtually impossible to know how much one can trust the information extracted from the Internet, 'age' being particularly unreliable. In extract (4), for example, the *YouTube* user BeraSk8, one of the commentators on US rapper Dr Dre in YTC020, purports to be 111 years old—and he is only one of many alleged 100+ users on *YouTube*.

Before we continue with the next challenge, here are some examples of transcripts from different parts of the corpus. In German examples, the English translations are given in italics.

(1) SMS transcript, German
<text ID TXT016G> <language German> <date 112011>
<user Philip,Marvin> <user age 26,25> <user sex m,m> <native language German,German>

<Philip> Hi Philip. Kann ich morgen deinen Ghettoblaster ausleihen?
<Philip> Hi Philip. Can I borrow your ghetto blaster tomorrow?

<Marvin> das tut mir leid der ist ja nicht von mir sondern von unserem Team [reg = u] und [\reg] zurzeit nicht in meiner Gewalt!
<Marvin> I'm sorry it doesn't belong to me but to our team and it's currently not under my thumb!

<Philip> [reg = aso] ach so [\reg]. Dachte wäre deiner. OK
<Philip> I see. Thought it was yours. OK

(2) SMS transcript, English
<text ID TXT003E> <language English> <date 102011>
<user Sean,Barry> <user age 20,21> <user sex m,m> <native language English,English>

<Sean> some burn on the rugby but on the other hand we're all off to poland

<Barry> some burn [reg=alrite] alright [\reg] haha. what you going there for? train ya? what part you going to?

<Sean> man for the euros in the summer!!

<Barry> haha ya man it's all about the soccer team. they'll probably get [reg=bate] beaten [\reg] by armenia the way things are going.

(3) *Facebook* posts, German
<text FBP007> <language German> <posting time 112011>
<user Kordula,Becky,Zack,Mira,Gordon,Carla> <user sex f,f,m,f,m,f> <user age ,,,,,>

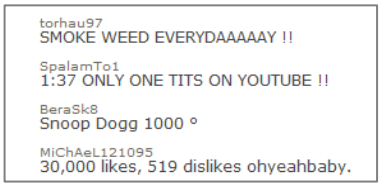
<Kordula> [26/11/2011 1:35pm]
ich will ans [emphcap] MEER [/emphcap]!!!!! Dicke Jacke, Gummistiefel, Schal, Mütze, Taschentücher, Geld für nen heißen Kakao und ab [reg=geeeehts] geht's [\reg]!
I want to go to the SEA!!!! Thick jacket, wellingtons, scarf, cap, tissues, money for a hot chocolate and off we go!

<Becky> [26/11/2011 1:36pm]
boah [reg= joo] ja [\reg], [reg= dat] das [\reg] [reg=wärs] wär's [\reg]
wow yeah, that would be great

<Zack> [26/11/2011 1:42pm]
wann [reg= solls] soll's [\reg] los gehen?
when do you want to go?

<Kordula> [26/11/2011 1:42pm]
hmmm. in [reg=ner] einer [\reg] stunde [em laugh] :D [\em laugh]
hmmm. in an hour :D
…..

(4) *YouTube* transcript plus screenshot



<text ID YTC020>
<topic music Dr Dre>
<language English>
<posting time 122010>
<user SpalamTo1,BeraSk8,…>
<user age 21,111,…>
<user country Poland,Brasil,…>
<URL http://www.youtube.com/watch?v=ejUARfOR7hE&feature=relmfu>

<user SpalamTo1>
<posting time 141210>
1:37 [emphcap]ONLY ONE TITS ON YOUTUBE[\emphcap]!!

<user BeraSk8>
<posting time 141210>
Snoop Dogg 1000[sym=°] degrees [\sym]
…..

4.2. Defining textual units

In his study "O brave new world, that has such corpora in it!", David Crystal remarks that, "[i]f there's one thing that unites all of us, in the field of corpus linguistics, it is that we assume we know a text when we see one" (Crystal 2011: 1). Unfortunately, but also intriguingly, this assumption is difficult to maintain when dealing with CMC genres like the ones discussed here. The traditionally definable properties of 'text'—such as spatial and temporal boundaries, and permanence—are hard to apply to newer media. The "stable, familiar, comfortable world" that corpus linguists once dealt with has changed, and research in digital discourse needs to rethink the notion of 'text' (Crystal 2011: 1).

In more concrete terms, we have to decide what to do with extra-textual elements, such as pictures, and textual elements that are not part of the main text or lead us to other texts, such as hyperlinks. During the compilation process, it soon became clear that the answers to these questions might vary, but the main criterion agreed upon by all student teams was that elements (both textual and extra-textual) should be included if, and only if, they are referred to in the main body of the text. In that respect, hyperlinks form part of the running text, but the texts to which they link do not.

Another question that had to be answered was at what point to cut off texts which lack the above-mentioned boundaries. Genres such as *Twitter*, for instance, have threads that can go on for a long time, often with extended intervals and, in most cases, these threads will continue after the collection of data for our project has ended. For the '*Twitter*' component, entire threads were obtained by clicking the 'all comments' view, on one specific date which is mentioned in the text header. This way, any thread could be chronologically extended in follow-up versions of the *DMC*.

4.3. Texts and pictures

In CMC genres such as blogs and image boards, pictures are regularly used to illustrate and comment (often humorously) or simply add visual impressions to the written text. In the texts themselves, these pictures are not always mentioned, but the connection is usually apparent. It was therefore decided, in both components, to include the pictures in the respective folders (see Figure 1), and to mark the original position of each picture with a *picture tag*.

The example in Figure 4 was taken from an American blog by an English native speaker, called dooce.com. The topics of this blog revolve around the author's everyday life, experiences and thoughts. Dooce.com has received numerous Weblog Awards for 'Best American weblog' (2005, 2008), 'Best-designed weblog' (2008), 'Weblog of the year' (2008), 'Most humorous weblog' (2005), 'Best writing of a weblog' (2005), and 'Lifetime achievement' (2008). In this example, the author writes about her dog, including pictures of him on the website. In the corpus transcript, these are indexed by consecutively numbered picture tags.

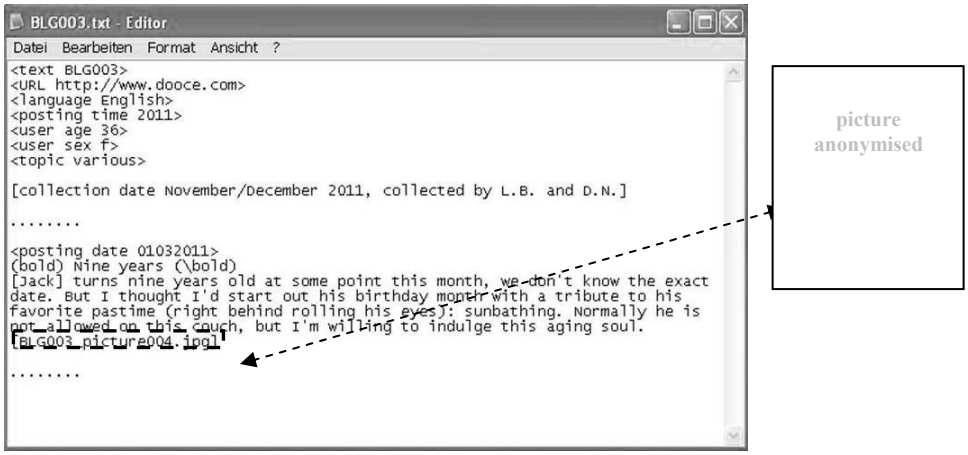


Figure 3. Blog transcript with picture tag (BLG003_picture004.jpg)

4.4. Consistent formatting

One of the goals of this project was to compile a consistently formatted CMC corpus for comprehensive analyses of new media language. In order to achieve this goal, multiple decisions had to be taken, once the data had been collected, in order to transfer them into a homogeneous format – always taking into account that the students had little or no experience in data processing.

First of all, it was decided that each transcript should be preceded by a header containing the basic text and user variables (using empty spaces for missing values). Due to differences in the accessibility of these variables, their number varies between the different genres, as shown in Table 3. In genres such as *Twitter* or *YouTube*, the personal details of the users are mostly unknown and cannot be deduced from the usernames (nicknames). The most anonymous genre –'Image boards' – naturally has the fewest variables; and '*Facebook* posts' is not specified for 'topic'. In *Twitter*, an additional distinction was introduced between the main author, i.e., the account holder (*user*), and the authors of other tweets, who are referred to as 'commentators'.

After determining the format of the headers, the issue of texts was addressed. Unlike more traditional genres, the ones included in this corpus exhibit features which compensate for prosody and other paralinguistic features typically associated with speech (see Crystal 2003: 291–293). In the texts this is, for instance, indicated by the use of emoticons (compensating for facial expressions and gestures), non-standard spellings (dialect features, slang, abbreviations), and different typographical conventions used to signal emphasis or a raised tone of voice. Other phenomena, such as the use of politically incorrect language and the frequent occurrence of orthographic mistakes, are linked to the spontaneity and the reduced level of formality in CMC. The different linguistic features tagged in the texts are described in the following sections.

BLOGS	IMAGE BOARDS	SMS	*TWITTER*	*FACEBOOK*	*YOUTUBE*
text ID	text ID	text ID	text ID	text ID	text ID
user		user	user	user	user
language	language	language	language	language	language
posting time	posting time	date	posting time	posting time	posting time
user age		user age	user age	user age	user age
user sex		user sex		user sex	
		native language			user country
topic	topic		topic		topic
			commentators		
URL	URL		URL		URL

Table 3. Text and user variables in the different *DMC* components

4.5. How to tag special symbols, icons and emoticons

The spontaneity, the relatively low level of formality, and the celerity with which language is used in digital discourse explain the frequent use of special characters and icons instead of fully written words. In the corpus files, we find

special symbols replacing words such as ‘and’ (&) or ‘degrees’ (°, see example (4)), different currency signs such as ‘dollar’ ($), the heart symbol used for ‘love’ (♥, also represented as “<3”), the *at* sign @ used for ‘directed at’ as well as local or temporal *at* in tweets and posts (English and other languages, too), and many more. The *Facebook* example in (10), for instance, shows how @ can be used to address more than one person in one post.

In all text files, special symbols were tagged with *symbol tags* in the way shown in examples (5)–(10) (once more, English translations are given in italics). This way, the symbols themselves are preserved in the text files whenever possible, but their transliteration into words is also given in order to facilitate word searches with text-based concordancers.

(5) <jasminlopez1970>
i will pray for you [reg=guyz] guys [\reg]..... just believe [sym=&] and [\sym] pray....
(*DMC*, YTC017)

(6) <katyperry 02112011>
It's [emphcap] HERE [\emphcap]! Grab [reg=urself] yourself [\reg] a horchata [sym=&] and [\reg] a churro cause the [sym=#] hash [\sym] CALIFORNIADREAMSTOUR goes to Mexico!
(*DMC*, TWT002)

(7) <Kordula> ... liebend gern, liebe Carla [sym=<3] heart [\sym]
<Kordula> ... I’d love to, dear Carla [sym=<3] heart [\sym]
(*DMC*, FBP007)

(8) <Timrath> Are they obliged to touch the dispatch box when they're speaking?
<Neighb0urhoodWatch> [sym=@] at [\sym] Timrath No
(*DMC*, YTC007)

(9) <MorseCoach> [em lol] lol [\em lol] guy sleeping [sym=@] at [\sym] 2:00
(*DMC*, YTC007)

(10) <Johannes> [12/12/2011 06:17pm]
[sym=@] at [\sym] flo: denkst wie dein [fl language] [reg=bro] brother [\reg] [\fl language] nur ans saufen [em crack up] xD [\em crack up]
[sym=@] at [\sym] flo: you always think of nothing but booze like your [reg=bro] brother [\reg] [em crack up] xD [\em crack up]
[sym=@] at [\sym] kathrin: ich bin grade in wien und mach ein praktikum für mein studium
[sym=@] at [\sym] kathrin: I’m in vienna just now doing an internship for my studies
(*DMC*, FBP012)

A regular strategy that users apply in order to reinforce the written comments and express their mood and emotions is the representation of facial expressions by so-called *emoticons*. Digital discourse is notorious for the use of predefined sequences of punctuation marks, such as the *smiley* :), which many programmes recognise and automatically convert into the corresponding pictorial representations (☺).

In the current version of the corpus, there are 37 different emoticons which had to be tagged accordingly, and more types will certainly appear as the corpus grows. Since each emoticon signals a different mood or facial expression, it was decided that each meaning would have to be specified within the *emoticon tag*.

Examples (11)–(14) show just a few occurrences of the many emoticons found in our data. Figure 5 shows the beginning of the extensive tag list that accompanies the corpus files.

(11) <calisunluvr> Awe. So cute and funny. [em smile] :) [\em smile]
(*DMC*, YTC023)

(12) <esa2go> the boy at the end could be little sheldon cooper [em laugh] :D [\em laugh]
(*DMC*, YTC027)

(13) <BrunoMars 15112011>
Damn we hit 5 Million! […] [reg=Lets] Let’s [\reg] take our shirts off [em nyah] :p [\em nyah]
(*DMC*, TWT003)

(14) <QuercusSola>
[sym=@] at [\sym] simbaglare714 I know what you mean. When I talk to neighborhood kids I have to switch to “dumb english”… [em lol] lol [\em lol] [em laugh] :D [\em laugh]
(*DMC*, YTC001)

1	Text/Symbol	Tag
2	lol	[em lol) lol [\em lol]
3	:)	[em smile] :) [\em smile]
4	:(	[em sad] :([\em sad]
5	;)	[em wink] ;) [\em wink]
6	:-)	[em smile nose] :-) [\em smile nose]
7	:-(	[em sad nose] :-([\em sad nose]
8	;-)	[em wink nose] ;-) [\em wink nose]
9	(-:	[em smile left] (-: [\em smile left]
10	|-)	[em hee] |-) [\em heehee]
11	|-D	[em hoo] |-D [\em hoho]
12	:->	[em smirk] :-> [\em smirk]
13	:-(	[em boohoo] :-([\em boohoo]
14	:-<	[em realsad] :-< [\em realsad]
15	:-|	[em hmm] :-| [\em hmm]
16	:-O	[em uhoh] :-O [\em uhoh]
17	:-o	[em shock] :-o [\em shock]
18	:-p	[em nyah nose] :-p [\em nyah nose]
19	:p	[em nyah] :p [\em nyah]
20	|-P	[em yuck] |-P [\em yuck]
21	:-X	[em lipssealed] :-X [\em lipssealed]
22	<:-)	[em dunce] <:-) [\em dunce]
23	>:-<	[em mad] >:-< [\em mad]
24	:-@	[em scream] :-@ [\em scream]
25	:-\	[em undecided] :-\ [\em undecided]
26	:-U	[em sarcastic] :-U [\em sarcastic]
27	:-D	[em laugh nose] :-D [\em laugh nose]
28	^ ^	[em grin] ^ ^ [\em grin]
29	B-)	[em cool] B-) [\em cool]
30	:D	[em laugh] :D [\em laugh]
31	XD (or xD)	[em crack up] XD [\em crack up]
32	;P	[em cheeky wink] ;P [\em cheeky wink]
33	-.-	[em irritated) -.- [\em irritated]
34	;D	[em wink laugh] ;D [\em wink laugh]
35	>.<	[em frustrated] >.< [\em frustrated]
36	=D	[em happy laugh] =D [\em happy laugh]
37	(=	[em happy smile left] (= [\em happy smile left]
38	O.O	[em shock] O.O [\em shock]
39	👍	[thumbs up]
40	👎	[thumbs down]

Figure 4. Screenshot of the first part of the *DMC* tag list

4.6. How to tag non-standard spellings, errors and abbreviations

In order to mark the great number of non-standard spellings and abbreviations found in the data, a special *regularisation tag* was introduced which permits to both keep the original item and insert a standardised variant. This way, dialectal, informal or slang realisations can be searched for directly or via the corresponding standard lexeme.

The examples shown in this section include all kinds of well-known phenomena which are usually associated with speech, such as the dropping of final *–g* in (15), but also creative innovations, such as the use of numbers or individual letters for homophonous words in (18), or extensive abbreviations which are not necessarily known outside a specific user community, as seen in (19). The latter are an especially frequent feature of image boards, which contain multiple abbreviations and figures of speech not found in other CMC genres. Two common abbreviations in image boards – *mfw* ("my face when") and *op* ("original poster") – are shown in (20) and (21), respectively.

The regularisation tag turned out to be one of the most frequent tags in the entire corpus; these are just a few examples.

(15) <beckymefford> So [reg=Freaken] Freaking [\reg] cute!!!!!! (*DMC*, YTC023)

(16) <justkidding93>And I'm a [reg=preachers] preacher's [\reg] kid to boot!
(*DMC*, YTC001)

(17) <Sean> some burn on the rugby but on the other hand we're all off to poland
<Barry> some burn [reg=alrite] alright [\reg] haha.
(*DMC*, TXT003E)

(18) <jasonderulo>
I'm excited [reg=2] to [\reg] [reg=b] be [\reg] going home [reg=4] for [\reg] thanksgiving! [reg=4] four [\reg] [reg=yrs] years [\reg] since I've enjoyed home [sym=&] and [\sym] [reg=fam] family[\reg] on t-day, not [reg=2] to [\reg] mention last [reg=yr] year [\reg] [sym=@] at [\sym] Ruby tuesday's! ha!
(TWT001)

(19) [reg=tihilw] this is how it looks worn [\reg] [BLG001_picture158.jpg]
(*DMC*, BLG001, referring to a picture in the blog)

(20) <No.2268571> [16:40] [pic 1324417255.jpg]
[reg=mfw] my face when [\reg] i see the body artist tucked in there
(*DMC*, IMB008)

(21) <No.2268577> [16:43] [2268564]
Have you seen the movie fight club [reg=op] original poster [\reg]?
(*DMC*, IMB00)

4.7. How to tag foreign language expressions

Another common feature in digital discourse are switches between languages. In our corpus, we found English words in German texts, Spanish words in English texts, and various other combinations. It was decided to mark these words in order to facilitate, for instance, the analysis of code-switching. The tag used here is a *foreign language tag* opening with the bracket [fl *value*], where *value* is the respective language of the tagged word or words. Foreign language expressions in our data range from individual words, as seen in the two German SMS in (22) and (23), to short phrases, such as (24), or even entire sentences, as seen in (25).

(22) <Christian> Wann seid ihr da, [fl English] guys [\fl English]?
When will you be there, guys?
(*DMC*, TXT105G)

(23) <Nena> Jetzt [reg=hab] habe [\reg] ich schon fast alle apps gelöscht [reg=nen] einen [\reg] viren scan gemacht und die scheiße schickt immernoch [fl English] fake [\fl English] nachrichten raus.
I have already deleted most of my apps, did a virus scan and this shit is still sending fake messages.
(*DMC*, FBP025)

(24) Tonight I am a Glamour magazine World's Most Beautiful All-Star Something-Something, and lovers, nobody deserves it [fl Spanish] mas que yo [\fl Spanish].
... more than me.
(*DMC*, BLG002)

(25) <anna10797> I defy anyone to say this lady [reg=isnt] isn't [\reg] talented! Feckin awesome!!
<diegohugostoso1> please! [fl Portuguese] alguém sabe o nome da primeira musica que ela cantou ??? [\fl Portuguese] thanks
... Does anybody know the name of the first song she sang?...
(*DMC*, YTC008)

Alongside such simple examples, we frequently find foreign language expressions which exhibit additional features requiring other tags. Just like any other passages in the discourse, interjections in a different language can contain non-standard spellings and abbreviations, and they can use the same typographical conventions, for example in order to signal emphasis as described in section 4.8. A combination of features can simply be marked by *nested tags*, as shown in (26) and (27) (repeated from (10)).

(26) Well Played, Jennifer Lopez "[fl Spanish] [emphcap] HOLA [\emphcap] [\fl Spanish]… [italics] sniffle [\italics]… [emphcap] LOVERS [\emphcap]".
(*DMC*, BLG002)

(27) <Johannes> [12/12/2011 06:17pm]
[sym=@] at [\sym] flo: denkst wie dein [fl language] [reg=bro] brother [\reg] [\fl language] nur ans saufen
[sym=@] at [\sym] flo: you always think of nothing but booze like your [reg=bro] brother [\reg]
(*DMC*, FBP012)

One of the most creative examples in the *DMC* is the mixed-code expression shown in (28), where German *viertel vor vier* 'quarter to four' becomes *4tel 4 4*. The author, *Philip*, uses digits instead of numbers to type in the time when he wants to meet. *Vier* 'four' becomes *4*. In addition, the preposition *vor* 'before/to' is represented by an English *4*, which is possible because of the near-homophony of the two expressions *vor* /fɔːɐ/ - *four* /fɔː(r)/. Note that the German word for number 4 is *vier* /fiːɐ/. While the first, second and fourth *4* in this message are pronounced in German, the third *4* must get the English pronunciation in order to make sense.

(28) <Philip>
Sorry aber das wird [reg=nix] nichts [\reg]. Sina kann erst doch um [reg=4] vier [\reg]. Also kannst länger arbeiten [em smile] :) [\em smile] bin [reg = 4tel] viertel [\reg] [reg=4] vor [\reg] [reg=4] vier [\reg] bei dir.
Sorry but I can't. Turns out Sina can only make it at 4. So you can work longer [em smile] :) [\em smile] I will be at your place at quarter to 4.
(*DMC*, TXT020G)

4.8. Typographical conventions signalling emphasis

Among the unique features that distinguish speech from writing is the use of prosodic elements, including emphasis through tempo and loudness (cf. Crystal 2003: 291). Different CMC genres have found a way to replace these elements by means of typographical conventions indicating increased emotivity and intensity. Overall, the two most widespread strategies – in texts which do not allow any other type of formatting – involve the use of capitalisation ([emphcap]) and asterisks ([emphast]), as shown in (29)–(33). Another convention which is found less frequently, additional spacing between letters ([emphspa]), has not occurred in our dataset so far.

(29) Something you collect: Monster bottle caps. Knowledge [emphcap] YEAH [\emphcap].
(*DMC*, IMG002)

(30) <katyperry 01112011>
Truly! RT [sym=@] at [\sym] Oh Ferras: tonight i'm going to wear…. [emphcap] NO MAKE UP! [\emphcap] best way to give [reg=y'all] you all [\reg] a fright!
(*DMC*, TWT002, RT 'retweet')

(31) <beautifulgirl95100>
[em lol] LOL [\em lol].[emphcap] EVERYBODY, THE FINEEE GUY AT THE END ON THE MOTORCYCLE, IS MICHAEL JACKSON'S NEPHEW, SIGGY JACKSON. [\emphcap]
< ShizzleKizzle07>
I will [emphcap] NOT [\emphcap] cry. [emphast] *sniffles and clears throat* [\emphast]
(*DMC*, YTC011)

(32) <mhairicatherine>
i [emphcap] LOVE [\emphcap] this!!!! seen it so many times and its utterly adorable
(*DMC*, YTC022)

(33) <SuperPeaceout14>
i love this video [em lol] lol [\em lol]
[…]
<zomgseriosuly>
Anthony Padilla? [emphast] *hungry face* [\emphast]
(*DMC*, YTC005)

Depending on the user interface, words can also be emphasised through a modification of the font, i.e., they can be underlined or set in bold type or italics—typographical conventions which are well known from writing. Since these changes are not displayed in plain text, we decided to mark them with the corresponding tags seen in Table 4.

Note that in the preliminary version of the *DMC* the use of emphatic asterisks is only found in *YouTube* posts, but it would probably not be restricted to this medium in a larger dataset.

Graphological conventions	*DMC* tags
capital letters used for emphasis/ shouting	[emphcap] … [\emphcap]
asterisks for emphasis	[emphast] … [\emphast]
letter spacing used for emphasis/ "loud and clear"	[emphspa] … [\emphspa]
underlined words	[underlined] … [\underlined]
words in italics	[italics] … [\italics]
words in bold type	[bold] … [bold]

Table 4. Graphological conventions signalling emphasis

4.9. Politically incorrect language: to tag or not to tag?

The final challenge in this project was the frequent use of swearwords and expletives, for instance in media such as *YouTube*, *Twitter* and image boards. It soon became apparent that most of the students involved felt uncomfortable including these words in the corpus without comment. Several solutions were proposed for tagging words such as *fucking*, *damn* and the like, but in the end it was agreed that, from a matter-of-fact linguistic perspective, there is no reason why these words should be distinguished from non-expletives.

In the future, the frequency of expletives in CMC, as compared to more traditional media, will certainly arouse some interest, and considering the widespread prejudices against certain CMC genres, this topic is in dire need of linguistic investigation, both qualitative and quantitative. It might therefore, at some point, make sense to introduce *expletive tags* in datasets such as the *DMC*.

5. Conclusion and outlook

The project presented in this paper proved to be a very positive experience, both from a didactic and from a corpus-linguistics point of view. Multiple challenges that were brought up during the collection and processing of the data were readily accepted by the students involved. Despite their lack of corpus and tagging experience, the students' familiarity with the genres at hand and the awareness that they could actively contribute to the production of "something new", more than outweighed the technical difficulties which are to be expected in this type of linguistic spadework.

On the basis of the continuous assessment of the tasks described in section 2 and the final course evaluation, it can be concluded that the students developed a firm understanding of human communication and of the differences between the various media and genres used to transmit information. In addition, the practical tasks in this seminar required particularly strong interpersonal skills. The communication and collaboration within the research teams provided an incentive for developing solutions in joint effort, completing assigned tasks within a given time frame, taking common decisions and sharing experiences.

As a final common task, the entire class wrote a corpus manual, comprising a general description of the textual markup and processing guidelines (written by the lecturer), as well as individual sections explaining the different components and their special characteristics. These sections were written by the students themselves – a task which proved more demanding than expected. Compared to the usual essays and term papers that students have to write during their studies, corpus manuals present a different genre with a very technical style and purpose. Thus, writing the manual presented an additional challenge and learning experience.

A strong motivation in this particular seminar was to create an "end product to share" (as mentioned in section 2), i.e., the corpus itself, which the course participants could subsequently use as an empirical basis for their own investigations. Regarding the student papers that resulted from the seminar, it is admittedly difficult to assess to what extent the didactic approach adopted in this project may have factored into the quality of the linguistic analyses. However, the general feedback from students who decided to write a term paper suggests that they felt more comfortable analysing data which they knew, and their newly obtained certitude as researchers who had been involved in the decision making and construction of their own database was positively reflected in how they construed their arguments in favour of the methodology and approach they chose for their investigations. A most encouraging response from various participants was their interest to continue contributing to the corpus afterwards.

In order to objectively assess the didactic value of the approach described in this paper, and in order to estimate its influence on student efficiency in corpus use, a special experiment would need to be designed to warrant the comparability with other corpus linguistic seminars. This could be implemented through a series of parallel or consecutive seminars on corpus linguistics using different didactic approaches for student groups with comparable computer skills and corpus experience.

With respect to the challenges discussed in section 4, the solutions proposed by the student teams were surprisingly similar to strategies known from established corpora. The intuitive response to problems posed by the data was generally unanimous across the different teams dealing with different CMC genres, for example, regarding the definition and delimitation of textual units, as well as the handling of linguistic features and typographical conventions that are not encountered in more traditional media. All of the solutions offered in this paper aim at facilitating the conversion of original CMC data into text-only files which can be searched with the usual concordance programmes. The tags proposed are straightforward and easy to implement in any type of digital discourse, allowing other datasets to be tagged along the same lines. Regarding the *DMC* itself, the design and markup opted for in the preliminary version will allow the corpus to expand and include further genres, and further languages, as the project continues.

References

Beißwenger, Michael and Angelika Storrer. 2008. Corpora of computer-mediated communication. In Anke Lüdeling and Merja Kytö (eds.), *Corpus linguistics: an international handbook. Volume 1*. Berlin: Mouton de Gruyter, 292–308.

Capraro, Robert M. and Scott W. Slough (eds.). 2009. *Project-based learning: an integrated science, technology, engineering, and mathematics (STEM) approach*. Rotterdam: Sense.

Crystal, David. 2003. *The Cambridge encyclopedia of the English language*. Second edition. Cambridge: Cambridge University Press.

Crystal, David. 2004. *A glossary of netspeak and textspeak*. Edinburgh: Edinburgh University Press.

Crystal, David. 2006. *Language and the Internet*. Second edition. Cambridge: Cambridge University Press.

Crystal, David. 2010. The changing nature of text: a linguistic perspective. In Wido van Peursen, Ernst D. Thoutenhoofd and Adriaan van der Weel (eds.), *Text comparison and digital creativity*. Leiden: Brill, 229–251.

Crystal, David. 2011. 'O brave new world, that has such corpora in it!' New trends and traditions on the Internet. Plenary paper to ICAME 32: Trends and Traditions in English Corpus Linguistics. Oslo, June.

Facebook. 2013a. Newsroom: Key Facts. *Facebook*. Webpage. <http://newsroom.fb.com/Key-Facts> (9th July 2013).

Facebook. 2013b. Information. *Facebook*. Webpage. <http://www.facebook.com/ facebook?v=info> (9th July 2013).

Ferrara, Kathleen, Hans Brunner and Greg Whittemore. 1991. Interactive written discourse as an emergent register. *Written Communication* 8/1: 8–34.

Herring, Susan C. 2002. Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology* 36: 109–168.

Herring, Susan C. 2007. A faceted classification scheme for computer-mediated discourse. *Language@Internet* 4. Article 1. <http://www.languageatinternet. org/articles/2007/761> (12/06/2013).

Owen, Paul and Christopher Wright. 2009. Our top 10 funniest YouTube comments – what are yours? Blog posting. *The Guardian* "Technology Blog", 3 November 2009. <http://www.guardian.co.uk/technology/blog/2009/nov/ 03/youtube-funniest-comments> (9th July 2013).

Peterson, Eric E. 2011. How conversational are weblogs? *Language@Internet* 8. Article 8.

Siebenhaar, Beat. 2006. Code choice and code-switching in Swiss-German Internet Relay Chat rooms. *Journal of Sociolinguistics* 10/4: 481–506.

Stoller, Fredricka L. 2002. Project work: a means to promote language and content. In Jack C. Richards and Willy A. Renandya (eds.), *Methodology in language teaching: an anthology of current practice*. Cambridge: Cambridge University Press, 107–119.

Wrigley, Heide Spruck. 1998. Knowledge in action: the promise of project-based learning. *Focus on Basics* 2/D: 13–18.

Yates, Simeon Y. 2001. Researching Internet interaction: sociolinguistic and corpus analysis. In Margaret Wetherell, Simeon Yates and Stephanie Taylor (eds.), *Discourse as data: a guide for analysis*. London: SAGE, 93–146.

Yus Ramos, Francisco. 2011. *Cyberpragmatics: Internet-mediated communication in context*. Amsterdam: John Benjamins.

RiCL Research in Corpus Linguistics

Hedging expressions used in academic written feedback: a study on the use of modal verbs

Kok Yueh Lee[1]
University of Birmingham / United Kingdom

Abstract – This paper sets out to answer a fundamental question: 'How do tutors hedge their comments using modal verbs?' A total of 126 feedback reports comprising 35,941 words were collected from two Humanities departments in a UK higher education institution. Although this is a relatively small corpus, it is a specialised corpus. The research focuses on a specific genre – written feedback –, thus the findings should be justifiable in relation to the hedging expressions used in giving feedback through the use of modal verbs.
A wordlist search of the nine core modal verbs (*can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*) was carried out with WordSmith Tools 5. The results show that *could*, *might* and *would* are the top three modal verbs, followed by *can*, *may*, *must*, *should* and *will*, all of which are used as hedging, although some level of certainties are higher than others. *Shall* was not found in the written feedback, since it is more commonly used in legal texts. The modal verbs *could*, *might* and *would* were used most often because of their lower levels of certainty. *Must*, *should* and *will* indicate the higher certainty level, more direct and less opted for.
The concordances for each modal verb were also further examined for their functions. The modal verbs were used to indicate criticism (*can*, *could*, *may*, *might*, *will* and *would*), suggestions (*could*, *may*, *might* and *would*), possibility (*may*, *might* and *can*) and necessity (*must* and *should*). Other functions included permission (*can*), certainty (*will*) and advice (*would*), all of which were of very low frequency. The results show that tutors tend to be more assertive or direct when commenting on mechanical aspects of writing (through *must* and *should*) and to use more hedging in criticising or offering suggestions.
The findings of this research aim to provide a feedback framework as a reference guide to teacher training programmes.

Keywords – academic English, hedging, modal verbs, written feedback

1. INTRODUCTION

This study developed from an initial exploration of genre patterns in written feedback. The findings from genre analysis revealed that a prevalent feature of feedback was the use of modals. Therefore, the general aim of this paper is to explore how modal verbs are used in the written feedback to express hedging. There are two types of modals, core modals and semi-modals (Biber 2006: 483–484; Carter and McCarthy 2006: 420, 922). The former include *can*, *could*,

[1] I would like to thank my supervisor, Professor Chris Kennedy, for his invaluable comments and thoughts throughout this entire research. I would also like to express my gratitude to CILC2012 conference's organisers, chairperson and participants for their helpful comments and thoughts of reflection. Finally, I would like to thank the Ministry of Education, Brunei Darussalam, providing the funding and opportunity for me to develop my research.

may, *might*, *must*, *shall*, *should*, *will* and *would*, and the latter, also called 'marginal modal verbs', include *dare*, *need*, *ought to* and *used to* (Carter and McCarthy 2006: 420, 922). Modals often embed a "degree of certainty and necessity" within them, whether something said or written is "real or true" or merely an assumption (Carter and McCarthy 2006: 638). This study only explores how hedging is expressed in written feedback through the use of core modal verbs (*can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*).

2. A BRIEF LITERATURE REVIEW

Written feedback is one of the main fundamental activities in universities for teachers (Parboteeah and Anwar 2009: 753). Giving feedback is also one of the important daily tasks tutors have to do (Ziv 1982: 2, F. Hyland 1998: 255; K. Hyland 2006: 103; Nicol and Macfarlane-Dick 2006: 200). It is part of "an educator's life" (Jackson 1995: 1). Keh (1990: 294) defines feedback as "input from a reader to a writer with the effect of providing information to the writer for revision". It consists of statements specifying the strengths and weaknesses of individual students, while offering ways in which students can improve in subsequent writings (Jackson 1995: 7; Harmer 2001: 99; Rust 2002: 152). Feedback is also one of the effective methods in enhancing writing competency (Ziv 1982: 2; K. Hyland 2006: 102–103). Feedback gives students information on their development an accomplishment as opposed to a summative form where students only learn whether they have passed or failed the task (Nicol and Macfarlane-Dick 2006: 212). Research by Lee (2003: 220) and others (see Jackson 1995: 2; Gibbs and Simpson 2004: 17; Nicol and Macfarlane-Dick 2006: 203) provide summaries on the main purposes of feedback, among which are helping students to improve writing competency, to become reflective learners and to recognise their errors by indicating to them their strengths and weaknesses in writing. This is summarised in Figure 1.

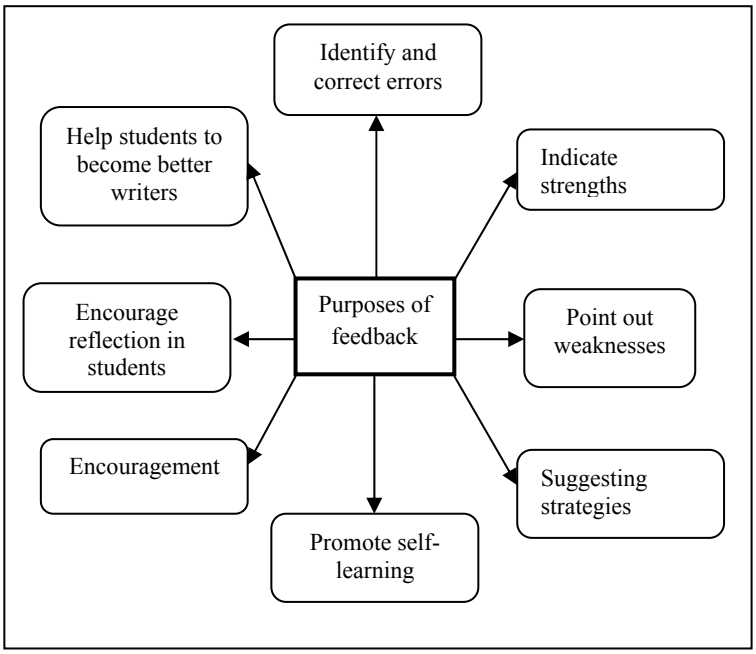


Figure 1. Purposes of feedback (adapted from Jackson 1995; Lee 2003; Gibbs and Simpson 2004)

Abundant research has been carried out on the area of feedback, either in native English language classrooms or in foreign classroom settings with native and non-native tutors and students. However, much of this research has looked into the effectiveness of feedback practices, the ways to deliver effective feedback, the common misperceptions of feedback among the tutors and students or the impact of feedback on students. This research does provide a good insight into understanding feedback and the extent to which feedback is useful and acted upon by students. Another type of research into feedback has also been carried out in higher education settings, focusing on undergraduates' (see Glover and Brown 2006; Stern and Solomon 2006) and postgraduates' level of study, including both master level (see Mirador 2000; Hyatt 2005) and doctoral level (see Kumar and Stracke 2007; Nkemleke 2011), or student-teacher training courses (see Farr 2011), some of which have explored the use of hedging expressions in the feedback reports (Farr 2011; Nkemleke 2011).

Likewise, much research into academic writing has been carried out in the area of hedging since its first introduction by Lakoff (1973: 175), who defines *hedging* as "words whose job is to make things more or less fuzzy". Swales (1990:

175) defines hedging as linguistic devices which express "honesty, modesty and proper caution in self-reports". Other researchers have also defined it as a linguistic strategy used by writers or speakers to express tentativeness and possibility with respect to the truth of propositions (Crismore and Vande Kopple 1988: 185; K. Hyland 1996a: 477; 1996b: 433; 1998a: 350; 1998b: 1). The term *hedging* itself is broad and multi-functional and often overlaps with other aspects such as modality, politeness, indirectness and vagueness (K. Hyland 1995: 34; Farr and O'Keeffe 2002: 26; Nkemleke 2011: 19; Salager-Meyer 2011: 36). Salager-Meyer (1994), in her study on medical English written discourse, has proposed five classifications of hedging which are used to represent the subcategories of hedging. Firstly, 'shields', which comprise modal auxiliaries or modals (*can*, *could*, *may*, *might*, *will* and *would*), epistemic verbs (*seem*, *appear*, *believe* or *suggest*), adverbs (*possibly* or *probably*) and their *related* adjectives. Secondly, 'approximators', which refer to quantity, degree, frequency and time (*approximately*, *usually*, *generally*, *somehow* or *somewhat*). Thirdly, phrases which express authors' personal doubt and involvement (*I believe* or *as far as I know*). Fourthly, 'emotionally-charged intensifiers', which express the writer's reactions (*extremely interesting*, *surprisingly* or *particularly encouraging*). Lastly, 'compound hedges' or 'strings of hedges', which could be double hedges (*it may suggest that*), treble hedges (*it would seem likely that*) or quadruple hedges (*it would seem somewhat unlikely that*) (Salager-Meyer 1994: 154, 155).

Arguably the classifications proposed by Salager-Meyer (1994) can be seen as rather stereotypical. However, they do provide a summary of the hedging strategies used by writers across disciplines. For example, Crismore and Vande Kopple (1988) found that hedging in the science and social-studies texts for ninth-graders is expressed through personal voice (*it seems to me* or *I suppose that*) and impersonal voice (*it seems that* or *it is supposed that*). K. Hyland has also done ample research on hedging in scientific research articles examining its functions and the grammatical features used to convey tentativeness. He looks at the use of lexical verbs, adverbials, adjectives, modal verbs and nouns in scientific research articles (1995; 1996a; 1996b; 1998b; 2000), and at the use of directives in various genres (K. Hyland 2002, 2005b). These taxonomies provide a good start-off point to understand hedging and its strategies. For this purpose, it is important to study the context in which the texts are produced. Hyland believes that a person's use of language is influenced by the discourse community (K. Hyland 1998a: 373; 1998b: 35). An author will write according to the expectations or 'norms' within his/her community (K. Hyland 1998b: 35) and in order to have a better understanding on the language use within a specific community, it is important to examine the contextual situation in which the texts are produced.

Hedging is also considered as a softening feature which mitigates a proposition by making it sound more tentative and less forceful (Carter and McCarthy 2006: 923) and which is expressed through modal and semi-modal verbs (*can*, *could*, *may*), lexical verbs (*wonder*, *think*, *hope*) (Carter and McCarthy 2006: 923) or stance adverbs (*perhaps*, *possibly*, *generally*) (Biber 2006: 101). Modals are generally used to express "degree of certainty" or "degree of obligation" (Carter and McCarthy 2006: 898). They are often used by writers (in this case, the tutors) to distance themselves from the reader (students) or, as Stubbs (1986: 1) has stated, to be "vague, indirect, and unclear about just what we are committed to". Modals are used to express various meanings in speech or writing. Coates (1983) has provided a detailed list of the range of meanings that modals convey, a summary of which is shown in Table 1 (see also Carter and McCarthy 2006: 642–656).

Modals	Meanings
can	ability, root possibility, permission
could	root possibility, epistemic possibility, ability, hypothesis
may	root possibility, epistemic possibility, permission
might	root possibility, epistemic possibility, permission, hypothesis
must	strong obligation, confident inference
shall	strong obligation, volition prediction, determination
should	weak obligation, tentative inference, hypothesis, necessity
will	volition, prediction
would	prediction, hypothesis, volition

Table 1. Summary of the meanings of modals by Coates (1983)

According to Nkemleke (2011: 20), "academic language is a world of indirectness and non-finality". Indirectness is regarded as a politeness strategy whereby the writer or speaker show respect to their reader or hearer (Upton and Connor 2001: 321). Myers (1989: 5) indicates that hedging is a politeness strategy in academic writing which forms an interaction between the writers and readers. Salager-Meyer (1994: 150) claims that writers or speakers use hedges to "convey (purposive) vagueness and tentativeness and to make sentences more acceptable to the hearer/reader, thus increasing the chance of ratification". In other words, hedges allows them to remain uncommitted (K. Hyland 1998b: 1; Downing and Locke 2006: 184), and at the same time gives them the opportunity to defend their status as academics (Lafuente Millán 2008: 68). With respect to written feedback, it is a strategy for tutors to be less assertive, or not "sounding too authoritative or direct" (Carter and McCarthy 2006: 906). Hedging is used as a softening feature, a mitigation strategy, to downtone negativity (Hyland and Hyland 2001), or to weaken a proposition to make it "more polite" (Carter and McCarthy 2006: 923).

3. DATA AND METHOD

3.1. Data for this study

The research data for the present study is a compilation of 126 feedback reports from two Humanities departments in a UK higher education institution (Departments A and B henceforth). The feedback reports were given by tutors to English degree undergraduates on their summative essays. The sum total of words was 35,941 (a small specialised corpus that will referred to as the *EdEng Corpus* henceforth). Students' names in Department A were all deleted when the 42 feedback reports were manually transcribed. Optical software was not used because all of the reports were relatively short (an average of 108 words per report, as shown in Table 2). Students in Department B used their student card's numbers, which were all deleted as well. The tutors' names were also deleted, after the number of participating tutors had been counted. No criteria were used for the collection of feedback reports. All feedback reports were used and analysed, and each report was assigned a number for cross-referencing (Text 1–Text 126). Table 2 shows the research participants and data for this study.

	No. of tutors	No. of students	Modules	No. of essays	Total no. of words	Average no. of words per report
Department A	10	6	12	42	4,527	108
Department B	1	42	6	84	31,414	374
Total	11	48	18	126	35,941	285

Table 2. Distribution of participants and data

It is open to argument when it comes to combining Department A and Department B feedback reports into one corpus instead of separating them into two corpora. However, these reports constitute a single genre—academic written feedback—and, therefore, I think it is unnecessary to discern both sets of reports. Nevertheless, when a particular feature is found to have been used only in Department B, since it only came from one tutor and is therefore an indication of idiosyncrasy, this will be mentioned in the results and discussion sections.

3.2. Method

The methodology for this study is based solely on text and corpus analysis. No follow-up study was carried out with the participants as they were hesitant to be interviewed, although they were aware that it would be an anonymous process. There were also very limited responses to the online questionnaire, which was discarded because no significant findings could be obtained from the results. This study started as a top-down approach (Biber, Connor and Upton 2007: 12), whereby hedging is set as the main linguistic function to explore.

This paper will therefore study the use of modals as hedging in feedback. *Wordsmith Tools*, a corpus programme on text analysis (Scott 2010), was used to search for the nine core modals (*can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*). Quantitative analysis showing the frequencies of occurrences was carried out to show the usage of these modal verbs in both departments. Alongside this, a study of the modals used in each department was also carried out to show if there were any discrepancies between the two departments as one corpus is slightly larger than the other. The main part of this study consists in identifying how hedging was used by means of the core modals (*can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*) and in implementing co-text analysis in order to derive the functions of each modal rather than interpreting them intuitively. Quantitative analysis was not carried out for every function of the modals as the main focus of this study is to explore how hedging was expressed through modals. Co-text analysis involves looking at the context of the word, that is to say, words that occur on either side of the word (Sinclair 1991: 172). The *Concord* Tool in *WordSmith 5* (Scott 2010) was used to retrieve the concordances for each of the modals. This has allowed us to see all the instances of the specified item in the corpus which can then be sorted left or right to identify significant textual patterns (McEnery and Hardie 2012: 241). The contracted negation modals (for instance, *can't* or *shouldn't*) were also searched for. Instances containing non-hedging features were extracted manually for each of the modals.

4. RESULTS AND FINDINGS

4.1. Quantitative results

The quantitative results demonstrate the frequencies of the use of modals as hedging in feedback, with an average of 3.5 occurrences per paper, about one every 81 words. Although both departments seemed to use modals equally (12.6 words per 1,000 in Department A and 12.2 words per 1,000 in Department B), as shown in Table 3, the use of modals in Department B (nearly 5 modals in every feedback report) is higher than in Department A (approximately 1 modal in every paper). This is mainly due to the amount of feedback given by the tutor in Department B (an average of 374 words per report as compared with Department A's feedback, an average of 108 words per report, as shown in Table 2 earlier).

	Total (Departments A + B)			Department A			Department B		
	Raw freq.	Modals per 1,000	Modals per paper	Raw freq.	Modals per 1,000	Modals per paper	Raw freq.	Modals per 1,000	Modals per paper
Modals	441	11.5	3.5	57	12.6	1.4	384	12.2	4.6

Table 3. Frequencies of occurrences of the core modals in Department A and Department B

Table 4 shows the frequencies of the nine core modal verbs. The most frequent modals in both departments were *could* and *would*, accounting for nearly 59% of all modals in the corpus (illustrated in Figure 2). *Should* and *must* had a similar frequency in both departments. Although the occurrences of *might*, *will* and *may* in the entire *EdEng Corpus* were very minimal (16%, 9% and 2%, respectively), they were found more often in Department B than in Department A. On the other hand, *can* was found more often in Department A (10% as compared with 3% in Department B, also shown in Figure 3). The modal *shall* was used in neither department.

	Total (Departments A + B)		Department A		Department B	
	Raw freq.	Words per 1,000	Raw freq.	Words per 1,000	Raw freq.	Words per 1,000
could	158	4.4	19	4.2	139	4.4
would	102	2.8	17	3.8	85	2.7
might	70	1.9	3	0.7	67	2.1
will	40	1.1	3	0.7	37	1.2
should	33	0.9	4	0.9	29	0.9
can	20	0.6	6	1.3	14	0.4
may	12	0.3	4	0.9	8	0.3
must	6	0.2	1	0.2	5	0.2
shall	0	0	0	0	0	0

Table 4. Frequencies of occurrences of the core modals in Department A and Department B

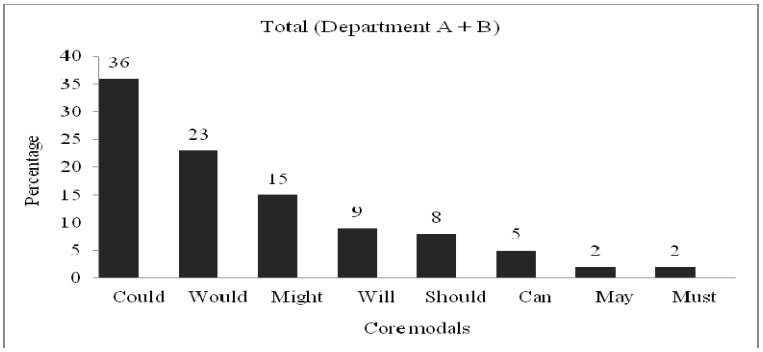


Figure 3. Percentage of occurrences on the use of modals in both departments

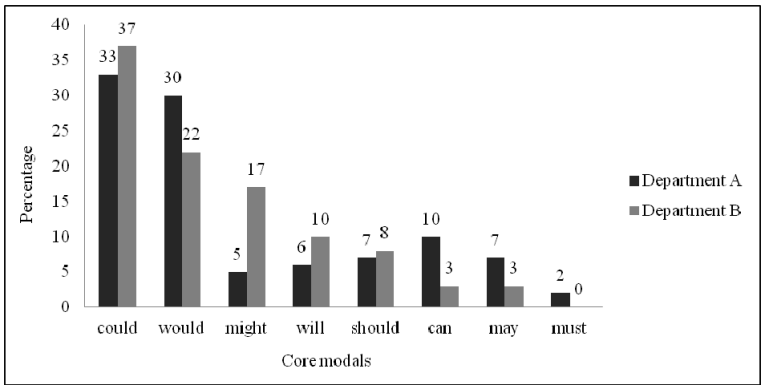


Figure 4. Percentages differences on the use of modals in Department A and Department B

4.2. Functions of modals

4.2.1. Introduction

Each concordance of the modals was thoroughly examined and identified for their respective functions. Seven functions were found to be associated with the modals in this study, as shown in Table 5. Categorising these modals into their respective function posed problems in some cases. There was a degree of fuzziness in the categorisations insofar as one modal could be classified as 'criticism', but at the same time, there was also an implicit suggestion. In this respect, K. Hyland (1996b: 437–438) states that the hedging devices can be rather 'polypragmatic', that is, they may have various meanings and it would therefore be impractical to group them categorically.

Functions	Modals
Criticism	*can*, *could*, *may*, *might*, *will*, *would*
Suggestion	*could*, *may*, *might*, *would*
Possibility	*can*, *may*, *might*
Necessity	*must*, *should*
Certainty	*will*
Permission	*can*
Advice	*would*

Table 5. Functions of modals

*4.2.2. Modals as criticism (*can*,* could*,* may*,* might*,* will*,* would*)*

It is worth mentioning that the modals do not imply criticism by themselves. Rather, it is the co-texts of the modals where the criticism is expressed. The examples shown below illustrate the co-texts of the modals expressing criticism, 'overly-reliant' in example 1, 'colloquial (chatty)' in example 2, 'overstating' in example 3, 'not clearly explained' in example 4, 'punctuation not needed' in example 5 and 'clarity problems' in example 6:

(1) Your style of writing can be rather colloquial (chatty) and you sometimes mix the present and past tenses. (Text 17)
(2) Most of the basic concepts in the essay appear to have been understood although they are not always explained as clearly as they could be (see next section). (Text 71)
(3) You tend to use a semi-colon when you may not have needed any kind of punctuation. (Text 82)
(4) There is a concern that you might be overstating some of your points. (Text 58)
(5) It is also not very clear in your introduction how you will approach and discuss the issue of language modification in the classroom. (Text 110)
(6) My only criticism would be that you are occasionally overly-reliant on your secondary sources—if you could integrate these more smoothly, while privileging your own, very promising, critical voice, this could be even stronger. (Text 10)

*4.2.3. Modals as suggestions (*could*,* may*,* might*,* would*)*

In addition to criticism, tutors also offer suggestions to students, often advising them to being cautious or maintaining solidarity:

(7) I think you could have developed this a little as it is an extremely good point especially since the colour and the message are working together to achieve an effect on the audience. (Text 78)
(8) Your structure is a little unusual—a more traditional introduction to ease the reader into your argument may have helped, for instance, and although the introduction of ideas of the semiotic versus the symbolic is fascinating, I don't think you elaborate on this in enough depth. (Text 3)
(9) One point I wish to make is that it might have been useful for you to have placed the information in the Appendix into the main body of your essay as this information was particularly useful for comparison purposes. (Text 89)
(10) I think it would have been useful to have mentioned the points on 'competency' vs 'fluency' at some point in the essay with reference to the use/learning of grammatical rules. (Text 58)

*4.2.4. Modals as possibility (*can*,* may*,* might*)*

The findings also show that tutors expressed possibilities, generally mitigating the feedback by being tentative and not committed to the feedback.

(11) Here are a few examples of language use that need revision: Page 3: can a parameter setting be 'uttered'? (Text 58)
(12) I think you could have developed on the 'symbolism of the apple' more in your discussion of advertisement 1 (page 5). There may be something to be said about the apple, the seductive appeal of the perfume and the colour green. (Text 78)
(13) Your essay indicates a well developed and quite detailed understanding of the progress of education in England and Wales, although, as you also indicate, 'progress' might not be entirely the right word. (Text 33)

Example 11 above shows an instance where its category adscription can be subjective due to the overlapping categories. "*Can a parameter setting be 'uttered'?*" can be classified as a criticism, but at the same time it might also be seen as indicating uncertainty, and may therefore be rephrased as "Is it possible for a parameter setting to be uttered?".

*4.2.5. Modals as necessity (*must*,* should*)*

Apart from tentativeness, there are instances where tutors were more forceful. These cases were normally found when feedback was on mechanical aspects of writing, such as referencing. Hedging is very minimal in these cases.

(14) You must pursue one line of reasoning, and signpost that throughout. (Text 31)
(15) The essay largely adheres to the guidelines stipulated in the Style Guide. However, you do need to note that the bibliography should be presented in alphabetical order. (Text 55)

*4.2.6. Modal as certainties (*will*)*

Another very minimal hedging sense lies in the function 'certainty' and is found to be used with *will* (example 17). Example (16) is more tentative, as it is initiated by another hedging marker, *I think* (Carter and McCarthy 2006: 223).

(16) I think your work will be enhanced by more research and advise you to develop this dimension of essay construction, especially given the positive qualities you display in other aspects of your writing. (Text 40)
(17) An alphabetical list with surnames first will be sufficient. (Text 120)

Will in these cases is still considered hedging, as it is slightly less assertive than *is/are*, which would be more forceful, as in "an alphabetical list is sufficient", which was also found in the feedback, all from Department B. Other forms which were found are listed below, ranging from the most tentative to simply assertive.

(18) An alphabetical list would suffice. (Text 115)
(19) Your Bibliography does not need to be bulleted; an alphabetical list would be sufficient. (Text 116)
(20) Just to point out that it would be sufficient for you to list your references alphabetically without the use of bullets. (Text 117)
(21) You do not need to bullet your Bibliography as an alphabetical list is sufficient. (Texts 120, 121, 122)

*4.2.7. Modal as permissions (*can*)*

The instances of *can* used in order to give permission were minimal (only 2 occurrences) and found only in Department B. There is very little hedging in this sense, as shown below:

(22) When you refer to sources, you only need to use their surnames and you can omit initials. (Text 63)
(23) You can replace 'a so' with 'an' for it to be accurate. (Text 64)

*4.2.8. Modal as advice (*would*)*

This is found only once in the *EdEng Corpus*, and was used in Department B.

(24) Your essay is well documented and adheres to the requirements in the Style Guide. I would only check the use of punctuation before a quotation. (Text 82)

4.2.9. Non-hedging instances

There were occurrences of modals which were not used for hedging purposes, such as expressing future intentions (example 25), and meta-statements (example 26). However, all of these non-hedging instances were only found in Department B.

(25) Please do take up my offer of discussing assignment 2 before you begin writing it as I think it will be helpful for us to meet. (Text 60)
(26) a. Here are a few points you could note. (Text 43)
b. There are a couple of points I would like to highlight. (Text 58)

Although examples in (26) express some form of hedging, these occurrences were not taken into account. Both these utterances were directing students to subsequent comments presented in bullet form, which is where the main area of investigation was in this study. This is one feature of feedback writing practices implemented by Department B's tutor.

4.3. Patterns of feedback

We have also investigated into the feedback patterns. Feedback was often given by highlighting the positive aspects (POS), indicating the problems or negativity (NEG) or giving suggestions (SUG). In these cases, either all the feedback instances were used alternately or else one or the other was omitted (for instance, POS + NEG + SUG; POS + SUG; NEG + POS, or NEG + SUG), as shown in the examples below.

(27) POS + NEG + SUG
[POS] You write fluently, [NEG] although a few grammatical errors creep in, [SUG] which perhaps a more stringent proofreading process would catch. (Text 6)
(28) POS + SUG
[POS] The move structure analysis is fairly well done [SUG] although it might have been more useful to have shown the analysis diagrammatically rather than through a discussion. (Text 114)
(29) NEG + POS (negativity does not lie in *could,* but in the co-text, as *could* actually mitigates the negativity)
[NEG] This essay has not answered the question as successfully as it could have [POS] although there is evidence of sufficient reading and an attempt at dealing with mostly relevant issues. (Text 80)
(30) NEG + SUG
[NEG] Your essay does not fully adhere to the guidelines stipulated in the Style Guide for in-text referencing. [SUG] In terms of presentation, you need to double space your essay and it might have also been better for you to have retyped some aspects of your appendix (for e.g. the models) than to have just put in the seminar handouts. (Text 101)

It is apparent from these examples that tutors tend to highlight the positive aspect in students' writing. Suggestions were hedged to make them less assertive. Negative comments were mitigated either through positive comments or by offering suggestions. The use of *although* is another salient feature, shown in examples 27–20 above. It generally followed a positive or negative comment, and involved an indication of politeness, mitigating the negativity either way.

5. DISCUSSION

This study has shown that, although tutors from Departments A and B used different templates to write their feedback, they generally hedged their comments by using modals and thus being more tentative. The corpus analysis shows that *could* and *would* were the two most frequent modals used in giving feedback, although *would* was slightly more frequent in Department A (12.3% more). Our findings confirm Farr's research on teaching practice feedback and also shows a high frequency of *could* and *would* in the spoken post-observation feedback. This is mainly due to the tentativeness of these modals and to the fact that they prove to be more polite as compared with the use of *should* or *must*, which are more direct (Carter and McCarthy 2006: 650, 652; Farr 2011: 120). Carter and McCarthy (2006: 640) imply that, by using the past form of the modals (*could* or *would*), they "express greater tentativeness, distance and politeness" between the writer and the reader or speaker and listener. This is the difference between, for example, *it will help your essay* and *it would have helped your essay*. The first utterance expresses a greater degree of certainty than the second utterance, which is more polite and less authoritative. Apart from being used for criticism and suggestion, *would* was also used to give advice, although it only appeared once in the entire corpus. Nevertheless, it showed another feature of hedging.

Another more apparent use of modals in Department A includes *can* (8.9% more than Department B) and *may* (6.2% more). From the analysis, we can see that *can* has more than one function (for instance, criticism and permission). *Can* as 'criticism' was found in both departments, whereas *can* as 'permission' was found only in Department B, which could imply idiosyncrasy. The occurrences were too few for us to provide any further explanations. On the other hand, the co-text analysis of our study has helped us categorise the modals into their respective functions. The function of 'criticism' does not lie within the modal itself, but it can be retrieved by looking at the co-text in which the modal occurs. Although *may* is also tentative, it is less frequent in the *EdEng Corpus* (2.7% in the entire corpus). This is possibly due to the extensive use of *could, would* and *might*, all of which are more tentative than *may*.

Might and *will* were more frequently used by Department B (12.5% and 4% more, respectively). Nevertheless, *might* was the third most frequent modal in the *EdEng Corpus*, as it is more tentative than *may* (Carter and McCarthy 2006: 647). Although both *might* and *could* express tentativeness (Leech 1987: 128), Gresset (2003: 96) stresses that *might* and *could* cannot be used interchangeably, as they are "not strictly synonymous". *Will* was found in the *EdEng Corpus* performing two functions, 'criticism' and 'certainty'. These functions were more definite or certain, thus hedging is very limited. It shows that tutors generally tend to be more direct when referring to the mechanical aspects of writing, such as references or presentation style indicated in the Style Guide or referencing booklet that students are expected to use.

The same holds for *must* and *should*, which are used to express necessity or obligation. Very little hedging was found as these convey a sense of confidence. Tutors seemed to display a higher level of confidence when they were commenting on the mechanical aspects of writing. Arguably, the uses of *should* (as shown in the results section) may be perceived as suggestions as they were proposing ways of improving. Since *should* is at the higher level end of certainty (see Figure 4), it is therefore an indication of necessity or obligation. The occurrences of *must* as necessity or obligation were limited in the *EdEng Corpus*, due to its high level of certainty and confidence (see Figure 4). In fact, tutors seemed to avoid using it, unless the proposition has been made very clear, such as the referencing style. Figure 4 shows the level of certainty and confidence of the modals. The scale of intensity is based largely on the findings of this research, as shown in the examples in the final column.

CONFIDENT	*must*	CERTAIN	which must be avoided
↕	*should*	↕	which should be avoided
	can		which can be avoided
	will		which will be beneficial
	may		which may be beneficial
	might		which might be beneficial
	would		which would be beneficial
DOUBTFUL	*could*	UNCERTAIN	which could be beneficial

Figure 5. Levels of certainty and confidence

We have not examined the use of other modals, such as *will*, to express future intentions, intentions and meta-statements in the feedback, since they were not means of hedging. In addition to this, these non-hedging expressions were all found within Department B's feedback reports, indicating the tutor's idiosyncrasy. The categorisation of modals into their respective functions can be fuzzy as they are often multi-functional, overlapping with other functions (K. Hyland 1996b: 437–438). The teachers' true intention when using each modal in a specific context is hard to be determined unless a follow-up study by means of interviews to tutors is carried out. Unlike Nkemleke's (2011) findings on the pre-defence reports of doctoral students, we have found that *could* and *might* were completely omitted by supervisor. *Can* was the most frequently modal used by supervisors to avoid ambiguity in the pre-defence reports. This

can also be seen in the use of *should*, the second most frequent. This seems to show that supervisors tend to be less ambiguous in their pre-defence reports. Tutors, on the other hand, are more cautious with their feedback.

Shall was completely omitted in giving feedback. The decline in usage of *shall* is highly evident in contemporary English, if compared to the Old, Middle and Early Modern English periods (Gotti 2003: 269). Gotti (2003: 268–269) shows that shall in is the least frequent of all modals (3.5% per 10,000 words). Leech (1987: 87) too has mentioned the decline of *shall* for prediction, expressing intention and volition, and its use only in "restricted linguistic contexts". These contexts are found more frequently in spoken and fictional registers (Gotti 2003: 269–271). Carter and McCarthy (2006: 650) also confirm that *shall* is more frequent in spoken than written texts, mainly because *shall* is used to "make suggestions or to seek advice", such as "*shall I/we…?*". *Shall* is considered to be very formal (Leech 1987: 87; CollinsCOBUILD 1990: 230, 233; Carter and McCarthy 2006: 650) and this is the reason why it is avoided in the feedback, since tutors tend to be constructive and tentative. (For the rarity of *shall*, see Coates 1983: 25; Biber, Conrad and Leech 2003: 486; Leech 2003.)

Based on our analysis, the patterns of feedback seem to be well-defined. Feedback was generally very positive (POS), although there were also some negatives comments (NEG) and suggestions (SUG). These three features were used at the same time or they were used alternately. One or the other feature could be omitted as well (for instance, POS + NEG + SUG; POS + SUG; NEG + POS; or NEG + SUG, also shown earlier in Section 4.3, examples 27–30). Explicit criticism was very rare in the feedback. Even when found, it was often heavily mitigated either with subsequent positive comments, or initiated with a positive comment before the problems were presented. As shown in this study, one way of hedging is through the use of modals. Nkemleke's (2011) study also reveals similar findings, whereby negative comments were rarely used in pre-defence reports and were also mitigated by positive comments when they were used. In addition to the uses of modals in their respective functions, they were also found in clusters in the feedback, or as "strings of hedges", as Salager-Meyer (1994) proposed. Examples from the *EdEng Corpus* include, *perhaps you could…* (Text 62), or *I think you could have said a lot more…* (Text 77). This further reaffirms the tutor's determination to remain as tentative as possible in giving feedback.

Although explicit criticism was rarely found, a thorough reading of all the feedback reports revealed one case in Department A in which all the feedback was negative. There was not a single positive comment or suggestion in the entire feedback report (extract 1 below). This was found only in the feedback report of weak essays which either failed or had a considerably low passing mark. As for Department B, since feedback was given on individual criteria, there were two failed essays, but the tutor did suggest a few recommendations for improvement (extract 2 below, suggestion underlined). There was also one occurrence where a negative comment was found in the criterion *'Overall'* in Department B (extract 3 below). Apart from these few occurrences of explicit criticisms, most feedback reports were generally positive. Negative comments would often be mitigated (as illustrated in extract 4, mitigation underlined).

Extract 1:
This is a very short piece of work and you do not seem to have put much effort into it. You do not answer the question—this is a very general essay without structure or focus. You do not provide supporting quotations from your chosen novels and much of your essay is spent retelling the narratives rather than analysing them. You speculate a great deal about the responses of child readers, but this is not part of literary criticism. The few critical quotations you include are general and you didn't engage with them. Your research has been ineffective. There are many critical works on Harry Potter but you haven't consulted any of them. Furthermore, at no point do you discuss the fantasy genre – particularly the position of these texts within the genre and the techniques they employ. (Text 37, Department A)

Extract 2:
The essay does not show a sufficient reading of a range of sources. The two books you have referred to are the core textbooks for the course but you needed to have read more widely to achieve a better understanding of the theories taught on the course. Only some of the information here is relevant and accurately interpreted. (Text 68, under the criterion 'Acquisition of knowledge', Department B)

Extract 3:
An essay that has not fully achieved the aims of the assignment. (Text 96, under the criterion 'Overall', Department B)

Extract 4:
The essay does not construct a convincing argument although you do show some indication of having understood some of the material. (Text 80, under the criterion 'Interpretation, analysis, construction of argument and relevance', Department B)

6. Conclusions and recommendations

This paper has looked at the hedging expressions in academic written feedback through a study of the modal verbs. Hedging is used to make a proposition more tentative and indicate a sense of possibility (Salager-Meyer 2011: 35). By implementing a top-down approach and combining it with a quantitative corpus approach, this study has set out to explore how hedging expressions operate within the nine core modal verbs (*can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*). Our analysis and findings show that tutors implemented substantial hedging devices through modals in giving feedback, in order to sound less assertive and soften their recommendations (Upton and Connor 2001: 319). Although the corpus is relatively small, it does show that tutors are on the whole very positive, except for a few occurrences of explicit criticisms in the case of weak essays. It is hoped that this paper has succeeded in showing how modals are used as hedging in giving feedback and how effective feedback-writing practices may be developed for teacher training programmes.

Although this study has tried to categorise the functions of modals in hedging, it is evident from the analysis that there remain some unsolved problems. Hedging is a broad area of investigation and modal verbs are but one part if. In our case, they were only a minor part of the entire feedback report. An investigation into other forms of hedging, such as stance adverbs, submodifiers and vague language, should also be examined thoroughly to see if there are other means by which tutors hedged their comments. Although the idea of investigating co-texts has led to a better understanding of each modal, this strategy is feasible only with a small corpus of samples. Random sampling could perhaps be carried out when dealing with a larger corpus to examine if the modals performed the same function, particularly in modals denoting criticism. Although modals do not denote criticism by themselves, looking at the left or right co-text of the modal may actually reveal their actual hedging function. Further research is needed in order to extend this study, for example by incorporating written feedback from other disciplines and institutional settings as these could possibly have an effect on the frequency of modals and on their hedging potential (Salager-Meyer 2011: 37). Examination of a larger scale corpus and of more instances of feedback would allow a better understanding of the hedging features tutors use when giving feedback, including the functions of modals.

Alongside this, a follow-up study is also needed to clarify the subjectivity problem as experienced in the present study. In addition, it is also possible to replicate previous research (Ziv 1982; Norton 1990; Norton and Norton, 2001; Lee 2003; Glover and Brown 2006) to the present study by investigating the extent to which students have found the feedback effective, or whether the students have taken into account the feedback in their subsequent writing. In fact, since it has shown that students do not necessarily understand the modals used in medical journals (Adams Smith 1984), it would be interesting to show if the same is true for other students. To take even a step further, the tutors' status, age and gender could also be investigated. Salager-Meyer (2011: 37) states that, apart from the discipline itself, these other variables may also contribute to the hedging practices of individuals. Writing is also largely affected by the "cultural contexts" that the writing is intended for (Upton and Connor 2001). Since hedging is a but minor part of writing, it could be affected by cultural issues as well (Salager-Meyer 2011: 37). Future research might take all these variables into account to broaden our understanding of feedback writing practice, hedging and the use of modals.

References

Adams Smith, Diana E. 1984. Medical discourse: aspects of author's comment. *The ESP Journal* 3/1: 25–36.

Biber, Douglas. 2006. Stance in spoken and written university registers. *Journal of English for Academic Purposes* 5/2: 97–116.

Biber, Douglas, Ulla Connor and Thomas A. Upton. 2007. *Discourse on the move: using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins.

Biber, Douglas, Susan Conrad and Geoffrey Leech. 2003. *Longman student grammar of spoken and written English*. Harlow: Longman.

Carter, Ronal and Michael McCarthy. 2006. *Cambridge grammar of English: a comprehensive guide. Spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

Coates, Jennifer. 1983. *The semantics of the modal auxiliaries*. London: Croom Helm.

CollinsCOBUILD. 1990. *Collins COBUILD English grammar*. London: HarperCollins.

Crismore, Avon and William J. Vande Kopple. 1988. Readers' learning from prose. The effects of hedges. *Written Communication* 5/2: 184–202.

Downing, Angela and Phillip Locke. 2006. *English grammar: a university course*. Second edition. New York: Routledge.

Farr, Fiona. 2011. *The discourse of teaching practice feedback: a corpus-based investigation of spoken and written modes*. London: Routledge.

Farr, Fiona and Anne O'Keeffe. 2002. *Would* as a hedging device in an Irish context. In Randi Reppen, Susan M. Fitzmaurice and Douglas Biber (eds.), *Using corpora to explore linguistic variation*. Amsterdam: John Benjamins, 25–48.

Gibbs, Graham and Claire Simpson. 2004. Does your assessment support your students' learning? *Journal of Teaching and Learning in Higher Education* 1/1: 1–30.

Glover, Chris and Evelyn Brown. 2006. Written feedback for students: too much, too detailed or too incomprehensible to be effective. *Bioscience Education e-Journal* 7/3: 1–14.

Gotti, Maurizio. 2003. *Shall* and *will* in contemporary English: a comparison with past uses. In Roberta Facchinetti, Manfred G. Krug and Frank R. Palmer (eds.), *Modality in contemporary English*. Berlin: Mouton de Gruyter, 267–300.

Gresset, Stéphane. 2003. Towards a contextual micro-analysis of the non-equivalence of *might* and *could*. In Roberta Facchinetti, Manfred G. Krug and Frank R. Palmer (eds.), *Modality in contemporary English*. Berlin: Mouton de Gruyter, 81–102.

Harmer, Jeremy. 2001. *The practice of English language teaching*. Third edition. Harlow: Longman.

Hyatt, David F. 2005. 'Yes, a very good point!': a critical genre analysis of a corpus of feedback commentaries on Master of Education assignments. *Teaching in Higher Education* 10/3: 339–353.

Hyland, Fiona. 1998. The impact of teacher written feedback on individual writers. *Journal of Second Language Writing* 7/3: 255–286.

Hyland, Fiona and Ken Hyland. 2001. Sugaring the pill: praise and criticism in written feedback. *Journal of Second Language Writing* 10/3: 185–212.

Hyland, Ken. 1995. The author in the text: Hedging scientific writing. *Hong Kong Papers in Linguistics and Language Teaching* 18: 33–42.

Hyland, Ken. 1996a. Nurturing hedges in the ESP curriculum. *System* 24/4: 477–490.

Hyland, Ken. 1996b. Writing without conviction? Hedging in science research articles. *Applied Linguistics* 17/4: 433–454.

Hyland, Ken. 1998a. Boosting, hedging and the negotiation of academic knowledge. *Text* 18/3: 349–382.

Hyland, Ken. 1998b. *Hedging in scientific research articles*. Amsterdam: John Benjamins.

Hyland, Ken. 2000. Hedges, boosters and lexical invisibility: noticing modifiers in academic texts. *Language Awareness* 9/4: 179–197.

Hyland, Ken. 2002. Directives: Argument and engagement in academic writing. *Applied Linguistics* 23/2: 215–239.

Hyland, Ken. 2005a. *Metadiscourse: exploring interaction in writing*. London: Continuum.

Hyland, Ken. 2005b. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7/2. 173–192.

Hyland, K. 2006. *English for academic purposes: an advanced resource book*. London: Routledge.

Jackson, Michael W. 1995. Making the grade: the formative evaluation of essays. *UltiBASE.* <http://ultibase.rmit.edu.au/Articles/jacks1.html> (12nd June 2013).

Keh, Claudia L. 1990. Feedback in the writing process: a model and methods for implementation. *ELT Journal* 44/4: 294–304.

Kumar, Vijay and Elke Stracke. 2007. An analysis of written feedback on a PhD thesis. *Teaching in Higher Education* 12/4: 461–470.

Lafuente Millán, Enrique. 2008. Epistemic and approximative meaning revisited: the use of hedges, boosters and approximators when writing research in different disciplines. In Sally Burgess and Pedro Martín-Martín (eds.), *English as an additional language in research publication and communication*. Bern: Peter Lang, 65–82.

Lakoff, George. 1973. Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic* 2/4: 458–508.

Lee, Icy. 2003. L2 writing teachers' perspectives, practices and problems regarding error feedback. *Assessing Writing* 8/3: 216–237.

Leech, Geoffrey N. 1987. *Meaning and the English verb*. Second edition. London: Longman.

Leech, Geoffrey N. 2003. Modality on the move: the English modal auxiliaries 1961–1992. In Roberta Facchinetti, Manfred G. Krug and Frank R. Palmer (eds.), *Modality in contemporary English*. Berlin: Mouton de Gruyter, 223–240.

McEnery, Tony and Andrew Hardie. 2012. *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press.

Mirador, Josephine F. 2000. A move analysis of written feedback in higher education. *RELC Journal* 31/1: 45–60.

Myers, Greg. 1989. The pragmatics of politeness in scientific articles. *Applied Linguistics* 10/1: 1–35.

Nicol, David J. and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education* 31/2: 199–218.

Nkemleke, Daniel A. 2011. *Exploring academic writing in Cameroon English: a corpus-based perspective*. Göttingen: Cuvillier Verlag.

Norton, Linda S. 1990. Essay-writing: what really counts? *Higher Education* 20/4: 411–442.

Norton, Linda S. and J.C.W. Norton. 2001. Essay feedback: how can it help students improve their academic writing? Paper presented at the International Conference of the European Association for the Teaching of Academic Writing Across Europe, Groningen.

Parboteeah, Sam and Mohamed Anwar. 2009. Thematic analysis of written assignment feedback: implications for nurse education. *Nurse Education Today* 29/7: 753–757.

Rust, Chris. 2002. The impact of assessment on student learning: how can the research literature practically help to inform the development of departmental assessment strategies and learner-centred assessment practices? *Active Learning in Higher Education* 3/2: 145–158.

Salager-Meyer, Françoise. 1994. Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes* 13/2: 149–170.

Salager-Meyer, Françoise. 2011. Scientific discourse and contrastive linguistics: hedging. *European Science Editing* 37/2: 35–37.

Scott, Mike. 2010. *WordSmith Tools (version 5.0)*. Oxford: Oxford University Press.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stern, Lesa A. and Amanda Solomon. 2006. Effective faculty feedback: the road less traveled. *Assessing Writing* 11/1: 22–41.

Stubbs, Michael. 1986. 'A matter of prolonged field work': notes towards a modal grammar of English. *Applied Linguistics* 7/1: 1–25.

Swales, John M. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Upton, Thomas A. and Ulla Connor. 2001. Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes* 20/4: 313–329.

Ziv, Nina D. 1982. What she thought I said: how students mispercieve teachers' written comments. Paper presented at the Annual Meeting of the Conference on College Composition and Communication, San Francisco, CA.

RiCL Research in Corpus Linguistics

Bill Louw's Contextual Prosodic Theory as the basis of (foreign language) classroom corpus stylistics research

Marija Milojkovic[1]
University of Belgrade / Serbia

Abstract – Corpus empiricism may alter the act of reading. This began as the reader searched a reference corpus for individual words and phrases. With the admission of lexicographers that intuition no longer suffices in providing a definition, corpus stylistics must go further by showing that a literary text can no longer be properly interpreted if not seen against the background of the wealth of recorded textual experience. This by no means suggests that a literary text may not have a satisfying impact on an individual reader; rather, corpus stylistics enhances our interpretation by means that are easily available. The core of Bill Louw's stylistic approach is his claim that prior knowledge is no longer perceived as concepts (unsatisfyingly intuitive).
Therefore, reference corpora may serve to enhance our stylistic interpretation of a literary text that was clearly written to be appreciated as a unique textual experience. Roughly, a large reference corpus will provide many parallel textual experiences, so that 'events' in the studied text are augmented by their counterparts in corpora. Thus, our understanding of the text will be augmented by what is absent from it, but present in the reference corpora. If, furthermore, our classroom is a foreign language one, the reference corpus will serve as missing language experience in the foreign language learner, even if the learner is very proficient.
After giving a brief overview of Louw's Contextual Prosodic Theory (CPT) and its implications for classroom corpus stylistics, the paper describes a study conducted with second-year students of English from the University of Belgrade. The aims of the study are to verify Louw's principle that text reads text and to test the proposed CPT-based methodology. The study consists of a quantitative part (where the learning phase is followed by a final test) and a qualitative part (questionnaire). The proposed methodology relies on confronting the subjects with concordance lines as a means of interpreting a collocation in a given short excerpt, with an absolute minimum of theoretical background. The subjects are tested on semantic prosodies, absent collocates and auras of grammatical strings, through tasks that vary in format. The results obtained are encouraging for CPT, despite the study's limitations, which are also discussed.

Keywords – Bill Louw, events, methodology, Contexual Prosodic Theory, corpus stylistics, semantic prosody, subtext

[1] I am grateful to the 2012 second-year students at the English Department, Faculty of Philology, University of Belgrade, for their willingness to participate in the research and provide honest and helpful feedback. I am also grateful to Bill Louw, University of Zimbabwe, for reading and commenting on this paper.

1. INTRODUCTION

1.1. The aim and significance of the study

The aim of this paper is to present a piece of classroom stylistics research based on the theoretical implications of Contextual Prosodic Theory (CPT), developed by Bill Louw from 1993 to date. As this is the first instance of practical implementation of CPT for the purposes of teaching corpus stylistics, the study will attempt to answer two research questions that are interrelated.

The first research question of this study is how well-founded Louw's claim is that "text reads text" (Louw, personal communication), meaning that language is its own instrumentation (Louw 2011: 174). This means that it is indeed sufficient to analyze the target text through similar 'events' in the reference corpus, or in the authorial corpus if the analyst is looking for private symbolism, without much recourse to theories and with no recourse to concepts. Theories and concepts are treated by Louw as an unnecessary imposition that obfuscates rather than clarifies. According to Louw, there should be no intermediaries between the researcher and raw data. He says (personal communication): "Concepts are there to explain data; but where data is plentiful, concepts become surplus to requirements, because collocation replaces and supplants intuition and intuitively derived concepts. Collocation becomes the instrumentation that 'concepts' thought they once were in such cases".

Secondly, the study being practical in nature, its other research question is to see how successfully a classroom stylistics methodology founded on Louw's theoretical views can be implemented. My hypothesis is that this methodology may be implemented in a university setting no matter whether the students are native or non-native speakers of the language of the studied text. In this particular case, the students are on a near-CPE[2] level, and their oral performance, for example, may often be described as native. This means that, for a variety of purposes, their feedback may be used to influence the methodology applied in a classroom of native speakers of English. It has been established by Louw as early as in 1993 that even a native speaker's intuition is insufficient to consciously process all the implications of a text (Louw 1993); however, because semantic prosodies (SPs) are frequent, second and foreign language learners also sometimes recognize them. Thus, although the non-native speaker will find more new information in a large corpus than the native speaker, the nature of that information is the same for both. The large reference corpus contains a greater number of encounters with a particular instance of language use than either a native speaker or a non-native speaker has experienced. In the case of a native speaker there may be more cases when the knowledge passes from the known, but not recognized, to the recognized, but the knowledge itself is always the same. That is why the words 'foreign language' are placed in brackets in the title.

However, it is my assumption that, although the principle must remain the same, the level of the tasks should be adjusted to the language level of the students. It is true that the corpus, whether it be a large reference corpus of the language or an authorial corpus, is in fact the sum of our possible encounters with the language/author, but in the case of less proficient students the gap between their personal experience of a pattern and the unedited reference corpus experience will be greater than in the case of their more proficient counterparts.

This study is significant because it puts to the test both theoretical and practical assumptions underlying Louw's CPT. The theory is only sound if it works in practice, and this particular theory insists on being 'instrumentation', on being equated with practice. In practical terms, it is possible for a theory to be sound, but to require a better explanation in the classroom – or a longer one, or more hands-on experience with actual data. This study adopts Louw's "text reads text" approach without much modification, and with the minimum of explanation.

1.2. Methodology

The whole study encompassed both quantitative and qualitative research. The quantitative research consisted of a 'learning phase' and a 'testing phase'. The 'learning phase' was conducted in five sessions. At the beginning of a regular lesson, after a short introduction, the students were given a short excerpt from a text and a concordance, with a particular question to answer. After the answers were written, they proceeded to discuss the text with the teacher and other students. The teacher – myself – gave her interpretation of the concordance lines and the studied text, encouraging the expression of individual opinion. Sometimes a spontaneous discussion of the text and the author's possible meaning ensued. I emphasized that the interpretation is not the teacher's, but should be based on the given concordance and, therefore, their personal analyses of the given concordance were the point of the session. To sum up, the students were learning 'by doing', while being encouraged to express what they saw in the concordance lines and the studied text. This strategy of instruction through practice implemented Louw's stance that a corpus stylistician relies on raw data.

---

[2] Certificate of Proficiency in English.

Each session of the 'learning phase' contained a different type of task. Each type of task was dealt with once. The students were asked to hand in their answers without making corrections after the discussion and to sign the papers, so that their responses could be marked and the subsequent progress could be monitored. During the first session the main corpus linguistics terminology was introduced and the students had their first encounter with concordance lines. The terms introduced were 'concordance', 'concordance line', 'node', 'collocates', '9-word window' and 'semantic prosody'. The first session only dealt with an authorial corpus (that of Philip Larkin) as it was deemed the easiest type of task. The next four sessions dealt with concordance lines taken from the reference corpus. The reference corpus used for the purpose of the study was the late Tim John's corpus of *The Times* newspaper of the year 1995, containing 44.5 million words. It was originally intended that a session should last up to 15 minutes, but in practice sometimes twice this amount of time turned out to be necessary.

The 'testing phase' was done in one sitting, without warning. The students were given a test of five tasks, mirroring the different types of tasks dealt with during the 'learning phase'. The estimated time of completion was assumed to be 45 minutes, but the subjects were urged to work at their own pace.

Both the results of the 'learning' tests and the final test were processed at the end. A uniform marking scheme had been established for each question. These were the approximate criteria:

- if the analysis fulfils the expectations the mark is 5 (also if the analysis is different from what was expected but excellent and contains detailed argumentation).
- if the analysis of the concordance lines has been done correctly but no connection between the lines and the text has been established, the mark is 4.
- if the lines were incorrectly interpreted, e.g. the student is misled by the first line or mistakenly interprets concordance lines due to a lack of experience, the mark is 3.
- if the analysis is wrong altogether the mark is 2.
- if no analysis was offered at all the mark was 'zero'.

Attached to the test was a questionnaire consisting of 11 questions. It aimed at getting feedback on the short course the subjects had undergone. This was the qualitative part of the research, designed to show what views the students had formed of the text-corpus interaction and of the course. It was conducted to find out if and to what degree the students had taken to the course – whether and how much the students appreciated the course, understood what was going on, found it useful and whether they would choose this subject if they were given the option. It benefited the students as well as the teacher, as it gave them a chance to express their opinions, given that the methodology must have come as a surprise to many of them.

1.3. Background to the study and its limitations

The subjects of the study were second-year students of English at the English Department at the Faculty of Philology, University of Belgrade. These were two groups out of the four comprising the current generation, referred to as group B and group D, and this is how they will be referred to in this study. They represented approximately half of all second-year students. One group was more proficient than the other at the entrance exam, and both may be considered representative of the level of language proficiency of the current generation, as the other half also consisted of one more proficient and one less proficient group. The research was conducted in February and March 2012, in approximately three weeks, within the framework of the second-year *Integrated Skills* course (the whole course was officially named *Contemporary English – G4*), during class time.

As some (though not all) second-year students of English are at the CPE level (some closer to Advanced, others closer to Proficiency), it was thought that their language knowledge was more or less sufficient to attempt the study of concordances taken from reference corpora with a view to interpreting poetic and other texts. None of the students had any previous practical knowledge of either corpus linguistics or stylistics, except for what is taught at secondary schools and on general undergraduate courses in linguistics and theory of literature. By the beginning of their fourth semester the students had completed courses in general linguistics, phonetics, morphology, and had become used to interpreting poetry and prose in their English literature classes. Consequently, their linguistic and academic background was deemed sufficient for them to attempt corpus stylistic interpretation without intervening concepts.

This arrangement had its faults. First of all, not all the students had the same level of language knowledge and yet all were equally tested. It is true that they were all second-year students, but, as stylistics is about nuances of interpretation, the step from CAE[3] to CPE could be crucial and no data of their language knowledge were available except their final results on the *Contemporary English – G3* course (reading, writing, listening, speaking and translation into English and into Serbian) in the previous semester. These results are not wholly reliable, as they partly depended on the students

[3] Certificate of Advanced English.

having learnt certain vocabulary items and grammatical structures (a very proficient student may have avoided the 'cramming' part, and they often do). It is known, however, that, of the two groups, Group B was more proficient on entering the first year of studies than group D. This, the exam results and the teacher's observation suggest that, on the whole, group B was more proficient than group D at the time of the experiment. The teacher's observation during one whole academic year suggests that group B, on the whole, was closer to CPE level and that group D, on average, was closer to CAE. Moreover, the stylistic analysis per se revealed the language level of certain students, particularly in cases where it proved to be lower than desirable. In this context, it was interesting to see whether the group whose level of English was higher and closer to that of the native speaker's would perform better.

Secondly, through force of circumstance, the research was conducted during the *Integrated Skills* course and not in a stylistics course. It was done this way simply because the researcher was currently teaching that course, and had an opportunity of providing instruction and receiving feedback. The main obstacle here was the fact that certain students lacked affinity for stylistic reading of literary texts, found it lacking in motivation and, for this reason, they might not be considered legitimate subjects. It stands to reason that one's performance ought to be motivated if it is to be successful. The students in question may have possessed all the necessary qualities and qualifications and still they may have underperformed through lack of interest. Nevertheless, all subjects were taken into account when the results were being processed. The final qualitative survey was also an attempt to throw light on the issue of the students' interest.

The students' responses varied, so an attempt was made to standardize marking as much as possible in these conditions. There were many variations, so gradations like 3.5 – and even 4.8 – were added. If an analysis exceeded the teacher's expectations in its acuteness, or if the student came up with a correct conclusion or interpretation that even the teacher herself had overlooked, the student was given 6 points – these were special cases. It seemed to me that the difference between a correct interpretation deserving the mark of 5 and an unexpectedly insightful one needed to be documented and taken into account.

No matter how nuanced the marking was, it could not take into account some important differences. First of all, the mark of 4 was given if the student analyzed the concordance correctly, but failed to see the connection between it and the excerpt studied. This included cases where the student was perfectly aware that a connection could be made, but refused to make it, maintaining that the poet meant precisely what he said (during class discussions group D in particular insisted on a poet's freedom not to be 'automated'). Secondly, in practice, since the mark of 2 was given for wrong interpretation, no one was given the mark of 1 – but the mark of 0 existed as part of the marking scheme and was given for no answer. Thirdly, the mark of 0 may have been earned through lack of motivation, as well as inability to offer interpretation. Finally, the teacher's subjectivity in the presence of so many variations is always a threat to standardized marking, despite her conscious efforts to reward similar answers similarly.

2. CONTEXTUAL PROSODIC THEORY TO DATE

2.1. Contextual prosodic theory: main areas of study and literature review

To date, four main areas of Louw's interest are the delexicalization-relexicalization continuum, semantic prosody, subtext and the implications of philosophy of language for collocation. These areas of study are interrelated and have collocation as their pivotal point.

The delexicalization/relexicalization distinction was first brought up by Louw as far back as 1991. This is Sinclair's summary of Louw's idea of delexicalization: "Words can gradually lose their full lexical meaning, and become available for use in contexts where some of that full meaning would be inappropriate; this is the so-called figurative extension" (Sinclair 2004: 198). Relexicalization comes about when a delexicalized word finds itself in the vicinity of a collocate which, purposefully or inadvertently, brings to mind the delexicalized word's literal meaning. For example, in Henry Miller's novel *Tropic of Capricorn*, the words *ghost* and *dead* are part of delexical expressions, but, within the Sinclairan 9-word window they relexicalize (Louw 2006): "Once you have given up the ghost, everything follows with dead certainty, even in the midst of chaos" (Miller 1966: 9).

The idea is explained in Louw (2007, 2008), the latter paper suggesting that "all devices relexicalise" (Louw 2008: 258), and proposing that all devices be given corpora-attested definitions. The 2008 paper also places collocation in the context of Firth's (1957) context of situation, but with emphasis on the provision of corpus-attested terminology.

Semantic prosody, first mentioned in Louw (1993: 157) as the "aura of meaning surrounding a word or phrase", was given a more comprehensive character in Louw (2000), where he first uses the term *Contextual Prosodic Theory*. In the paper he claims that Contextual Prosodic Theory is confirmatory of the Firthian tradition rather than new. A semantic prosody is here defined as

> a form of meaning which is established through the proximity of a consistent series of collocates, *often* characterisable as positive or negative, and whose primary function is the expression of the attitude of the speaker or writer towards some pragmatic situation. (Louw 2000: 56; my emphasis)

According to Louw, the key to semantic prosodies is Firth's taxonomy for the context of situation. A semantic prosody arises from a fractured context of situation, and by *fractured* Louw means either under- or overprovided one. The approach falsifies Halliday's grammatical metaphor (Louw, personal communication).

Both these aspects – relexicalization and semantic prosody – are mentioned in Louw (2010a), the very title of the paper suggesting that "collocation is instrumentation for meaning". The paper proposes to dispense with concepts, stating that collocation alone interprets both fact and fiction, and, for the first time, introduces the notion of *subtext*, rooted in the work of analytic philosophers: Frege, Carnap, Wittgenstein and Russel. Co-selection chunks states of affairs, while subtext (quasi-propositional variables) provides the underlying argument, in which the grammatical pattern collocates with the author's lexical choices, falsifying the Vienna Circle's assumption that logic and metaphysics must never be separated. Subtext continues to be studied in Louw's (2010b) examples from Yeats. The application of subtext to prose is illustrated at length in Milojkovic (2013), and to poetry in Louw and Milojkovic (2014).

Milojkovic's contribution to Louw's CPT is the application of semantic prosodies and subtext to Russian (Milojkovic 2011a), pointing to the theory's universality, and the notion of 'grammatical strings' (as opposed to 'lexical items') having an aura of meaning (e.g., *but when did* in Milojkovic 2012). The only claim that a grammatical string may have a prosody, and not using this terminology, was made by John Sinclair in 2006. The grammatical string chosen by him (*when she was*), which was arrived at by choosing the next most frequent collocate, in the end turned out to contain two opposite distinct semantic prosodies, depending on the context of situation (Sinclair 2006). In fact, the two specific fractured prosodies found by Sinclair and embedded in the context of situation prove CPT. My other, very small contribution, is the term 'prosodic clash', describing a situation when a particular writer's or speaker's use of collocation is remarkably different from the prosody established through a reference corpus. It is basically the same as Louw's 'fractured prosody', but Louw's term works within Firth's context of situation, whereas a 'clash' emphasizes the discrepancy between the writer's use and general usage. Within Louw's theoretical framework the term 'prosodic clash' is more telling than 'collocational mismatch', for example. A prosodic clash is an indication of either irony or insincerity in the dichotomy first described in Louw (1993).

Thus, CPT is supported by confirming the thinking of the analytic philosophers and founded entirely on collocation. Co-selection chunks states of affairs in terms of context, produces literary devices in terms of expression, and constantly creates new meaning in terms of semantic prosody (lexical co-occurrence) and logical prosody (subtext). All three need to be viewed on the higher level of events within Firth's context of situation. The implications of CPT are not limited to stylistics, as its terminology may be used to interpret events both fictional and real, by comparing the event in the studied text with similar ones in the huge reference corpus.

To my knowledge, apart from Louw, only one author uses reference corpora in the analysis of (literary) texts, namely, Bettina Fischer-Starcke in her work on Jane Austen (Fischer-Starcke 2010).

2.2. The implications of Contextual Prosodic Theory for classroom stylistics

This section points at those aspects of Louw's theory that are particularly relevant to the present practical study, and will in part draw on examples selected for the subjects' interpretation.

The underlying principle of Louw's stylistics is that "text reads text". It means that no concepts are necessary in order to interpret literary or non-literary texts, but that all we deal with is the reference corpora as the norm against which we judge the text's deviation. Concepts, that is, "the ideas meaning of words", according to Louw (2008: 248, following Firth 1957: 181), are an unnecessary imposition that obfuscates the meaning of a text, while all we need for its successful interpretation is raw data accessible through large reference corpora. This is summarized by the title of one of Louw's papers, "Collocation as instrumentation for meaning" (Louw 2010a), where collocation is seen as a tool which constantly creates meaning through co-occurrence. Situational meaning (meaning in the context of situation and culture) is created in the form of events. A target text contains an event, comparable against similar events in the reference corpus and, therefore, in the world as represented by a balanced and representative reference corpus, created especially for the sampling of the world and creating its dictionary. A line of best fit usually subsists between the target event and those in the reference corpus.

According to Louw, any text can be read against the background of similar texts and the events they represent. Contextual interpretations of keyness are still in their infancy (Louw, personal communication). What is the similarity that qualifies corpus data to act as a background to such a reading? There can be many instances of this, but let us look at a few.

A key word in an authorial text may be 'checked' in the reference corpus to see in what sort of contexts it tends to occur. This analysis may result in revealing a semantic prosody. The author's usage may be clarified by establishing that there is a positive, negative or specific prosody in the language. Alternatively, a clash between the author's use and general usage may improve interpretation in discovering a conscious irony or a subconscious insincerity. In the well-known (to those familiar with the term 'semantic prosody') example of David Lodge's *Small World*, it is said that conference goers are bent on self-improvement (Lodge 1984), whereas in a reference corpus the prosody of *bent on* is

negative and points to destructive intentions. Many semantic prosodies, though perhaps not all, have been mentioned in dictionaries as part of definitions. As native speakers do not consult dictionaries, they may thus remain unaware of certain semantic prosodies in the language, as the example with *cook up* will later show. It is a logical assumption that, where a search in a reference corpus detects a semantic prosody that is used to create irony, irony in the target text cannot be ruled out, as an ironic intention is a definite likelihood.

A similar analysis studying a key word within an authorial corpus may reveal a semantic prosody that is indicative of the author's attitude throughout the corpus of his work, e.g., when Larkin uses the word *day* in prevalently negative contexts and *night* in more positive ones. An absence of a word from an author's corpus may also contribute to the understanding of his work, e.g., Larkin uses the word *night* 72 times, but the word *hope* only nine times. The case of semantic prosody means that collocates of the node in the reference corpus, however diverse, throw light on its usage in the language, if the 'aura' is persistent. In the case of absent collocates, however, there are specific collocates that keep re-appearing. These 'usual', expected collocates may be 'replaced' by an unusual one in the authorial text. For example, in Adrian Henri's poem entitled "Drinking Song" in the line *as the afternoon wore off*, a native speaker will feel that drugs usually wear off and afternoons usually wear on. A non-native speaker will feel this to the extent of his/her proficiency in the language. The non-native speaker will feel this as some kind of word play underlying a metaphor, but a corpus will show what exactly has been replaced. In the quoted line, for example, even a native speaker, when reading, may only notice the abrupt ending of the afternoon, but not the 'drug' part of the pun, which makes it metaphorical. Absent collocates thus contribute to the interpretation of a text and may also help in subtler cases that are not so obvious to the 'naked eye', not as yet armed with corpus experience. It is conceivable that second or foreign language speakers may have been less exposed to the full range of collocates in specific situations. This will mean an increased level of difficulty in dealing with events fractured because of the omission, so they need collocation lists and contexts to unpack them. An increased level of difficulty presents no problem in the 21st-century because of the availability of corpora.

Not only lexical items, but grammatical strings too, may have fairly specific prosodies of their own (Milojkovic 2012). If we look at the case of *but when did*, it becomes crucial to our interpretation of Larkin's notion of love in the following line, which in its context seems to be positive:

> Admitted; and the pain is real.
> ***But when did* love not try to change**
> **The world back to itself** – no cost,
> No past, no people else at all –
> Only what meeting made us feel,
> So new, and gentle-sharp, and strange?

Studying the line starting with *but when did*, we discover that *but when did* introduces angry rhetorical questions, like *But when did a car salesman ever tell you that you had better walk or take a bus?* The semantic prosody is one of futility.

I believe that the angry and rhetorical part in Larkin's line is subconscious, and that it is a variation on the general irony/insincerity division. It is insincere inasmuch as it is not what the person really feels. Larkin *wants to* believe in love's power to change the world and the rhetorical anger is for the most part subconscious. By the way, when asked to judge if Larkin's line in question is or is not optimistic, out of 52 second-year students of English, 37 (71%) replied that it is.

Interestingly, *but when did* found its way into David Lodge's *Small World* too:

> "I have just one question," said Philip Swallow. "It is this: what, with the greatest respect, is the point of our discussing your paper if, according to your own theory, we should not be discussing what you actually *said* at all, but discussing some imperfect memory or subjective interpretation of what you said?"
> "There is no point," said Morris Zapp blithely. "If by point you mean the hope of arriving at some certain truth. **But when did** you ever discover *that* in a question-and-discussion session? Be honest, have you ever been to a lecture or seminar at the end of which you could have found two people present who could agree on the simplest precis of what had been said?"
> "Then what in God's name *is* the point of it all?" cried Philip Swallow, throwing his hands into the air.
> "The point, of course, is to uphold the institution of academic literary studies."
>
> (Lodge, *Small World*, my emphasis in bold)

Therefore, this classroom corpus stylistics research focuses on semantic prosodies (whether they surround a word, a phrase or a grammatical string), which may or may not be consciously felt by native speakers. They also may or may not have been fully captured by dictionaries, which is a point worth dwelling on. Apparently, there are two factors which might restrict the elaboration of semantic prosodies in dictionaries. The first one is obviously space. The second has to do with the capacity of a definition to fit in with any instance of a word's or expression's use. If we suggest that

but then, or *but then again* in the majority of cases has a narrower meaning to it than expressing a slight contradiction or lack of surprise (COBUILD), or even than lessening the meaning of a previous sentence or lack of surprise (Longman), the explanation must fit every text we encounter. But it is in the nature of a prosody not to extend its aura to every single case. Also, thorough research would need to be carried out on *but then* and *but then again*, involving the study of contexts of situation of specific sentences, which might justify devoting a whole paper to it. Naturally, it may be impracticable when compiling a dictionary, although perhaps not inconceivable in the future. For example, *but then* and *but still* are different, and not only in the degree of contradiction. Since creating a dictionary involving all nuances of meaning is practically impossible, that is where corpus stylistics proves useful in relation to a text's interpretation. It is hypothetically possible that the aura of the comforting *cest la vie* meaning of *but then* is extended to the remaining 32% not carrying it at first sight, if one studied the contexts of situation for the remaining 32%. This needs to be researched in the future, and will hinge upon discovering collocates that differentiate these two senses. A question of what is and what is not to be found in corpus-based dictionaries and why could come from students studying a given concordance. For example, some sceptical and conservative subjects of this study were surprised to learn that the negative prosody of *bent on* is recorded in the COBUILD dictionary.

The other methodological question that bears upon lexicography, corpus stylistics and corpus stylistics pedagogy has to do with the amount of percentage of occurrences of a specific tendency in meaning. How far does the aura have to spread for the word or phrase to be considered as bearing a prosody? Is 60% enough? 65%? 75%? How much is enough for a lexicographer? How much for a stylistician? What do we tell the students in the classroom who attempt to interpret texts? Do we tell them to rely on corpus-based dictionaries for prosodies in order to avoid error?

The answer to these questions has to do with the term 'events' (Louw and Milojkovic 2014). The mere proportion of lines showing a prosody is not sufficient proof that a text should be interpreted in a certain way. In the tradition of the philosophy of language and its postulates (Russel 1948) and according to the notion that a reference corpus is a representation of the world and its dictionary, we must look for *similar* events. The lines that carry the investigated pattern, whether based on a grammatical string or a lexical collocation, and describe similar events are the ones to be considered. This advanced notion was not shared with the second-year students, who were instead simply at the mercy of concordance lines. However, 10% of them did mention that in a concordance certain lines corresponded in meaning to the one in the studied text, which must have involved their studying contextual clues in these concordance lines. This finding shows that, at least to these students, text does read text, as they could see the notion of events without previous instruction.

3. QUANTITATIVE RESEARCH

3.1. The learning phase

This section of the paper will describe the tasks given to students during the 'learning phase' of the experiment, the instructions they were given and their response. There were five learning sessions on the whole, one task per session. The sessions took place at the beginning of regular *Integrated Skills* classes.

At the very first encounter the students were given full concordances from the authorial corpus of Philip Larkin with the nodes *day*, *night*, *light* and *God*.[4] As Larkin's use of these words differs drastically from the conventional – *day* is viewed pessimistically, *night* brings relief, *light* appears as dark, and *God* is doubted –, this was a good opportunity to illustrate to the students how meaning is created in context through co-occurrence with other collocates. The students were given basic corpus linguistics terminology – the 'node' and the 'collocates', as viewed in corpus linguistics, and Sinclair's '9-word window'. The relevant collocates were given in bold to facilitate comprehension at this first encounter. For example, this is the concordance with *God* as the node:

[4] My paper on Philip Larkin (Milojkovic 2011b) is available online, with the four concordances given in full: <http://www.belgrade.bells.fil.bg.ac.rs/Bells 3.pdf>.

MicroConcord search SW: god
80 characters per entry
Sort : 1L/SW unshifted.

1 that inspired it all, And made him **a** god. No, he would never fail. Others, of c
2 ortraits of Sex Sun. Tree. Beginning. God in a thicket. Crown. Never-abdicated c
3 e, **musty**, **unignorable silence**, **Brewed God knows** how long. Hatless, I take off My
4 the sky, Asking **to die**: 'To **die**, dear God, before a **scum of doubt Smear** the whol
5 pausing, goes into a prayer **Directing** God about this eye, that knee. Their heads
6 any nights, as many dawns, **If finally** God grants the wish. ~2 February ~950 Dece
7 go on before us, they Are sitting in God's house in comfort, We shall see them
8 ey need; And famous lips **interrogated** God Concerning **franchise** in eternity; And
9 And thought, That'll be the life; **No** God any more, or **sweating** in the dark Abou
10 ' Let it be understood That **'somehow'** God plaits up the threads, Makes 'all for
11 , and lips bleeding. Yes, gone, thank God! Remembering each detail We toss for h
12 tor **clenched** his **fists** And **swore** that God exists, **Clamping** his features **stiff** wi
13 adio's altarlight The **hurried** talk to God goes on: Thy Kingdom come, Thy will be
14 mit with his gown and dish Talking to God (who's **gone too**); the big wish Is to h
Data from the following files: ZARKIN.CTX

Figure 1. Concordance for *God* as node

At the second encounter, the students were given the excerpt from David Lodge's novel *Small World* quoted in Louw (1993). Unlike the first introductory session, the second one involved a task that was to be completed individually in a test-like format before the feedback and the 'right' answers were given. In the task, the students were first asked several comprehension questions, and then given chosen concordances from *The Times* corpus, printed in a bigger size than usual and with the significant collocates in bold. The concordance had been edited to facilitate understanding, without the students' knowledge, as the unedited concordance might have discouraged them. In this session, the students were taught the notion of semantic prosodies and explained that the prosody of *bent on*, which they had just discovered, is in the dictionaries. After it was elicited from the class that Lodge is in fact being ironic, Louw's irony/insincerity dichotomy was explained to the students. This is the format of the task:

Session 2 (Semantic Prosody)

Consider the following short passage from the novel *Small World* by David Lodge:

> The modern conference resembles the pilgrimage of medieval Christendom in that it allows the participants to indulge themselves in all the pleasures and diversions of travel while appearing to be austerely *bent on self-improvement*.

1. Explain the meaning of the passage in your own words, either in English or Serbian.

2. Explain the meaning of the phrase *bent on self-improvement*, either in English or Serbian ... and translate it into Serbian ☺

3. Look at the concordance lines of *bent on* taken from a large reference corpus:

1 in a society **hell** bent on achievement. Mutable thinkers don't
2 werful enchanter, bent on **bringing Arthur to his ruin** this dev
3 iding **donkeys** all bent on business, they were forcibly impress
4 cter development. Bent on change, **even to the point of shatter**
5 r world she seems bent on **conquering**. Well, I suppose that you
6 overnment is **hell** bent on demanding greater and greater protec
7 of Yoller's wood, bent on **destroying all survivors** before purs
8 he people who are bent on doing good **they can be the danger**, s
9 stic **savagery** and bent on **engulfing and drowning trapped men** a
10 **sonal safety** and bent on **escaping not only the enemy**, but the

Judging by the random lines taken from the reference corpus, how is *bent on* usually used? How does it influence your understanding of the line? What is the author implying? Do you think that intuitively you felt this at the first reading?

Figure 2. Session 2 (semantic prosody)

The third session focused on the notion of 'absent collocates'. This time the relevant collocates were not given in bold and the concordances were not edited. The task proved a success not only because the students could see the role played by absent collocates, but also because learning phrasal verbs is always of interest to non-native speakers. Whereas during the previous session the subjects reacted with surprise to the notion of semantic prosody, this time they

understood the practical advantage of establishing the meaning of a word through its collocates in the reference corpus. This is the format of the task, whose impact is briefly discussed in Section 2.2:

Session 3 (absent collocates)

DRINKING SONG (Adrian Henri)

He became more and more drunk
As the afternoon wore off.

```
                              MicroConcord search SW: wore off
                                   80 characters per entry
                                   Sort: 1R/SW unshifted.
 1 ut after a few minutes the stinging wore off and I began to enjoy the exquisite
2 y to stare at him. The novelty soon wore off, however, as Smythe persistently re
 3 lbeing that it was months before it wore off. I am still trying to remember wh
 4 have knocked him out for half a day wore off in a fraction of the time, and for
5 ing glissandi. But the novelty soon wore off. Michael Thomas disarmingly explain
6 men using implants said the effects wore off more quickly, and 29% said they nee
7 ut eventually, inevitably, the drug wore off. Some say it was Fortensky who call
  8 increase the dosage as the effects wore off. "What we have done is to establi
 9 his clean-cut approach, the novelty wore off when they realised how much pocket
                        Data from the following files: TIMES95.TXT

                              MicroConcord search SW: wore on
                                   80 characters per entry
                                   Sort : 1R/SW unshifted.
 1 red, four years ago, and as the day wore on a repeat looked ever more probable.
2 ed Thatcher and Major. As the night wore on a swing to the right, whether or not
 3 only obscured my face but, as time wore on,had a horribly isolating effect on m
  4 things were to change as this game wore on. After 17 minutes, Durrant put the
 5 on to the bat and off it as the day wore on. Although he found life more diffi
 6 officials reported that as the day wore on an ever-growing crowd of terrified o
7 Corsie's form improved as the match wore on, and to whitewash a player of Schuba
8 w more perfunctory as the afternoon wore on and finally ended up with Stewart ha
 9 was only sustained, as the evening wore on and got colder, by the particular in
 10 tive electoral history. The night wore on and the flow of Tory setbacks mounte
  11 did get harder, and time certainly wore on. And on... But first, Wimbledon. T
12 d dummy runs got better as the game wore on and some of his early ones were no m
13 had become even firmer as the match wore on and were not going to be pulled arou
14 ut he played ever better as the day wore on and, after Ilott had returned to dis
                      Data from the following files: TIMES95.TXT
```

Comment:

Figure 3. Session 3 (absent collocates)

The fourth session focused on grammatical strings and their prosodies. The subjects were first asked whether they perceived the lines containing the grammatical string under study as positive or negative. Then, they were given the whole concordance from *The Times* corpus together with the wider contexts of four chosen concordance lines. The wider contexts were provided in order to better illustrate the relationship that is usually established between the clause starting with *but when did* and the surrounding text. The wider contexts of lines 1, 2, 3 and 6 were chosen because they were easier to understand than others, since newspaper language with its ironies, sarcasm and sophisticated vocabulary is not always easy for non-native speakers. This is the format of the paper:

Philip Larkin

When first we met, and touching showed
How well we knew the early moves
Behind the moonlight and the frost
The excitement and the gratitude
There stood how much our meeting owed
To other meetings, other loves.

The decades of a different life
That opened past your inch-close eyes
Belonged to others, lavished, lost;
Nor could I hold you hard enough
To call my years of hunger-strife
Back for your mouth to colonise.
Admitted; and the pain is real.
But when did love not try to change
The world back to itself – no cost,

Is this an optimistic reference to love? YES/ NO

No past, no people else at all –
Only what meeting made us feel,
So new, and gentle-sharp, and strange?

Comment (optional)

```
                    MicroConcord search SW: but when did
                          80 characters per entry
                          Sort : 1R/SW unshifted.
 1 there's nothing wrong with that. But when did a car salesman ever tell you that
 2 s may make little economic sense, but when did economics really come into the eq
 3 yd and Rob Lowe also participate, but when did either last make a prudential car
 4 ties in both manager and country. But when did England last have success or a pu
 5 be the logical time to bow out. But when did football, life, logic, Charlton o
 6 tiality, and has become a cliche. But when did that deter anybody? </Group>
 7 abs. The rot set in after that. But when did the present system start, and why
 8 company, making £2billion a year. But when did you last hear critics sounding of
 9 hormone is, of course, a cop-out, but when did you last hear of a netball crowd
                       Data from the following files: TIMES95.TXT
```

1 Banks, insurance brokers and estate agents sell their products and there's nothing wrong with that. *But when did a car salesman ever tell you that you would be better off walking or taking a bus?*

2 A politically imperilled Government will probably still opt to cut taxes instead. This may make little economic sense, *but when did economics really come into the equation so close to a general election?*

3 TOMMY BOY, 97 mins, PG

After Dumb and Dumber, we now have Dumbest to date. Starring Chris Farley, yet another dubious Saturday Night Live Graduate, this is not so much a comedy of errors as an error of comedy as our hero takes over the family car-brake business when his father (the much-abused Brian Dennehy) dies from over-exertion caused by marrying Bo Derek. Dan Aykroyd and Rob Lowe also participate, *but when did either last make a prudential career move?*

4. Two related programmes on BBC2 focus on elephants and their would-be savior, Richard Leakey. *The Savage Paradise* (Monday, BBC2, 8pm) is billed as an intimate portrait of an elephant herd in Botswana, while Leakey is profiled in *Africa's Wildlife Warrior* (Wednesday 9.30 pm).

The green devotees will doubtless tune in to *Witness: Beyond the Rainbow* (C4, Wednesday, 9 pm), in which the daughter of a photographer killed in the sinking of the Rainbow Warrior embarks on a quest to find out more. This approach to documentary-making virtually ensures partiality, and has become a cliché. *But when did that ever deter anybody?*

What is the tendency of meaning in sentences starting with *but when did*? How does it influence your understanding of Larkin's lines? Is your perception different now?

Figure 4. Session 4

As mentioned earlier, on the first reading, 71% of the subjects saw the reference to love in the poem as positive. Of these students, only one fourth (24.7%) changed their views completely after studying the concordance lines and the wider contexts, and concluded that the implication of the lines was in fact negative. During the discussion time, the views expressed by groups B and D differed substantially. Group B claimed that the poet was intentionally ironic. Group D claimed that the concordance lines in the reference corpus, especially as it was a newspaper one, had nothing to do whatsoever with the poem and the poet. The poet, they claimed, was free to use the language as he pleased. My suggestion that a grammatical string is a basic unit in a language and may therefore be studied in a newspaper corpus as

well as anywhere else, failed to convince. Neither Group B nor Group D agreed with my hypothesis that the poet intended to make a positive statement while subconsciously he did not believe in the power of love to change the world back to itself.

The fifth section of the experiment will not be described here in detail as it was unsuccessful. During the fifth section the students were introduced to the idea of 'subtext' on the basis of John Donne's poem "The Good Morrow". The poem was chosen because the students were well familiar with it from their literature class. From the point of view of subtext, the poem is not easy to interpret. On the other hand, the choice of Yeats's "Sailing to Byzantium", for example, quoted in Louw (2010b), would have been more appropriate from the point of view of subtext, but difficult to deal with in the classroom, as the students had not yet studied it in their literature class. It was also my impression that the students may have received too much new information in a very short time. The time was limited and I did not pursue the notion of subtext further. Had the experiment been conducted as part of a stylistics course, I would have been justified in spending more time on teaching subtext to students. As it was taking up the time of an altogether different course, with different aims and a set curriculum, I abandoned subtext until a better opportunity arose. This does not disprove the principle of "text reading text", but rather calls for more time spent on explanation and classroom practice.

3.2. The testing phase

All tasks in the first four sections were mirrored in the final test. In the first task, the students were given an edited concordance with the node *hope* from the authorial corpus of Philip Larkin. The second task had to do with a negative semantic prosody of the phrasal verb *cook up*, the prosody not being obvious solely on the basis of the context. The text itself came from Leo Jones' *New Progress to Proficiency* (2001: 112) and had recently been studied in the classroom, so the subjects were familiar with the wider context as well, although this knowledge was not strictly necessary. The third and the fourth tasks focused on the grammatical strings *but then* and *but what is*, respectively. In the third task the students were invited to analyze wider contexts from *The Times* corpus, and in the fourth the grammatical string needed to be interpreted on the basis of a concordance.[5] The fifth task was based on the third step in the learning phase and had to do with the absent collocates of a phrasal verb. In this case, in order to save time and also to vary the tasks, the concordance itself was skipped. The students were given the result of the analysis of the concordance and asked to connect it to the studied excerpt from another poem by Larkin. Here is the format:

1. What is Philip Larkin's view of hope? You have before you the contexts in which he used the word *hope*. In his authorial corpus there exist nine lines overall. Lines 6 and 9, from a birthday poem to a friend, and lines 1 and 4, from a jocular last will and testament, have been omitted as they belong to occasional poetry, and therefore not likely to express the poet's true attitude.

```
                    MicroConcord search SW: hope
                       80 characters per entry
                       Sort : 1R/SW unshifted.
  2 signalled in attics and gardens like Hope, And ever would pass From address to
  3 claims The end of choice, the last of hope; and all Here to confess that somethi
  5 what I desired - That long and sickly hope, someday to be As she is - gave a fli
  7 e Through doubt from endless love and hope To hate and terror; Each in their dou
  8 it's a different country. All we can hope to leave them now is money. 10 lanuar
                    Data from the following files: ZARKIN.TXT
```

Comment:
1. Read the following familiar text from *New Progress to Proficiency* by Leo Jones. Then read the given concordance lines from *The Times* reference corpus.

> The idea of preserving biological diversity gives most people a warm feeling inside. But what, exactly, is diversity? And which kind is most worth preserving? It may be anathema to save-the-lot environmentalists who hate setting such priorities, but academics are starting ***to cook up answers***.'

```
                    MicroConcord search SW: cooked up
  1 and Pacific supermarket, and so I cooked up a story called A & P. I drove my da
  2 h of Euro-scepticism, and you have cooked up a crisis." Tory Euro-sceptics wil
  3 a stream of mixed notices, having cooked up a storm in America. "Crime in exces
  4 ister for the energy industry, had cooked up a £1.2 billion payout to them from
  6 fact that this whole exercise was cooked up by a record company executive, and
  7 tarting to resemble a cynical ploy cooked up by lenders to force the government'
  9 st demand for tax-planning schemes cooked up by Jenkins and his colleagues, whic
 11 liance claimed the affair had been cooked up by the Russians in an attempt to de
 13 e fallen for every publicity stunt cooked up by the lawyers in the Simpson case
```

[5] All three concordances in the final test were edited – given that the time was limited – and some unmotivating lines were removed, but the necessary level of difficulty was preserved.

```
 17 ese than the Mayan extravaganza he cooked up. Certainly the claim that the build
 18 mmary of the predicament Slovo has cooked up for her headstrong part-time detect
 20 has proved controversial, but was cooked up in close consultation with Major. "
 22 one knew that. Whatever scheme was cooked up, London would rally to the common c
 23 mmon murderer. Salvatore Cammarano cooked up the plot and later provided somethi
 25 , 20, said: "It was the father who cooked up the plot to say the car was stolen
 27 was it. Yet the officials who had cooked up this crass plan bounced councillors
 29 ant turns out to be less than it's cooked up to be, and Connie's disillusionment
 30 ed up the National Lottery (I said cooked up, you sniggering lot) and departed b
                    Data from the following files: TIMES95.TXT
```

2. Read the following poem by Philip Larkin, REASONS FOR ATTENDANCE:

The trumpet's voice, loud and authoritative,
Draws me a moment to the lighted glass
To watch the dancers – all under twenty-five –
Solemnly on the beat of happiness.

– Or so I fancy, sensing the smoke and sweat,
The wonderful feel of girls. *Why be out here?*
But then, why be in there? Sex, yes, but what
Is sex? Surely to think the lion's share
Of happiness is found by couples – sheer

Inaccuracy, as far as I'm concerned.
What calls me is that lifted, rough-tongued bell
(Art, if you like) whose individual sound
Insists I too am individual.
It speaks; I hear; others may hear as well,

But not for me, nor I for them; and so
With happiness. Therefore I stay outside,
Believing this, and they maul to and fro,
Believing that; and both are satisfied,
If no one has misjudged himself. Or lied.

The following contexts are taken from the 1995 *The Times* corpus:

It's just that art students, and art critics for that matter, spend a lot of time in galleries thinking about sex. *But then*, everyone used to go to galleries to think about sex.
Ibsen himself was subject to fits of depression, so he wasn't one for light entertainment. *But then*, few Norwegian entertainers are.
'That reminds me,' he said, 'did you translate the poem?'
I brought out a grubby piece of paper, made soft by much handling, [and read my translation].
'It's not bad,' said Daniel, 'but you didn't do the rhymes.'
'Are you kidding? Look at the rhyme scheme: a,b,b,a,b,b,c,d,c,d. It's impossible.'
Daniel sniffed. 'Paul-Jean Toulet did it,' he said. '*But then*, French is a richer language than English.'
The end of the Mozart story is tragic and you may even weep, as I did, as you read this affectionate account of his last days. Mozart's life could easily have been so much happier. *But then*, considering those 626 works in the Kochel catalogue (= a complete, chronological list of Mozart's works), would we really have things otherwise?
He found her beautiful and alluring. *But then*, eligible man-about-town Hewitt finds many women beautiful and alluring. Nobody has denied, however, that it was Diana who started the serious flirtation that led him to her bedchamber.
'Our love keeps us going.' It's not easy living in a Frankfurt jail, *but then again* it wasn't easy living with the guilt and angst of running the error account on Barings Bank and meddling with millions, other people's millions, as if they were Mars bars.
'Sure, I might meet someone nice, *but then again* I might meet someone I don't want to meet.'

Take into account the contexts you have just read. How do you understand the *but then* line in Larkin? Give reasons derived from the *The Times* contexts.

3. Now read the contexs of *but what is* found in the *The Times* 1995 corpus:

```
                        MicroConcord search SW: but what is
                              80 characters per entry
                              Sort : 1R/SW unshifted.

  1 f of 1% of total public spending. But what is a majority taste? Nothing, really,
  2 been there the old money of Eton, but what is a school to do with a boy who, rec
  3 r desk pontificating arrogantly". But what is a columnist for if not to pontific
      4 onOfPaper> <Story> <Group> But what is a beer without a hangover? Wine-
 36 th the price war with Wordsworth. But what is going on here? If selling 99p book
   42 are ball into the six-yard box. But what is he meant to do, other than act as
 43 preference holders. Clear enough; but what is he doing upping his stake in anoth
  56 aling, though possibly necessary. But what is it all for? "Have some knowledge
 58 ibition spaces in central London, but what is it beyond that? The Academicians t
```

```
 73 r remains far removed from normal but what is normal behaviour for a king? Seizi
 79 oes not look like a comic genius, but what is one supposed to look like? As he l
 80 enstern, all we get is incidents; but what is our role in these incidents? Have
  95 claims are "exorbitant demands". But what is reasonable? Sybil Gooldrich, one
 97 rought down to earth with a bump. But what is risk, and how can you avoid it? Th
 112 deaf members of the audience." But what is the point of interpreting opera fo
 114 ood as Claridge's or at my house, but what is? The clientele were an odd mix,
116 ould love to lead the Government, but what is the point if the party is too asha
117 he accused sold were not genuine. But what is the difference between a genuine l
118 ng as many sights as time allows. But what is the rush? Rome was not built in a
119 ow inflation and a trade surplus. But what is the point? This is a question whic
 127 s is impressive and he is excited but what is the reality? Market forces in ed
 130 dangerous Rollerblades can be. But what is the attraction? Unlike traditional
139 ically/she's using him fiscally." But what is this thing called friendship? When
144 ays before an international game. But what is to stop the clubs refusing to sign
                    Data from the following files: TIMES95.TXT
```

After you have read the concordance lines, how do you understand the *but what is* line in Larkin? Give your reasons, basing them on the concordance lines.

4. And everywhere the stifling mass of night
   Swamps the bright nervous day and puts it out.

The lines come from the poem 'Midsummer Night' (Philip Larkin again) which deals with transition between day and night. The phrase *it out* was searched in *The Times* corpus, and 195 lines were found. *Out* was mostly a particle belonging to a phrasal verb, with *it* as its direct object, like *carry it out*, *pull it out*, *sort it out*. Mostly the underlying argument in the concordance lines was that the action described by the phrasal verb was intended to solve a problem. Four concordance lines contained *put it out*. In all the four lines what needed to be put out was a great fire.
How would you apply this knowledge to the interpretation of the lines from the poem?

Figure 5. Final test

3.3. Discussion of results in the quantitative phase

All the students' tests, in both the learning and the testing phases, were marked. In the tables below, the tasks in the learning phase are marked as 'a', for example, 2a, 3a, etc. They are juxtaposed with the results from the testing phase, marked as 1b, 2b, etc. Column '1a' is empty because the subjects' answers were not graded during the first session. Column 4a is empty as task 4b is focused on a grammatical string and as well as 3b.

These results were processed in the following way. The scores of both groups were entered into separate tables. The students were given marks from 0 to 5. Grade 0 was given to students who were present, but left a blank space instead of doing the task. In practice, since a completely wrong answer was given the mark of 2, no one was given the mark of 1. In cases of exceptionally acute judgement the mark was 6 out of 5. The average results per each task were calculated, for the two groups separately as well as for all subjects together. The average marks were calculated per each group and for all the subjects both in points and percentages (the mark 5 points for all the students was accepted as 100%). Statistical treatment was performed in an *Excel* spreadsheet. The final marks were plotted in a diagram, for the two groups separately, as well as for all the students together.

Group B	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	sum	%
Average		5	5	4.2	4	4		4.3	4.6	4.1	21.38	85.509
Standard deviation		0.7	0	0.7	1	1		1	0.4	1.1	2.20	8.8095
Number of students present		22	9	22	20	22		22	18	22	22	22

Table 1. The results of Group B

Group D	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	sum	%
Average		4.5	5	4.1	3	3		4.1	4	3.7	19.68	78.714
Standard deviation		1.1	0	0.8	1.3	1		1.4	1	1.8	3.246	12.982
Number of students present		28	10	28	24	28		28	23	28	28	28

Table 2. The results of Group D

Group B + D	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	sum	%
Average		4.7	5	4.14	3.33	3.58		4.14	4.28	3.87	20.4	81.7
Standard deviation		0.95	0	0.76	1.28	0.86		1.22	0.86	1.51	2.93	11.74
Number of students present		50	19	50	44	50		50	41	50	50	50

Table 3. The results of the two groups together

The difference in the marks of the two groups for the 'b' tasks is statistically significant (p=0.041). It is obvious that the subjects from group D scored fewer points than those from group B. Figure 6 shows the plotted curves of the distribution of marks for both groups and for all students together.

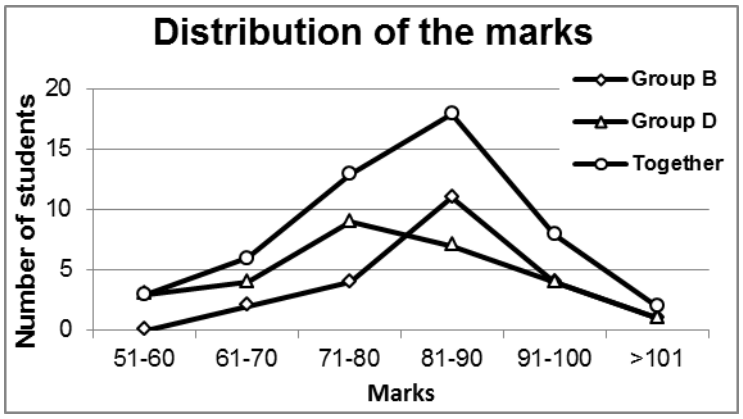


Figure 6. The distribution of final marks

Although the results scored in group D are lower than in group B, which confirms the initial assumption that the more proficient group would score better results, the results of both groups suggest that the difficulty of the tasks was adequate, since all three curves could have been obtained after any undergraduate course of moderate difficulty.

As the first research question is to see whether "text reads text" for Belgrade students of English, it is important to establish the percentage of students who scored the highest marks on the final test (5 or 6). Given the study's limitations, the figures seem to suggest that the research question has been answered positively:

- Question 1b (*hope*) 62%
- Question 2b (*cooked up*) 34%
- Question 3b (*but then*) 14%
- Question 4b (*but what is*) 30%
- Question 5b (*put it out*) 32%

The percentages of students who scored the mark of 5 or 6 are even more important for our research question than the previously given tables and plots, as the mark of 4 was given to students who correctly interpreted the concordances, but could not or would not see the connection between the concordances and the studied text.

Another finding which is relevant to the research question is that five students out of fifty (10%) made comments when studying concordance lines in 4b (*but what is*) that could be construed as attempts to look for similar events in the reference corpus. As such a method of interpretation had not been mentioned in the classroom, this also confirms Louw's stance that "text reads text".

4. QUALITATIVE RESEARCH

Attached to the final test was a questionnaire consisting of 11 questions whose aim was to see how well the students understood the point of the course, whether they found it useful, what they thought of the methodology, whether they enjoyed it and whether they would choose it if it were on offer. Here is the format of the questionnaire:[6]

Questionnaire

Please read all the questions first before answering.

1. What is corpus linguistics?
2. What is corpus stylistics?
3. What is stylistics?
4. Would you have appreciated being given more terminology and background when doing classroom corpus stylistics?
5. Do you feel you have learnt something from this course? What?
6. What can a foreign student at your level of knowledge learn from this course?
7. In your view, what can a native speaker learn from this course?
8. What was your overall view of the teaching methodology?
9. Do you feel you have been encouraged to develop your own opinion?
10. How difficult did you find the course? What might have caused this?
11. Did you enjoy the course? If corpus stylistics was on offer at this department, would you consider choosing this subject?

Figure 7. Final questionnaire

The first three questions were aimed at discovering what definitions of 'corpus linguistics', 'corpus stylistics' and 'stylistics' the students would give after being exposed to a short course that, in itself, was based on the principle that "text reads text". The researcher was curious to see how the subjects understood these three disciplines. Any answer that was not wrong was marked with a 'yes', wrong answers were marked with a 'no' and the absence of the answer was marked with a '0'.

Out of 50 students, 52% gave acceptable definitions of 'corpus linguistics'; 48% defined 'corpus stylistics' (6% more defined it by means of the word *style*) and 32% defined 'stylistics' (16% more defined it by means of the word *style*). I have separated the answers which depended on the word *style* as it seems too vague in the circumstances of this particular research, so it is not certain what the subjects actually meant and how they defined style as such.

In the subsequent questions, positive answers were marked with a 'yes' and negative ones with a 'no'. To preserve this principle of describing answers as 'yes' if the feedback is positive and 'no' if negative, in question 10 'yes*' denotes that the course was not found difficult by the student, and 'no*' means that it was found difficult.

Half of the subjects (50%) suggested they would have liked more terminology and background, and 30% said they would not have wanted more. To all three subsequent questions (5, 6 and 7) as many as 70% of the subjects replied in the affirmative, while showing sufficient understanding of the point of the course (positive answers that showed that the student did not understand the main point of the course were not marked with a 'yes'). The adopted teaching methodology was approved of by 72% of the subjects, and 48% stated that the course was not difficult.

In the last question, consisting in fact of two, one related to the enjoyment of the course and the other to whether the student would choose it if it were on offer, the adopted description of the answer was, e.g., 'yes/no – the student enjoyed it but would not choose it'. This is the distribution of answers: 'yes/yes': 36%'; 'yes/maybe': 16%; 'yes/no': 22%; 'no/no': 14%; 'no/yes': 4%; and no answer: 8%.

The results of the qualitative survey suggest that the impact of the course was overall significant, and the subjects' reaction was positive. More than 70% of the subjects claimed it was useful, approved of the methodology and stated they had enjoyed it. These percentages would have been higher had the subjects' affinities and interests been consulted. All this has a bearing on the second research question asked in this study – whether the CPT-based methodology proposed in this paper proved to be successful.

5. CONCLUSION

The two interrelated research questions posed in this study are a) if "text reads text", i.e., if reference corpora alone and without theoretical concepts can help interpret authorial text, and b) if the proposed CPT-based classroom stylistics

[6] The questionnaire was in part inspired by Burke (2004).

methodology can be successful. After the quantitative and qualitative survey conducted in this study, there is sufficient foundation for the claim that both research questions have been answered positively. The percentage of students who completed the final test successfully, together with the fact that there was almost no theoretical instruction, proves that text does read text for the non-native students of English at the Belgrade English Department, so far as it can reasonably be expected. The fact that the more proficient group achieved noticeably better results suggests that native speakers would have been even more successful, but that the principle is the same regardless of the level of proficiency. The implication is that the degree of tasks' difficulty may – or should – vary depending on the students' or pupils' proficiency. The feedback gained on the qualitative part confirmed the success of the teaching methodology. Both quantitative and qualitative results in fact exceeded the researcher's expectations, given the study's limitations, which deserve some dwelling upon.

Firstly, had only the motivated students been tested, the final results might have been even more encouraging. As things stood, a fair amount of students were not particularly interested in poetry, corpora or stylistics. Another issue is the level of personal, and not linguistic, maturity, which will be reflected in a stylistic interpretation of the poem or of the lines. Personal maturity in the issues of love, hope, despair or resignation is a factor altogether different from, for example, the inability or refusal to see that a reference corpus does have a bearing on a particular author's meaning. The subjects, in this case young adults, may not have the experience that the (middle-aged) author has tried to convey and, therefore, cannot interpret what they have not understood. When interpreting a poem or part of it through concordance lines, it is first necessary to check each student's literal understanding and the degree of their appreciation of the text as readers. This was not done and all students were tested in the same way. Thus, the findings may shed light on how an average generation of second-year students may react to this sort of course, but for finer nuances of the process of interpretation a more detailed study ought to have been conducted.

It is also worth noting that no proper course of corpus stylistics would have been founded on such a minimum of instruction. Therefore, the point of this research, again, is to see how the subjects react to texts. However, in real life, more students would have responded to this kind of teaching positively after reading on semantic prosody. Some students may have understood the point, but were too conservative to believe it, as it may not have fitted into the manner of their dealing with text and meaning in their previous schooling. They could have changed their minds after reading a couple of papers containing examples. As things stand, some of them may have been too conservative to start seeing things differently. For example, one of the students, generally proficient and hard-working, could only see grammatical usage in the lines, but never auras of meaning. On the other hand, they might have complied with the expectations of a regular course unquestioningly, whereas the adopted way of learning showed how they felt when not under pressure.

It is obvious from the above that the marks should not be interpreted only as comments on the quality of the students' answers, but rather as a way of showing what kind of feedback a subject gave. A difference must be made between a student who is not capable of perceiving how a concordance can assist the interpretation of the studied text and a student who can see how it can be done but refuses to accept that a reference corpus can be allowed to read text. Both students were given the mark of 4 and the distinction is not reflected in the results. Besides, from their comments it was sometimes difficult to see why exactly the connection between the concordance and the studied text had not been made – whether the student was not able to make the connection or refused to make it.

Another restriction has to do with the fact that the final test was slightly more difficult than the 'learning' tests, because the answers were slightly less obvious. The students may have expected prosodic clashes where there were none and may not have been prepared for the other option, namely, that interpretation may be deepened when not changed by the concordance, especially in the case of non-native speakers of English who do not have the native speakers' accumulated experience. However, many students commented that, after reading the concordance lines, they understood the studied line in the text better. Ironically, in some cases they claimed it even if objectively they misunderstood the line.

A lack of basic skills in reading concordances on the part of the subjects was another important limitation. Overgeneralization was one of the observed errors – sometimes the first concordance line influenced the interpretation of the whole concordance. Another interesting error was misinterpretation based on the subject's personal experience of life. Also, the subjects of the study lacked experience in making sense of the syntactic structure of a concordance line, namely, they were used to the sentence, clause or syntagm as units of interpretation, rather than a concordance line that could begin or end at any point in a sentence, clause or phrase. All these issues would have been addressed on a proper corpus stylistics course.

With hindsight, my general impression is that the final test might have been too difficult for a fair number of my students, due not exactly to lack of linguistic proficiency, but to a combination of not enough English to understand all of the text and not enough general critical skills to interpret the English that they understood. It would have been sensible either to give them a poem that had already been interpreted in their English literature class (it would have had to be one by John Donne, for example) or to go through the poem with them first to ensure comprehension (which would have been difficult because of the lack of time). The results are therefore a mixed picture of enough or not enough comprehension, enough or not enough critical skills and enough or not enough of corpus stylistics performance. This is, after all, how it would turn out in a real-life situation, but one wishes for more concrete findings that would

have taken more time than originally planned. Nonetheless, the results of the quantitative part show that the tasks were not too difficult, and those of the qualitative survey suggest that the course was appreciated by the majority of its participants.

REFERENCES

Burke, Michael. 2004. Cognitive stylistics in the classroom. *Style* 38/4: 491–510.

COBUILD. 1998. *Collins COBUILD English dictionary*. London: HarperCollins.

Jones, Leo. 2001. *New progress to proficiency. Student's book*. Cambridge: Cambridge University Press.

Firth, John R. 1957. *Papers in Linguistics 1934–1951*. Oxford: Oxford University Press.

Fischer-Starcke, Betina. 2010. *Corpus linguistics in literary analysis: Jane Austen and her contemporaries*. London: Continuum.

Henri, Adrian, Roger McGough and Brian Patten. 1967. *The Mersey sound*. First edition. London: Penguin.

Larkin, Philip. 1988. *The collected poems of Philip Larkin.* London: Faber and Faber.

Lodge, David. 1984. *Small World.* Harmondsworth: Penguin.

Louw, William E. 1991. Classroom concordancing of delexical forms and the case for integrating language and literature. *English Language Research Journal* 4: 151–178.

Louw, William E. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds.), *Text and technology: in honour of John Sinclair*. Amsterdam: John Benjamins, 157–176.

Louw, William E. 2000. Contextual Prosodic Theory: bringing semantic prosodies to life. In Chris Heffer and Helen Sauntson (eds.), *Words in context. In honour of John Sinclair*. Birmingham: ELR, 48–94.

Louw, William E. 2006. Collocation as the determinant of verbal art. In Patrick D. Miller and Monica Turci (eds.), *Language and verbal art revisited: linguistic approaches to the study of literature*. London: Equinox, 149–180.

Louw, William E. 2007. Truth, literary worlds and devices as collocation. Closing keynote presentation at TaLC6 on 7th July 2004. In Encarnación Hidalgo Tenorio, Luis Quereda Rodríguez-Navarro and Juan Santana Lario (eds.), *Proceedings of the Sixth Conference on Teaching and Language Corpora*. Amsterdam: Rodopi, 329-362.

Louw, William E. 2008. Consolidating empirical method in data-assisted stylistics: towards a corpus-attested glossary of literary terms. In Sonia Zyngier, Marisa Bortolussi, Anna Chesnokova and Jan Auracher (eds.), *Directions in empirical literary studies. In Honour of Willie van Peer*. Amsterdam: John Benjamins, 243–264.

Louw, William E. 2010a. Collocation as instrumentation for meaning: a scientific fact. In Willie van Peer, Sonia Zyngier and Vander Viana (eds.), *Literary education and digital learning*. Hershey, PA: IGI Global, 79–101.

Louw, William E. 2010b. Automating the extraction of literary worlds and their subtexts from the poetry of William Butler Yeats. In Marta Falces Sierra, Encarnación Hidalgo Tenorio, Juan Santana Lario and Salvador Valera Hernández (eds.), *Para, por y sobre Luis Quereda*. Granada: Universidad de Granada, 635–657.

Louw, William E. 2011. Philosophical and literary concerns in corpus linguistics. In Vander Viana, Sonia Zyngier and Geoff Barnbrook (eds.), *Perspectives on corpus linguistics*. Amsterdam: John Benjamins, 171–196.

Louw, William E. and Marija Milojkovic. 2014. Semantic prosody. In Peter Stockwell and Sarah Whiteley (eds.), *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, 263–280.

Miller, Henry. 1966. *Tropic of Capricorn.* London: Panther.

Milojkovic, Marija. 2011a. Semantic prosody and subtext as universal, collocation-based instrumentation for meaning and literary worlds. In Victor P. Zakharov (ed.), *Трудыы международной конференции Корпусная лингвистика*. St Petersburg: St Petersburg State University, 47–52.

Milojkovic, Marija. 2011b. Quenched light, or seeing through a glass darkly – a collocation-based view of Larkin's atheism and depression. In Zoran Paunović (ed.), *Belgrade English language and literature studies*. Volume III. Belgrade: University of Belgrade, 127–144.

Milojkovic, Marija. 2012. Time and transitions in Larkin's poetry. *NAWA: Journal of Language and Communication* 6/1: 102–126.

Milojkovic Marija. 2013. Is corpus stylistics bent on self-improvement? The role of reference corpora 20 years after the advent of semantic prosody. *Journal of Literary Semantics* 42/1: 59–78.

Russell, Bertrand. 1948. *Human knowledge: its scope and limits.* London: Routledge.

Sinclair, John M. 2004. *Trust the text.* London: Routledge.

Sinclair, John M. 2006. *Phrasebite*. Pescia: TWC.

RiCL Research in Corpus Linguistics

A corpus-based analysis of language ideologies in Hungarian school metalanguage

Tamás Péter Szabó[1]
Research Institute for Linguistics, Hungarian Academy of Sciences / Hungary
University of Jyväskylä / Finland

Abstract – The main goal of this paper is to present a recently built interview corpus called *Corpus of Hungarian School Metalanguage – Interview Corpus* (*CHSM-IC*) and its potential in language ideology studies. This corpus was compiled during a broad survey on Hungarian school metalanguage carried out in 2009 and was recently made available for a wider research community within the CESAR (Central and South-East European Resources) project.
The study investigates interactional routines used in metadiscourses on language use. Printed texts cited from prestigious handbooks and interview data from *CHSM-IC* are compared. Thus, widely used, culturally-inherited text fragments are detected and confronted with the interviewees' narratives on their own communicational experiences. A case study on the discourse marker *hát* ('so', 'well') illustrates that there is a conflict and often a controversy between language ideologies disseminated by the Hungarian school system and the linguistic self-representation in the interviewees' narratives.
Combining Language Ideology, Conversation Analysis, Discourse Analysis and Discursive Social Psychology frameworks, the paper presents a detailed description on the emergence of metadiscourses in a school setting. The paper concludes that metalinguistic utterances (e.g., answers on grammaticality, statements on linguistic accuracy, etc.) and observable, spontaneous (or semi-spontaneous) language use patterns are regularly not in accordance with each other.

Keywords – discourse marker, L1 education, language ideologies, standardist language culture

1. Introduction

The present paper summarizes the results of a survey based on a communication-oriented approach of metalanguage. The analysis of interview data taken from a recently built corpus describes the dynamics and interactional structure of school metalanguage, illustrating how language ideologies emerge in metadiscourses. The presented theoretical background and methodology can be widely applied in the analysis of educational communication.

The importance of a corpus-based investigation of Hungarian school metalanguage becomes clear considering the standardist nature of Hungarian language culture (cf. Milroy 2001). In Hungary, curricula used in formal education contain prescriptivist and descriptivist elements. This heterogeneous and often controversial design is broadly criticized

[1] I would like to express my gratitude to my anonymous reviewers, to Petteri Laihonen and Alejandro Alcaraz Sintes for their careful reading and useful comments.

by sociolinguists, such as Kontra (2006). As part of this critical discourse, extensive research on school metalanguage has been conducted (for references, see, among others, Csernicskó and Kontra 2008), but interview analysis remained marginal. For such an analysis, interview corpora are needed and the present corpus aims to fill this gap.

Hungarian attitude and ideology research tradition is basically normative. Papers often conclude that attitude A is false and to be avoided, while attitude B is to be disseminated in education. In contrast, the goal of the present survey is not to evaluate ideologies and attitudes, but to investigate them as they emerge in discourses. For the construction and the analysis of the corpus, the methods of Conversation Analysis (CA) and Discourse Analysis (DA) have been followed.

In the second part of this paper, a case study on discourse marker *hát* ('so', 'well') illuminates the benefits of CA and DA approaches. The frequent use of *hát* in a turn-initial position is mentioned as an example of erroneous talk in research interview data, even by teachers and their students. This conception belongs to the prescriptivist tradition. At the same time, *hát* is present in almost all of the interviewee's verbal production in a turn-initial position, regardless of whether a given speaker stigmatizes *hát* or not. That is, the usage of *hát* does not predict its legitimization at an ideological level, nor can the negative evaluation of *hát* foretell its absence in speakers' utterances. The analysis of corpus data suggests that ideologies describing school-related expectations of linguistic accuracy, and observable patterns of spontaneous (or semi-spontaneous) performance are regularly not in accordance with each other.

2. On metalanguage

2.1. Approaches to metalanguage

According to Van Leeuwen (2004), at least two groups of definitions of metalanguage can be identified. Definitions from the first group claim that metalanguage is a specific register of language use with a scientific nature. Thus, metalanguage is described as a representation of cognitive representations, a tool for making theories, or telementing[2] inner beliefs or views. In the second group, communication-oriented definitions like Laihonen's can be found. Laihonen (2008: 669) argues that "[f]rom an interactional point of view, talk about language is a part of conversational action, such as answering, defending, blaming, accusing and apologizing". Following this approach, the present paper counts metalanguage as a part of ordinary communication and uses the methods of CA and DA for a dynamic description of metadiscourse patterns. Such analyses are important, because language ideologies emerge during metadiscourses.

Language ideologies can be briefly defined as statements on language use (for a detailed overview of the notion, see Laihonen 2009). These are indispensable elements of a community's language culture and, thus, are part of the linguistic socialization of a speaker acting in the given community.

Metalinguistic utterances in everyday communication routine have their evident antecedents in a huge tradition of metadiscourses. This tradition offers questions and often ready-made answers reflecting everyday language use. Socialization in metalanguage is the exploration of this tradition: speakers are exposed to a high amount of metalanguage (see Berko Gleason 1992) and simultaneously they learn various ways of formulating statements on language. One specific scene of this socialization is formal education.

In schools, several methods are used in order to shape students' metalinguistic performance. Some ideologies are rejected, while others are considered desirable by teachers, who are very influential actors in a formal educational setting. They are positioned as more competent speakers (role models), primary knowers and discourse managers (Lee 2007). Students meet a variety of metalinguistic practice during their training years and they learn how to construct language ideologies by explicit linguistic explanations, evaluations or other- and self-repair (repair often being an implicit way of building ideologies; cf. Laihonen 2008, 2009[3]). Students can assimilate to or differ from practices used in schools in their own communication.

Following a communication-oriented definition of metalanguage, some notes are necessary for the aims of metalanguage studies. Research on attitudes or ideologies is often legitimized by its presupposed usability in changing social structures and maladaptive behavioral patterns (among others, see Bohner 2001). For example, a study on school metalanguage might be considered capable of changing some false practices by teachers. But this argumentation is plausible only if one assumes that the change of attitudes or ideologies can change behavior. However, such a hypothesis involves the oversimplification of the results of cognitive psychology. In this oversimplified approach, reality is the input of cognition, while action is the output: one can recognize something in his or her environment, and then this recognition can be the basis of his or her decisions on behaving in a certain way. Mainstream cognitive

---

[2] The ideology of telememtation is described by Coulter (2005) as a conception which emphasizes the primacy of thoughts. From this point of view, thoughts and feelings readily exist before speaking, and speaking just gives form to them.

[3] Following Schegloff, Jefferson and Sacks (1977: 363), I will prefer the more general term 'repair' rather than 'correction' in data analysis, but when citing interview excerpts, I will use 'correction' because the words *javít*, *javítás* ('to correct', 'correction') were used in the transcribed texts.

psychology rejects this approach, evaluating it as a folk theory based on the Carthesian tradition (see Győri 2008). Győri (2008) argues that awareness is not an antecedent, but an outcome of behavior.

While folk cognitivist tradition explains behavior presupposing the following sequence 'cerebral activity → behavior → awareness', Discursive Social Psychology (DSP) does not investigate cerebral activity or mental processes (Potter and Edwards 2001, 2003). According to DSP, one learns to do something somehow, and then learns ideologies explaining, legitimizing or illegitimizing that behavior. These ideologies make behavior meaningful, but they do not govern acting. In a DSP description, ideologies are dynamic in nature and are never stable or finished: they are always maintained by people who construct, reconstruct or deconstruct them.

2.2. The role of metalanguage in a standardist community

Hungarian is a standard language culture (see Milroy 2001; Kontra 2006; Sándor 2006). Among others, it means that the cult of a privileged dialect called *standard* is observable. Efforts are continuously made to generate an idealized use of language. Besides, standardist movements aim for a uniform metalanguage as well. To reach this goal, various tools, such as dictionaries, prescriptivist handbooks of good usage and language etiquette (e.g., Grétsy and Kovalovszky 1980,1985) are widely available.

Standardization is a never-ending process, because language is basically heterogeneous. That is why a homogeneous language use, which would form the basis of a standard, cannot exist nor be practised (Milroy 2001). Nevertheless, several communicative habits show that the so-called standard presents an ideal for the majority of Hungarian speakers, and the school system plays a central role in the construction of this position: in formal education curricula, prescriptivist ideologies are disseminated, and language awareness activities are cultivated.

The above-mentioned dominance of standardist ideologies does not mean that there is a total absence of contestants. Descriptivist and prescriptivist ideologies are taught simultaneously, in different contexts. For example, while topicalizing the variety and richness of Hungarian, several vernacular inflections and phonemes are discussed and highlighted. However, these very same features are often evaluated as erroneous and are the object of other-repair, even in Hungarian language and literature lessons. This dynamism shows that language ideologies are embedded in various situations, and that they are used for very different purposes. When the curricular goal is to disseminate information on language varieties in order to develop tolerance, the evaluation of vernacular (dialectal) features are positive, but basically different ideologies are constructed, often implicitly, to legitimize the continuous regulation of classroom discourse, on the grounds that dialectal forms are inadequate in a formal education context.

This duality uncovers a feature of standardist cultures: linguicism, which can be overt or hidden. The notion of linguicism has been defined by Skutnabb-Kangas and Phillipson (1989: 455) as "ideologies and structures which are used to legitimate, effectuate, and reproduce unequal division of power and resources (both material and non-material) between groups which are defined on the basis of language".

Linguicist practices need legitimizing ideologies and these ideologies are formulated through metalanguage. As Milroy (1998: 64-65) summarized, "[i]n an age when discrimination in terms of race, colour, religion or gender is not publicly acceptable, the last bastion of overt social discrimination will continue to be a person's use of language".

Linguistic evaluations and repair may thus serve linguicism, and the Hungarian school system, as an influential institution, routinely disseminates standardist ideologies and repair techniques. Students learn how to reproduce overt defensive statements against another person or against themselves (e.g., other- and self-repair, public negative evaluation of others' speech or of one's own speech), and, what is more, they simultaneously learn practices to hide the nature of this language culture. In other words, students might in practice learn, for example, how to repair dialectal forms in interaction, even though learning to appreciate dialects in principle. The positive evaluation of dialects as archaic and regionally/temporally valuable varieties seems to be a kind of tolerance, but restricting the usage of dialects to a really small segment of communication is a type of linguicism. This practice limits the chances of dialect users to use their vernacular without being stigmatized.

3. THE *CORPUS OF HUNGARIAN SCHOOL METALANGUAGE* (*CHSM*)

3.1. A need for a corpus-based investigation

A systematic description and analysis of implicit and explicit language ideologies in education is needed to achieve an understanding of Hungarian metalinguistic socialization in schools. For this description to be empirically based, the project requires a corpus. The available significant interview corpora of contemporary spoken Hungarian were compiled for other purposes by the researchers at the Research Institute for Linguistics. *Budapest Sociolinguistics Interview* (BSI-2; cf. Kontra and Váradi 1997) was a complex survey on the attitudes and speech performance of the habitants of Budapest, carried out between 1987 and 1989. Another larger project at the same institution is *BEA – A*

Hungarian Spontaneous Speech Database. Its data collection began in 2007. However, these materials do not include recordings from educational settings, therefore a new corpus was needed.

3.2. Data collection

Building the *CHSM* was part of the PhD project of the present author. This project targeted a complex, multiperspectivist investigation of school metalanguage. Thus, data collection followed three different methods in Hungarian elementary schools, vocational high schools and grammar schools, in grades 1-4, 7 and 11. In the Hungarian educational system, children of the same age can learn in different types of schools, so that even in a grammar school very young children can be found in grades 1-4 (aged 6-11), older children in grade 7 (aged 13-14) and in grade 11 (aged 17-19). The sampling does not represent the demographic average of Hungary, but fits the requirements of qualitative representativity (Sántha 2006): it is rather heterogeneous and represents both typical and extreme cases.

3.2.1. Questionnaires

A questionnaire was used to gather background information on the students, concerning (1) the consumption of various cultural goods and language practices in different genres ("How often do you read a book/watch television/write an e-mail/write an official letter/read anything in foreign languages?", etc.); (2) consumption of standardist culture goods, such as spelling dictionaries, comprehensive dictionaries or television programmes/websites on linguistic accuracy; (3) demography characteristics (age, sex, etc.); and (4) improvement of students at school.

Discussions were initiated on the evaluation of different speech varieties, e.g., dialects, slang and others. Other- and self-repair was recurrently topicalized (e.g., "Do you correct others' language use? Are you corrected yourself? How? How often?", etc.).

Some questions generated lively debates in the class community. For example, the students were asked to compare two imaginary young girls who were described by their language use. One of the students was the speaker of an unidentified dialect, and the other was a so-called standard-language speaker. The goal of this task was to initiate a discussion on the possibilities, relevance and legitimacy or illegitimacy of evaluating and classifying others by their use of language. Being an exciting issue, this topic served as an ideal introduction in the interviews.

1,195 students filled in the questionnaires in grades 7 and 11. During the quantitative analysis of the questionnaire data, question–response sequences were taken as turns in a mediated discourse between the researcher and the informants, so questionnaires can also serve the goals of a detailed interactional analysis of Hungarian school metalanguage.

3.2.2. Notes on classroom observations

Notes on classroom observations (grades 7 and 11, N=62 lessons) focused on the organization of a lesson, and on the patterns of teacher–student communication. Emphasis was on interactional routines for regimenting classroom discourse. Scenes from the observed lessons, which were considered as interesting from the interviewer's angle, were mentioned in the interviews in order to initiate conversation on the given problem (e.g., "Why does your teacher punish slang speakers?"; "What do you think about the student's response to the teacher in this case?", etc.).

The corpus contains ca. 29,000 words and it is stored in XML format, ready to be analyzed with corpus linguistics tools, such as CLaRK (Simov et al. 2001).

3.2.3. Interviews: CHSM-IC

Semi-structured research interviews were made with students and their teacher of 'Hungarian Language and Literature'. Students were interviewed by the present author, generally in the company of one or two of their peers. Regularly, students reacted on the statements of their peers, and their co-constructive routines, having a high importance in the emergence of language ideologies, were observable. Being the largest subcorpus of *CHSM*, the *Interview Corpus* (*IC*) was published under the name of *CHSM-IC* (*Corpus of Hungarian School Metalangiage – Interview Corpus*) in 2013, as part of a European Union-funded international project called *CESAR* (*Central and South-East European Resources*). The publication process was supervised by Tamás Váradi, chair of the CESAR project. An online registration, a research plan and a declaration on ethical issues is required from future users.

Some of the basic topics of the interviews were 'stereotypes', 'language rules', 'other-repair', and the linguistic evaluation of certain language varieties such as dialects (unidentified), slang, obscenity and impediment in speech. Tasks differed according to the topic investigated. Students had to read a text initiating the topic and then answer a few questions. For instance, in the case of stereotypes, the questions were the following: "If you meet somebody speaking a rural dialect / slang / profane words / in a tongue-tied way, what do you think of him or her? Would you like to be his or her friend? Do you like the way he or she speaks? Would you evaluate or correct explicitly his or her language use?"

Or, in the case of rules in general: “What do you think is a rule? Do you know optional rules? Do you know obligatory rules? Give examples!”. Students were invited to deliver narratives on language use and also to explain the situation described.

Teachers answered questions on the evaluation of their students’ linguistic performance, textbooks and other materials used during classroom activities, and on other methodological and pedagogical issues. These interviews could be used as background material for the analysis of classroom discourse or student interviews.

From the viewpoint of language ideology research, the most important subcorpus of *CHSM* is the *IC*. Since it is an annotated transcription of spoken metalanguage, it can be used for various purposes. Data in the *IC* can be searched along two types of annotations:

(1) Thematic annotations marking the topic of the conversation. These are used primarily when seeking for ideologies emerging on a certain question, e.g., ideologies on other-repair, or, in the case of teachers, ideologies on teaching principles and methods, or the evaluation of children’s knowledge, etc. Annotation is multilevel: within categories, certain subcategories can be found (e.g., other-repair, either initiated by the interviewee or initiated by a communication partner, etc.).

(2) Annotation of the characteristics of spoken language. This can be useful when analyzing the interactional features of the emergence of an ideology. Comparative studies can also be made, e.g., what interviewees say about repair and how they actually do repair during the recorded conversation. For a detailed presentation of interview structure, see Appendix 1 and Appendix 2.

Table 1 shows the number of interviewees. Interviewee selection targeted a comparative study on metalinguistic performance at different levels of formal schooling. As already mentioned, students were interviewed in small groups of two or three, and for this reason the number of interviewees is much higher than that of the interviews (cf. Table 2). For the selection of interviewees, the so-called *snowball* technique was used (interviewees suggested other potential interviewees), so that the size of the subsamples is not equal.

The interview corpus contains ca. 47.7 hours of recorded speech (346,500 words; see Table 3). Transcription is stored in XML format, ready to be analyzed with corpus linguistics tools such as CLaRK (Simov et al. 2001). Transcript annotation fits the standards published in TEI guidelines (Burnard and Bauman 2012). Personal data such as names, location of the interview, address or residential city of the mentioned persons, etc. are masked. At the present state, the audio files recorded by a digital device have not yet been annotated.

Interview collection of children aged 6 to 11 (grades 1-4) was not transcribed word by word in full length, because discourse topic often deviated from the interview outline to small talk. These secondary topics were summarized in brief notes.

Groups and classes	Capital	City	Village	Σ
Students, grades 1–4.	28	13	6	47
Students, grade 7	26	11	2	39
Students, grade 11	13	10	–	23
Teachers, grade 7	7	3	1	11
Teachers, grade 11	8	5	–	13
Σ	**82**	**42**	**9**	**133**

Table 1. Number of interviewees

Groups and classes	Capital	City	Village	Σ
Students, grades 1–4.	15	5	3	23
Students, grade 7	11	4	1	16
Students, grade 11	7	5	–	12
Teachers, grade 7	7	3	1	11
Teachers, grade 11	8	5	–	13
Σ	**48**	**22**	**5**	**75**

Table 2. Number of interviews[4]

[4] A teacher working in the capital city speaks about his experience with grade 7 and 11 students in the interview. This implies that this interview is counted twice.

Groups and classes	Capital	City	Village	Σ
Students, grades 1–4.	492	149	87	728
Students, grade 7	450	173	31	654
Students, grade 11	348	221	–	569
Teachers, grade 7	265	106	21	392
Teachers, grade 11	316	204	–	520
Σ	**1871**	**853**	**139**	**2863**

Table 3. Length of interviews (minutes)

Interviews were not based on elicitation but had a conversational design. Both the interviewer (the present author) and the interviewees participated intensively, as Table 4 shows. In grade 7, the interviewer often ended up explaining tasks in a more detailed manner than in grade 11. That is the main reason why the proportion of the interviewer's speech is higher in grade 7. Speaking with teachers, the interviewer was not so dominant, at least according to quantitative indicators. As trained and experienced professionals, teachers delivered detailed narratives and explicit evaluative comments, and they often kept the floor for several minutes.

Groups and classes	Number of words in subcorpora	Proportion of words uttered by the researcher
Students, grades 1–4.	46 236	—[3]
Students, grade 7	91 185	42%
Students, grade 11	82 054	31%
Teachers, grade 7	55 350	15%
Teachers, grade 11	71 675	15%
Total:	**346 500**	**Average[4]: 28%**

Table 4. Number of tokens in the interview collection[5,6]

The proportion of the researcher's words is not higher than in other Hungarian corpora of semi-structured interviews. In the second version of the *Budapest Sociolinguistics Interview*, 35% of the transcribed corpus is the speech of the interviewers (cf. Borbély and Vargha 2010), while in a segment of *BEA*, this proportion is 23% (cf. Bata and Grászi 2009).

4. A CASE STUDY: DISCOURSE MARKER *HÁT* ('SO', 'WELL') IN A TURN-INITIAL POSITION IN SPOKEN HUNGARIAN

4.1. Traditions in grammar studies: from prescriptivism through controversies to descriptivism

Any kind of linguistic description is a language ideology, but grammars have special importance. Metatexts, published in an academic context, are prestigious sources of language ideology construction: they can serve as a basis of argumentation in metadiscourses. Therefore, it is important to investigate characteristic approaches to *hát* through the history of Hungarian grammars before analyzing the data recorded in *CHSM-IC*.

Hát ('so' or 'well' in English) has important interactional functions in turn-taking. In the mainstream Hungarian literature of *hát,* a duality of prescriptivism and descriptivism can be found. Authors from both streams claim that *hát* is used mainly for marking the intention of speaking (the speaker's wish to participate in a conversation), but they differ in the evaluation of the item.

---

[5] Since not the whole material was transcribed word by word, the last column does not show the proportion of utterances for students, grades 1-4.
[6] The total average is for students (grades 7 and 11) and teachers (grades 7 and 11).

A very typical example of the prescriptivist approach comes from *Nyelvművelő kézikönyv* ('Handbook of Language Cultivation'; Grétsy and Kovalovszky 1980, 1985). The two thick volumes of this prestigious handbook can be found in almost every school library in Hungary, and its content was widely popularized through grammar textbooks or various journal and newspaper articles. The viewpoint of this handbook is characteristic to the normative approach which was dominant in Hungary for the second part of the 20th century. (Most of the manuscript was written in the 1960s, but its publication and editing lasted until the 1980s.) The excerpt cited below can be found in the manual's entry for *beszédtöltelékek* or 'filler words':[7]

> Megfigyelhetjük, hogy sokan, főképp az élőbeszédben, társalgásban s különösen értekezleteken, vitában, felszólaláskor, úgy szerkesztik mondataikat, hogy **teletűzdelik** őket **tartalom és hasznos nyelvi funkció nélküli,** ill. funkciójukat vesztett fölösleges elemekkel: töltelékszavakkal, szókapcsolatokkal, mondattöredékekkel. Ezek többnyire **csupán** arra valók, hogy a beszélő időt nyerjen mondanivalójának megfogalmazására, megtartsa beszédének **(látszólagos)** folyamatosságát, ill. megakadályozza, hogy a beszélgető társ elvegye tőle a szót. **Nemegyszer azonban a gondolatok kialakulatlanságából, zavarosságából, esetleg hiányából ered a használatuk**.
>
> We can observe that many people, especially in spontaneous speech, conversations and particularly at meetings, debates or speeches, lard their sentences with elements **without content and useful linguistic function**, or with unnecessary elements of lost function. These are filler words, phrases and fragments of sentences. Generally, these are used **just** to gain time for the speaker for constructing the message and to maintain the **(illusionary)** continuity of the speech and to arrest his or her communication partner in continuing. **But, frequently, the motivation of their use is the primitivity, confusion or lack of thoughts.**
>
> (Grétsy and Kovalovszky 1980: 323; emphasis added)

There is no evidence that this text directly shaped metadiscourses, but it appeared in a prestigious handbook, published by the Hungarian Academy of Sciences and edited by reputed linguists, so that it could indeed have an impact. Appearing in a synthetizing handbook, this text is not without antecedents: it summarizes a folk tradition. The text describes *hát* as an element which is insignificant and lacks a function, but which can disturb communication processes (e.g., the flow of discourse). However, the passage is contradictory: some functions are mentioned ("gain time for the speaker for constructing the message and to maintain the (illusionary) continuity of the speaker's speech"), but they are not evaluated as "useful".

As predecessors of present-day corpus-based linguistic analyses, studies on dialectal texts have had a long tradition in Hungarian linguistics. Based on tape recordings, grammars of regional varieties were elaborated. The author of such a regional grammar, Ittzés (1981) examined the case of *hát*. She writes about the rich functionality of *hát* and notes that most often it occurs in a turn-initial position, but evaluates it as a filler word.

Using another corpus of recorded and transcribed speech gathered at Eötvös Loránd University, Keszler (1983) notes that every linguistic element has a function, so a zero function (or functionless) can hardly be thought to exist in language use. She adds that the term *töltelékszó* 'filler word' had not been defined in detail in previous Hungarian studies. Keszler's notes foreshadow the basic notions of purely descriptivist sociolinguistics, but surprisingly, as a contradiction, the last note in her paper—and the conclusion of her later account (Keszler 1985) – is that *hát* is often a "totally functionless (...) unnecessary (...) filler word" (Keszler 1983: 178).

In recent papers, a zero function is mentioned as nonsense in linguistic description. Using a heterogeneous terminology, several authors – mainly from the framework of speech act pragmatics – conclude that *hát* has an important and multiple function in the organization of spontaneous speech as discourse marker, mainly in the signaling of turn-taking. Although this approach can now be considered mainstream in the community of Hungarian linguists (cf. Dér 2010; Dér and Markó 2007; Schirm 2011), the case is not the same in public formal education, as interview data will show in the next section.

4.2. A corpus-based investigation of explicit language ideologies on *hát*

In order to investigate language ideologies emerging in research interviews, we can observe explicit evaluations and narratives on communication practice and then compare the findings to the performance of the interviewees. Thus, explicit and implicit ideologies can be compared, because the usage of *hát* can be counted as an implicit ideology claiming that *hát* performs different functions.

The topic of *hát* arose in various contexts during the research interviews, but most often it was framed by narratives on other-repair. As a part of the interview protocol, the following questions concerned this topic: "Have you been corrected for your language use? How was it done? How was it reasoned? What was the problem with the corrected

---

[7] In the indented examples, Hungarian originals come first, and then my own translations follow.

word/expression?" These questions were targeted to all of the participants in the interview, which is why it was possible to observe the reflections between participants. These cases show the dynamics and interactive characteristics of ideology construction very well. The following excerpt illustrates a widely occurring ideology:[8]

701: Osztályfőnök azt mondta, hogy »*hát*tal nem kezdünk mondatot.«

701: Our teacher said "we don't start a sentence by *hát*".

(Csongrád county, city, grammar school, grade 7, female)

In this case, a prescriptive statement on language practice was quoted by the interviewee, and the source (the teacher) was identified. Quoting is a very common form of making ideologies on language use: by mentioning an authority as a source, the relevance and value of the given statement can be increased. Quoting plays an important role in metalinguistic socialization: it marks the first step in the acquisition and internalization of a given metalinguistic procedure, e.g., the construction of an ideology (cf. Aro 2009, 2012).

In the previous excerpt, a direct quotation is formulated as a statement on a "we-group". This group avoids using *hát.* This common identity can be characterized through other narratives on "how we use language" as well. "We-groups" are often in an opposition to "they-groups": "their" language use differs from "ours". This dynamics is crucial in a standardist community as a basis for establishing distinctions between groups.

The following excerpts illustrate that even in the first stage of formal schooling, developed practices of language ideology construction are learnt. The interviewer had a discussion with two girls in grade 2 in a Baranya county elementary school. The evaluations used by teachers during repair were discussed:

IV: [Mit mondtak a tanárok?]
451: == Hát azt mondja, ne kezdd hogy »ne kezdd úgy mindig a mondatunkat [!], hogy *hát*« ==
452: == mondja, hogy (2 mp) »ne kezdd úgy (5 mp) *hát* meg *öööö*« ==
451: hogy meg azt is mondják, hogy »nem így kezdjük a mondatokat« .
IV: Ühm. És mit gondoltok, hogy miért nem úgy kezdjük a mondatokat? (2 mp) Azt azt nem mondják el, hogy miért?
451: Nem.
452: Nem szokták.
IV: Aha. És ö ti mit gondoltok? Van ezzel kapcsolatba valami ötletetek, hogy mi lehet ennek az oka? (8 mp – suttognak egymás között)
452: Van.

IR: [What did teachers say?]
451: == **Hát** she says, dont't start that »don't start our sentences in a way that *"hát..."*« ==
452: == says that (2 secs) »don't start in a way (5 secs) *hát* and *öööö*« ==
451: and they say as well that »we don't start sentences this way«.
IR: Uh huh. And what do you think, why we don't start sentences that way? (2 secs) Don't they tell you why?
451: No.
452: They don't.
IR: Yup. And er you, what do you think? Do you have an idea on what could be the reason of that? (8 secs – whispering between each other)
452: We have.

(Baranya county, village, elementary school, grade 2, females)

This excerpt illustrates the process of common ideology construction. The girls in the excerpt cooperate in answering the interviewer's questions. They quote their teachers in a similar way and the girl numbered 452 often follows her classmate, reformulating her statements. It may also be observed that the interviewer adopts his speech in the use of personal pronouns to the teachers quoted by the girls, e.g., "Why *we* don't start sentences that way?" and then switches to the students' position, e.g., "And er *you*, what do you think?".

After a discussion between each other, the girls create an ideology on why *hát* should not be used. The interviewer continues ideology construction, initiating another perspective on erroneous talk, and asks the girls to make an "ordinary" utterance (in their terminology: *sentence*) "not ordinary". Girls solve the problem by inserting *hát* and hesitation marker *öööö* to positions where they are normally used in spoken Hungarian:

---

[8] In the excerpts, bold fonts signal that *hát* is used as a discourse marker in its primary function (many occurrences of *hát* display a secondary, metalinguistic function, as subjects of evaluations). *Öööö* is used for transcribing a common hesitation marker; it is not translated when used as a subject of evaluation. *IR* (in the Hungarian text: *IV*) stands for "interviewer", a three-digit number identifies interviewees. == marks overlaps. Between » and «, statements counted as quotations can be read. | | marks reiteration and correction.

IV: No, mi az oka? Mire jöttetek rá? (4 mp)
451: Hát az, hogy a *hát*tal meg az *öööv*el meg az ilyenekkel nagyon nem lehet szép, rendes mondatot == alkotni. ==
452: == Alkotni. ==
IV: Ühm. És milyen az a rendes mondat? (1 mp)
451: Mondjuk ha *én elmentem a fagyizóba, vettem egy csokis csokis fagyit.*
IV: Ühm. ez egy rendes mondat? És milyen lenne ez úgy, hogyha ez nem rendes mondat lenne? (4 mp) Hogyan mondanátok ezt nem rendesen?
451: Hogy == *hát én elmentem a fagyizóba ööö vettem egy* (4 mp) *egy gombóc ö fagyit.* ==
452: == *Én elmentem a | jé | fagyizóba ööö vettem egy* (2 mp) *gombóc* (4 mp) *fagylaltot fagylaltot* ==

IR: Well, what's the reason? What have you arrived at? (4 secs)
451: **Hát** that by *hát* and *ööö* nice, ordinary clauses are not really possible to == be created ==.
452: == Be created. ==
IR: Uh huh. And what does an ordinary clause look like? (1 sec)
451: Let's say, *én elmentem a fagyizóba, vettem egy csokis csokis fagyit* ['I went to the ice cream shop and I bought a scoop of chocolate ice cream']
IR: Yeah, is it an ordinary sentence? And what would it be like if it weren't an ordinary sentence? (4 secs) How would you say it in a not ordinary way?
451: That == *hát én elmentem a fagyizóba ööö vettem egy* (4 secs) *egy gombóc ö fagyit.* == ['I *hát* went to the ice cream shop *ööö* I bought an ice cream']
452: == *Én elmentem a | jé | fagyizóba ööö vettem egy* (2 secs) *gombóc* (4 secs) *fagylaltot fagylaltot* == ['I went to the ice cream shop *ööö* I bought a scoop of ice cream']

(Baranya county, village, elementary school, grade 2, females)

At this point, the interviewer changes to linguistic stereotypes concerning the usage of *hát* and *ööö*. The girls prepare the answer together for some 15 minutes and then construct an ideology: as *hát* users, they can be evaluated as bad communicators in a conversation:

IV: Ühm, tehát akkor ilyen lenne a nem rendes. És mit gondolsz, mondjuk hogyha így beszélgetnétek velem és és öö úgy beszélgetnétek velem, hogy mondanátok, hogy *hát elmentem és ööö vettem egy fagyit,* akkor én mit gondolnék rólatok vagy gondolnék-e valamit egyáltalán, hogy ti most ilyen nem rendes mondatban válaszoltatok?
451: Igen.
452: Igen.
IV: Mit gondolnék rólatok? (8 mp – suttognak) Hm? (7 mp – suttognak)
451: Hogy mi nem tudunk nagyon == beszélgetni az emberekkel == .
452: == Beszélgetni az emberekkel ==

IR: Uh huh, so this would be the not ordinary. And what do you think, let's say, if you would speak that way to me, and and er you would speak to me that way that you would say *hát elmentem és ööö vettem egy fagyit* ['*hát* I went and *ööö* I bought an ice cream'], then what should I think about you or should I think anything about you because you have answered in a not ordinary sentence?
451: Yes.
452: Yes.
IR: What should I think about you? (8 secs –girls are whispering) Huh? (7 secs – girls are whispering)
451: That we can't really == get a conversation with people ==.
452: == Get a conversation with people ==.

(Baranya county, village, elementary school, grade 2, females)

In the above cited language ideologies, a rather negative evaluation of *hát* can be found in explicit statements. Further investigation proves that similar ideologies were created at different levels of education as well: these can be seen as popular. For example, in another interview done to two girls in grade 7, the students claimed that *hát* marks uncertainty and lack of communication skills.

In another conversation, one of the interviewees claims that she does not like utterances starting by *hát* or *és* 'and'. Subsequently, assimilating to the ideology constructed by this student, the interviewer asks for a legitimization of the negative evaluation of these lexemes:

IV: [...] miből gondoljátok, hogy ö hogy rossz dolog *és*sel vagy *hát*tal kezdeni mondatot? (3 mp)
092: Hát ö igazából ezzel nem kezdünk mondatot.
091: *[nevet]*
IV: *[nevet]*
092: Már úgy kezdődik.

IR: [...] how do you know that it is a bad thing to start a sentence by *hát* or *és*? (3 secs)
092: **Hát** er actually we don't start a sentence by this.
091: *[laughs]*
IR: *[laughs]*
092: It is started that way, then.

(Budapest, elementary school, grade 7, females)

The student, by creating a "we-group", answers that utterances (in her terminology: *sentences*) are not started by *hát* (or *és*), but betrays her own tenets by starting her utterance precisely by *hát.* To this self-contradiction the other interviewee and the interviewer respond with laughter and then 092 herself reacts by laughing, too. This can signal that she behaved differently from the members of the previously constructed, ideal "we-group".

It is not by accident that students make negative evaluations on turn-initial *hát*. As a source of high formality, the teacher's statements contain a rejection of *hát* as well. In a narrative, constructed upon her own practice as a teacher, the interviewee cited below claims that students should not use *hát* in a classroom context. Her opinion is that students should know answers by heart and that is why *hát* is not acceptable:

671: A *hát,* minden mondatot háttal kezdünk, én is sokszor, (1 mp) ö ha szabadon beszélgetünk, persze nem olyan nagy baj. De amikor a diákokat kérdezem, elvileg már őneki meg kellett tanulnia azt a választ, tehát nem kezdi [nyújtva *hát*] ilyen időnyerő válasszal, de mindig *hát*tal kezdik, és akkor azt is [...] fölírom, hogy *hát,* és akkor áthúzom. *[nevet]*

671: *Hát,* we start every sentence by *hát,* including me, many times, (1 sec) er when we talk spontaneously, of course, it is not a big problem. But when I examine students, they have had to learn the answer – in principle – so s/he wouldn't start by *"hát…",* which is a tool for gaining time. But they always start by *hát* and then [...] I write *hát* up on the blackboard and then I score it out *[laughs]*

(Csongrád county, city, grammar school,
teacher of Hungarian language and literature in grade 11, female)

The ideology presented is educational for at least two reasons:

(1) The quoted teacher notes that students should know answers "by heart". This legitimization of school practices supports a traditional assimilation method (cf. Aro 2009, 2012) that gives only one main task for students: the reproduction of normative texts disseminated by the school.

(2) A narrative on a repair method can be observed as well. The teacher writes examples of erroneous talk on the blackboard and then she scores them out. By using this method, she enforces her verbal instructions with visual ones.

The following excerpt is a typical example of the dynamics of common ideology construction in a research interview context. Talking about phenomena evaluated as errors (turn-initial *hát* and *és* 'and', and definite article before proper names), the interviewer starts to investigate the reason of students' statements:

211: Erre nincs írott szabály,
IV: Ühm.
211: ezt mindenki tudja magától, hogy *[nevet* nem szabad.*]*
IV: Ühm. És akkor hogyha mindenki tudja magától, akkor valóban így is beszélnek? Tehát akkor nem is ö kezd senki *hát*tal mondatot?
211: De.
212: De.
IV: *[nevet* És mi*]* lehet ennek az oka, hogy hogy tudják, s akkor mégis használnak *hát*tal mondatot, vagy kezdenek *hát*tal mondatot vagy *és*sel, kirakják a név elő a névelőt?
211: Valaki kérdez valamit == ö a választ azt ==
212: == Igen, és ezzel időt nyerünk. ==
211: majdnem mindig *hát tudod, a nem tudom, mi, hát nyolckor, hát este,* hát majdnem mindenki így beszél szerintem.

211: There is no written rule for that,
IR: Yeah.
211: everybody knows by himself/herself that *[laughs* it mustn't be done.*]*
IR: Uh huh. And if everone knows it by himself/herself, do they speak this way in practice? So then nobody starts a sentence by *hát,* huh?
211: They do.
212: They do.
IR: *[laughs* And what*]* can be the reason of that? They know it and they still use sentences by *hát* or start sentences by *hát* or *és* and they use names with an article…
211: Somebody asks something == er the answer is ==
212: == Yes, and we gain time by this. ==
211: in almost every cases *hát tudod, a nem tudom, mi, hát nyolckor, hát este* ['*hát,* you know, I don't know what, *hát,* at eight, *hát,* at the evening'], **hát** almost everybody talks this way, I guess.

(Pest county, city, grammar school, grade 11, females)

Students claim that every speaker knows the rules governing the use of these words and that these words are incorrect. Following this line, the interviewer – in a naive manner – supposes that speakers who know these rules never use the mentioned words. Students make narratives on everyday communication to prove that the opposite is true. By doing so, they construct a linguistic description that is at odds with real life usage. This is a kind of implicit language ideology as well: prescriptivist metalinguistic tradition does not conform to everyday practice.

4.3. An analysis of *hát* in the speech of *CHSM-IC* interviewees

The data stored in *CHSM-IC* can be analyzed from a quantitative approach as well, which, like the qualitative one, also uncovers differences between (1) a school tradition of metatexts, (2) narratives on one's own communication practice and (3) the performance recorded in the interviews. From the interviews cited in section 4.2, cases were collected where *hát* was used in a turn-initial position in its primary function as a discourse marker, and not as the object of a linguistic evaluation. These data show how recorded performance differs from what is presented as an ideal. Table 5 shows the number of turns transcribed word by word from the selected interviewees' speech. This number indicates the size of the subcorpus of the given speaker. Another number shows the occurrence of *hát* in turn-initial position as a discourse marker. Speakers presented in Table 5 were all students, only participant 671 was a teacher. The data in Table 5 confirm Labov's claim that speakers, even teachers, may produce variants they evaluate negatively (Labov 1972; Nardy and Barbu 2006).

Interviewee's ID	**Number of turns transcribed word by word, uttered by the interviewee**	**Frequency of turn-initial *hát* (count)**
091	170	44
671	*141*	*31*
211	156	29
092	117	26
701	171	25
212	129	8
451	29	2
452	27	1

Table 5. They said "we don't use *hát*..."

A general percentual description of the use of *hát* in interviews is summarized in Table 6. The goal of this analysis is to illustrate the proportion of *hát* in the speech of the interviewees. The interviewer's utterances were filtered out. Definite and indefinite articles *(a*, *az* 'the' and *egy* 'a/an'), connectives such as *és* 'and', *hogy* 'that' and *is* 'too', and the negator *nem* 'no, not' were not considered. The analysis was made without lemmatizing subcorpora, since, in interactional studies, the form of a word is important from the point of view of agency analysis and routinized interactional patterns. Table 6 shows the proportion of the most common words occurring in *CHSM-IC* subcorpora.

Students				Teachers			
Rank	**Grades 1–4**	**Grade 7**	**Grade 11**	**Rank**	**Grade 7**	**Rank**	**Grade 11**
1	*hát* (2,71%)	*hát* (3,06%)	*hát* (2,01%)	1	*tehát* ('therefore', 'so', 1,78%)	1	*tehát* ('therefore', 'so', 1,65%)
2	*akkor* ('then', 2,46%)	*akkor* ('then', 1,92%)	*akkor* ('then', 1,59%)	2	*ez* ('this', 1,45%)	2	*akkor* ('then', 1,36%)
				4	*akkor* ('then', 1,33%)	4	*ez* ('this', 1,30%)
				10	*hát* (0,91%)	9	*hát* (0,99%)

Table 6. Interviewees' most frequently used tokens (percentage of the total number of tokens)

As may be seen in Table 6, *hát* is the most common token in each student subcorpus and is quite common in teachers' language use as well. It is noteworthy that *hát*, *akkor* 'then', *ez* 'this' and *tehát* 'therefore' all play a significant role in discourse organization and that *tehát* displays functions similar to those of *hát*.

5. Summary and future plans

As a part of a broad survey on metalanguage, I compiled the *CHSM*. This is a complex research tool which can be used for research on language ideology, for studies on spoken Hungarian or for educational purposes in general. The *CHSM-IC* lends itself to the analysis of the emergence of language ideologies in spoken discourses, with a CA approach (Laihonen 2008).

In the present paper, a case study on *hát* concluded that there exists no consistency between ideologies on language use and language use itself (cf. Krashen 1982). That is, language ideologies have no bearing on performance and *vice versa*: even teachers, who are perhaps the most prestigious authorities in prescriptive discourses, have used *hát* as a discourse marker regularly.

A collection of additional data from regions outside Hungary would be needed for a better understanding of Hungarian metalinguistic socialization in different cultural settings. To reach this goal, fieldwork in a bilingual context should be carried out. Another avenue for further research would be to study institutional multilingual policies in dual language schools. A third study, placing the present discussion into a wider cultural context, should deal with the linguistic landscape of educational spaces (e.g., pictures, cultural symbols, summaries of grammar instructions on the school walls) and its impact on the assimilation of linguistic evaluations (Brown 2012). The *CHSM-IC*, combined with additional material from the *BEA* and *BSI-2* corpora, can be the basis of a detailed CA description of discourse markers such as *hát*.

The applicability of the *CHSM-IC* is versatile, especially for educational purposes:

(1) A corpus-based analysis of interview discourse could be conducted while dealing with language-planning or sociolinguistics in the classroom. Excerpts from the *CHSM-IC* can serve as models for discussion on the topic and as experiments for teachers while planning their classroom activities.

(2) Alternatively, the *CHSM-IC* can help both teachers and students to observe spoken Hungarian. Tasks should be given to students, e.g., an analysis of an excerpt with special attention to sociolinguistic variables, such as status (student, teacher and researcher), age, gender, etc. In this case, the *CHSM-IC* can be used as a corpus of spoken language and its metalinguistic character would have a secondary importance.

(3) A systematic analysis of language ideologies emerging during interview discourses in the *CHSM-IC* can be used for decision-making in language policy and educational policy.

References

Corpora

BEA = Beszélt nyelvi adatbázis / A Hungarian Spontaneous Speech Database. Budapest: Research Institute for Linguistics of the Hungarian Academy of Sciences. Project chair: Mária Gósy. Available at <http://metashare.nytud.hu/repository/search/>

BSI-2 = Budapesti Szociolingvisztikai Interjú / Budapest Sociolinguistics Interview. Version 2. Budapest: Research Institute for Linguistics of the Hungarian Academy of Sciences. Project chairs: Miklós Kontra and Tamás Váradi. Available at <http://buszi.nytud.hu/>

CESAR = Central and South-East European Resources. Project chair: T. Váradi. Available at <http://cesar.nytud.hu/>

CHSM-IC = Magyar Iskolai Metanyelvi Korusz – Interjúkorpusz / Corpus of Hungarian School Metalanguage – Interview Corpus. Budapest: Research Institute for Linguistics of the Hungarian Academy of Sciences. Project chair: Tamás Péter Szabó. Available at <http://metashare.nytud.hu/ repository/search/>

Secondary sources

Aro, Mari. 2009. *Speakers and doers. Polyphony and agency in children's beliefs about language learning.* Jyväskylä: University of Jyväskylä.

Aro, Mari. 2012. Effects of authority: voicescapes in children's beliefs about the learning of English. *International Journal of Applied Linguistics* 22/3: 331–346.

Bata, Sarolta and Tekla Etelka Gráczi. 2009. A beszédpartner életkorának hatása a beszéd szupraszegmentális jellegzetességeire ['Impact of the communication partner's age on the suprasegmental features of speech']. In Borbála Keszler and Szilárd Tátrai (eds.), *Diskurzus a grammatikában – grammatika a diskurzusban*. Budapest: Tinta, 74–82.

Berko Gleason, Jean. 1992. *Language acquisition and socialization.* Boston: Boston University.

Bohner, Gerd. 2001. Attitudes and attitude change. In Miles Hewstone and Wolfgang Stroebe (eds.), *Introduction to social psychology*. Third edition. Oxford: Blackwell, 239–282.

Borbély, Anna and Andras Vargha. 2010. Az *l* variabilitása öt foglalkozási csoportban. Kutatások a Budapesti Szociolingvisztikai Interjú beszélt nyelvi korpuszban. ['Variability of *l* in five professional groups. Studies on the corpus of Budapest Sociolinguistics Interview']. *Magyar Nyelv* 106: 455–470.

Brown, Kara D. 2012. Minority languages in the linguistic landscape**,a**. In Heiko F. Marten, Durk Gorter and Luk van Mensel (eds.), *Linguistic landscapes and minority languages*. New York: Palgrave, 281–298.

Burnard, Lou and Syd Bauman (eds.). 2012. *TEI P5: guidelines for electronic text encoding and interchange by the TEI Consortium*. Charlottesville, VA: TEI Consortium.

Coulter, Jeff. 2005. Language without mind. In Hedwig te Molder and Jonathan Potter (eds.), *Conversation and cognition*. Cambridge: Cambridge University Press, 79–92.

Csernicskó, István and Miklós Kontra (eds.). 2008. *Az Üveghegyen innen. Anyanyelvváltozatok, identitás és magyar anyanyelvi nevelés.* ['Varieties of Hungarian, identity and Hungarian mother tongue education']. Ungvár–Beregszász: PoliPrint–II. Rákóczi Ferenc Kárpátaljai Magyar Főiskola.

Dér, Csilla Ilona. 2010. "Töltelékelem" vagy új nyelvi változó? A *hát, úgyhogy, így* és *ilyen* újabb funkciójáról a spontán beszédben. ['"Filler words" or new linguistic variables? Newer functions of *hát*, *úgyhogy*, *így* and *ilyen* in spontaneous speech']. *Beszédkutatás* 2010*:* 159–170.

Dér, Csilla Ilona and Alexandra Markó. 2007. A magyar diskurzusjelölők szupraszegmentális jelöltsége. ['Suprasegmental markedness of Hungarian discourse markers']. In Tamás Gecső and Csilla Sárdi (eds.), *Nyelvelmélet – nyelvhasználat*. Székesfehérvár–Budapest: Kodolányi János Főiskola–Tinta, 61–67.

Grétsy, László and Miklós Kovalovszky (eds.). 1980. *Nyelvművelő kézikönyv.* ['Handbook of language cultivation']. Vol. 1. Budapest: Akadémiai.

Grétsy, László and Miklós Kovalovszky (eds.). 1985. *Nyelvművelő kézikönyv.* ['Handbook of language cultivation']. Vol. 2. Budapest: Akadémiai.

Győri, Miklós. 2008. Tudatosság és megismerés. ['Consciousness and cognition']. In Valéria Csépe, Miklós Győri and Anett Ragó (eds.), *Általános pszichológia*. Vol. 3. Budapest: Osiris, 267–297.

Ittzés, Nóra. 1981. *Szövegszerkesztési kérdések Esztergom regionális köznyelvében.* ['Features of text construction in the regional standard of Esztergom']. MA thesis. Budapest: Eötvös Loránd University.

Keszler, Borbála. 1983. Kötetlen beszélgetések mondat- és szövegtani vizsgálata. ['Syntactic and discourse features in spontaneous speech']. In Endre Rácz and István Szathmári (eds.), *Tanulmányok a mai magyar nyelv szövegtana köréből*. Budapest: Tankönyvkiadó, 164–202.

Keszler, Borbála. 1985. Über die Verwendung der Füllwörten. *Annales Universitatis Scientarium Budapestinensis de Rolando Eötvös Nominatae. Sectio Linguistica* 15: 11–26.

Kontra, Miklós. 2006. Sustainable linguicism. In Frans Hinskens (ed.), *Language variation – European perspectives*. Amsterdam: Benjamins, 97–126.

Kontra, Miklós and Tamás Váradi. 1997. *The Budapest sociolinguistics interview: version 3.* Budapest: Research Institute for Linguistics of the Hungarian Academy of Sciences. <http://www.nytud.hu/buszi/ wp2/index.html> (12 June 2013).

Krashen, Stephen D. 1982 [2009]. *Principles and practice in second language acquisition.* Oxford: Pergamon Press. The 2009 online edition is the author's own: <http://www.sdkrashen.com/Principles_ and_Practice/Principles_and_Practice.pdf> (12 June 2013).

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania Press.
Laihonen, Petteri. 2008. Language ideologies in interviews: a conversation analysis approach. *Journal of Sociolinguistics* 12/5: 668–693.
Laihonen, Petteri. 2009. *Language ideologies in the Romanian Banat. Analysis of interviews and academic writings among the Hungarians and Germans*. Jyväskylä: University of Jyväskylä.
Lee, Yo-An. 2007. Third turn position in teacher talk: contingency and the work of teaching. *Journal of Pragmatics* 39: 180–206.
Milroy, James. 1998. Children can't speak or write properly anymore. In Laurie Bauer and Peter Trudgill (eds.), *Language myths*. London: Penguin, 58–65.
Milroy, James. 2001. Language ideologies and the consequences of standardization. *Journal of Sociolinguistics* 5/4: 530–555.
Nardy, Aurélie and Stéphanie Barbu. 2006. Production and judgment is childhood. The case of liaison in French. In Frans Hinskens (ed.), *Language variation – European perspectives*. Amsterdam: Benjamins, 143–152.
Potter, Jonathan and Derek Edwards. 2001. Discursive social psychology. In W. Peter Robinson and Howard Giles (eds.), *The new handbook of language and social psychology*. Chichester: Wiley and Sons, 103–118.
Potter, Jonathan and Derek Edwards. 2003. Sociolinguistics, cognitivism, and discursive psychology. *International Journal of English Studies* 1: 93–109.
Sándor, Klára. 2006. Nyelvtervezés, nyelvpolitika, nyelvművelés ['Language planning, language policy and language planning']. In Ferenc Kiefer (ed.), *Magyar nyelv*. Budapest: Akadémiai, 958–995.
Sántha, Klára. 2006. *Mintavétel a kvalitatív pedagógiai kutatásban* ['Sampling in qualitiative surveys on pedagogy']. Budapest: Gondolat.
Schegloff, Emanuel A., Gail Jefferson and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53/2: 361–382.
Schirm, Anita. 2011. *A diskurzusjelölők funkciói: a hát, az -e és a vajon elemek története és jelenkori szinkrón státusa alapján* ['The function of discourse markers: the history and present synchronic status of the Hungarian elements *hát*, *-e* and *vajon*']. PhD dissertation. Szeged: University of Szeged.
Simov, Kiril, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov and Atanas Kiryakov. 2001. CLaRK – An XML-based system for corpora development. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster: Lancaster University, 553–560.
Skutnabb-Kangas, Tove and Robert Phillipson. 1989. 'Mother tongue': the theoretical and sociopolitical construction of a concept. In Ulrich Ammon (ed.), *Status and function of languages and language varieties*. Berlin: Walter de Gruyter, 450–477.
Van Leeuwen, Theo. 2004. Metalanguage in social life. In Adam Jaworski, Nikolas Coupland and Dariusz Galasiński (eds.), *Metalanguage. Social and ideological perspectives*. Berlin: Mouton de Gruyter, 107–130.

APPENDIX 1. MAIN INTERVIEW TOPICS IN GRADES 1-4

- Do you like talking (with friends, family, classmates)? Do you talk a lot (with friends, family, classmates)?
- Is there any difference between talking at school/kindergarten and talking at home? Is there any difference between talking to children/adults?
- Is there anything you can do/say at home/with friends but not at school/kindergarten (and vice versa)? Have you ever been told you should not talk like that or should not say something?

APPENDIX 2. MAIN INTERVIEW TOPICS IN GRADES 7 AND 11

The main interview topics in grades 7 and 11 are presented in the following table, where (+/–) signals that the given topic emerged (+) or did not emerge (–) regularly in interviews made with students and teachers. Abbreviations were used during the XML annotation of the transcription.

Abbreviation	Focus	Students	Teachers
Metadisc	Metadiscourse: speech on the current interview discourse	+	+
CorrInstr	Narratives and ideologies on repair strategies (own/other) Self- and other-repair activity during the interview	+	+
Attitudes	Evaluation of other speakers' practice (dialects, slang, curse, impediment	+	–
Grammar	Narratives on grammar courses	+	–
Rules	Explicit description of different rules (not linguistic)	+	–
Attitudes	Evaluation of students' speech	–	+
TeachingExp	Teaching experience	–	+
TextBook	Evaluation of textbooks and other materials	–	+
Grammar	Evaluation of students' activity during grammar classes. Ideologies on the meaning and goals of grammar lessons. Own motivations.	–	+

RiCL Research in Corpus Linguistics

Phrasal Verbs in learner English: a semantic approach. A study based on a POS tagged spoken corpus of learner English

Joanna Wierszycka[1]
Adam Mickiewicz University / Poland

Abstract – Phrasal Verbs (PVs), understood as a verb and a particle, though very common in native speech, are reportedly difficult to learn by non-native speakers (NNSs) of English (see Celce-Murcia and Larsen-Freeman 1999). The hypothesis is therefore put forward that for Polish learners of English too the range of PVs is generally significantly smaller than for English native speakers (NSs) and that their degree of use of the semantic categories of PVs is inversely proportional to the PVs' level of idiomaticity. In other words, Polish learners have little trouble with transparent verbs, more with semi-transparent and most with opaque ones (see Dagut and Laufer 1985). In order to verify this hypothesis, we have used the evidence from the *PLINDSEI* corpus, that is, the Polish part of the *LINDSEI* (*Louvain International Database of Spoken English Interlanguage*), containing advanced English as spoken by NSs of Polish, and from the *LOCNEC* (*Louvain Corpus of Native English Conversation*), which we have used as a reference corpus. The comparison of PV usage by Poles as NNSs of English and by English NSs has been performed employing the scheme of contrastive interlanguage analysis (Granger 1996). We show learner over- and underuse of items and illustrate the searches conducted for identifying patterns of use. The methodology applied consists in a partially automatic extraction and a subsequent manual filtering of PVs from a POS-tagged NNS corpus and its reference NS corpus. A semantic analysis of the extracted PVs based on the notion of compositionality (see Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004) has been performed and the hypotheses verified.

Keywords – phrasal verbs, POS tagged corpus, semantic analysis, learner language, spoken learner corpus

1. INTRODUCTION

The aim of the paper is to verify the hypothesis that learners use idiomatically opaque Phrasal Verbs (PVs) less frequently than transparent items, assuming a linear scale of idiomaticity. The classification of PVs is performed by providing a semantic analysis of those items as used by Polish learners of English. The analysis has been done on a POS (part of speech)-tagged NNS corpus of oral English, *PLINDSEI* (that is, the Polish part of the *LINDSEI – Louvain*

[1] This project was funded by the National Science Centre, Poland (grant no. 3787/B/H03/2011/40). Special thanks are due to Paul Rayson for providing the tools and expertise needed to work with CLAWS4. I wish to thank Dr. Alejandro Alcaraz Sintes for his many suggestions and criticisms of an earlier draft of this paper.

International Database of Spoken English Interlanguage) and the *LOCNEC* (*Louvain Corpus of Native English Conversation*), which we have used as a reference corpus.

In Section 2 we provide an introduction to the grammatical annotation of the NNS or spoken learner corpus (Gilquin, De Cock and Granger 2010) that we have used for the purpose of the analysis. In the next section, we offer a more detailed analysis of the non-native use of PVs (Celce-Murcia and Larsen-Freeman 1999 and Armstrong 2004), as attested in the corpus data.

For the purpose of this paper, PVs are defined as a union of a lexical verb and a following particle[2] (see Quirk et al. 1972: 1150ff). PVs are to be distinguished from Prepositional Verbs (e.g., *call* [*on* NP] ('visit NP'), *cope* [*with* NP], *laugh* [*at* NP], *provide* NP [*with* NP]) and Phrasal-Prepositional Verbs (e.g., *face up* [*to* NP], *cut down* [*on* NP, e.g., *on expenses*], *fall back* [*on* NP, e.g., *on your wife's money*] (cf. Cappelle 2005). PVs as defined above are equivalent to, for example, Dehé's (2002) particle verbs or Pelli's (1976) verb-particle constructions. They "are sometimes [also] called two-word verbs because they usually consist of a verb plus a second word (...) a particle (...) to distinguish it from prepositions and other adverbs, although we acknowledge that (...) the same word can fit into more than one category" (Celce-Murcia and Larsen-Freeman 1999: 426). However, since the learner is the center of our research, the term most frequently used in pedagogical approaches and EFL course books, the label 'phrasal verbs', was decided upon.

The verb and the particle operate as a PV not only when they fulfill certain structural criteria, but also when they are found to function together semantically as a unit. The semantic unit features one of the following configurations (cf. Jackendoff 1997; Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004):

- Both the verb and the particle retain their literal meanings, the particle often indicating geographical direction, e.g., *come back.* These are directional PVs and are semantically transparent.
- The verb has a literal meaning and the particle provides an aspectual meaning, which is redundant, cf. Dehé (2002) and Hampe (2002) respectively, e.g., *read through.* These are aspectual PVs and are semantically semi-transparent.
- The verb and the particle have an idiomatic meaning, e.g., *come across.* These are idiomatic PVs and are semantically opaque.

Although very common in native speech (see Biber et al. 1999: 408), PVs are reportedly difficult to learn for NNSs of English (see Celce-Murcia and Larsen-Freeman 1999). However, Marks (2005: 12) points out that, against the common supposition of students and even teachers, the meaning of PVs is not illogical and random. What is more, the meaning can often be understood if learners recognize metaphorically extended meanings of particles and verbs. The underlying reason of the learners' difficulty with PVs might stem from the rule that regulates the native use of these items, namely, that the use of PVs by NSs is directly proportional to the PVs' level of idiomaticity.

Taking the above-mentioned factors into account, our hypotheses of PV use by Poles are the following:

- In the first place, the number of PVs used by learners tends to be significantly smaller than that of NSs.
- Secondly, PV use by learners is inversely proportional to the PVs' level of idiomaticity. In other words, Polish learners should have little trouble with semantically transparent verbs, more with aspectual verbs, and most with idiomatic ones. This would be contrary to how NSs use their language, as their use of idiomatic PVs grows together as their level of idiomaticity increases (see Celce-Murcia and Larsen-Freeman 1999).
- Finally, we suspect that the number of PVs used by Polish learners should vary according to their degree of exposure to the English language (number of years studying English in a classroom situation and number of months spent in an English-speaking country).[3]

All three hypotheses are summed up in Table 1.

---

[2] This particle is referred to as 'adverbial particle' in the CLAWS POS tagging system. For examples, see <http://ucrel.lancs.ac.uk/bnc2sampler/guide_c7.htm#m3prepadv-prep> (7 May 2012).

[3] The *LINDSEI* transcripts come along with learner metadata in the form of 'learner profiles' (see Gilquin et al. 2010). This has given us access to information on the time spent abroad and in the classroom by the Polish learners. However, it did not contain more specific information, such as, for example, the number of movies without voice-over watched by the learners.

HYPOTHESIS NO.	DEFINITION
H1	The number of PV tokens is significantly lower in NNSs than in NSs.
H2	The distribution of semantic categories of PVs in NNS is inversely proportional to the PV's level of idiomaticity as observed in NS use. (Polish EFL speakers underuse the idiomatic semantic category of PVs most).
H3	Differences in L2 exposure among NNS do matter. It is assumed that a longer exposure to the English language will produce more PVs in the learners' speech.

Table 1. Hypotheses pertaining to learners' PVs use

2. POS TAGGING OF THE *PLINDSEI* CORPUS: RESOURCES AND REASONS

In the case of this study, a part-of-speech (POS)-tagged corpus constituted an important, yet independent, part of the methodology. The two corpora that have been POS-tagged for the purpose of this study are *PLINDSEI* and *LOCNEC*. Both comprise subcorpora of the mother project, the *LINDSEI*. This is a corpus of advanced English learner speech produced by young adults with different mother languages (Gilquin 2012). *PLINDSEI* is the Polish IL[4] component of *LINDSEI*. Alongside the non-native varieties of English, a native English comparative corpus, the *LOCNEC* was compiled, in order to carry out contrastive interlanguage analyses both across various NNS categories, and between any of the NNS speech and the NS speech, since the *LOCNEC* was compiled following the same criteria as all *LINDSEI* subcorpora (see Figure 1). The *LINDSEI* language data came from informal interviews conducted in a question-answer format and constitutes continuous discourse. The target size of each subcorpus was aimed at approximately 100,000 words, but the subcorpora usually consist of a larger amount of data.[5] For the exact numbers, see Gilquin et al. (2010).

Although data annotation is the natural next step after data collection and transcription,[6] there were no available POS taggers trained on learner data when the *LINDSEI* data were annotated (academic year 2008–2009). It was therefore decided to test CLAWS (Garside 1995), a well-known tagger which had already been successfully trained on native spoken language on the *British National Corpus*.

For reasons of space there is no room to discuss at length the whole process of POS-tagging *PLINDSEI*. Still, it is important to stress that the main aim of tagging *PLINDSEI* was to find and pinpoint learner dysfluencies which caused erroneous POS tagging, in order to train the tagger on this difficult kind of input, and therefore improve its accuracy. We managed to achieve an overall tagger accuracy of 98.5%. Since it was the first attempt at POS-tagging a spoken learner corpus (see Mukherjee 2007; Aijmer 2009), it is hoped that other scholars engaged in this field of study will profit from our experience. What must be remembered is that the *PLINDSEI* corpus was POS-tagged with no particular grammatical or vocabulary item, such as, for example, PVs, in focus. This approach, that is, giving the same importance to all parts of speech, will greatly enhance further research in ways that the people responsible for tagging the corpus had not projected. In fact, as Leech (2004)[7] wrote, "no one in their right mind would offer to predict the future uses of a corpus".

At the same time it is vital to notice that there are pre-assigned grammatical categories, such as infinitives, adverbs or prepositions, which are essential to define the tags, as it is impossible to approach corpus tagging in a theory-free manner.

Once an effective POS tagging is done, it enhances grammatical searches enormously. General searches for PVs (i.e., any form of verb plus adverbial particle) would have been impossible without it. However, in searches for concrete examples of particles, or whole PVs, no POS tagging is needed. In the latter approach the researcher has a raw corpus at his or her disposal. Research on learner PV use so far has been based on raw corpora and, as a consequence, has been conducted on selected PVs only (e.g., Gilquin 2012). Therefore, the searches for PVs for the purpose of this study were done with the use of POS tags and with the help of *WordSmith Tools* software (Scott 2008).

3. METHODOLOGY

The investigation presented in this paper is based on the evidence of advanced spoken English by NSs of Polish, as described in section 1, and was performed using the scheme of contrastive interlanguage analysis (Granger 1996), the

---

[4] Granger's (1996) NL (native language) corresponds to our NS (native speaker), and her IL (interlanguage) corresponds to our NNS (non-native speaker).

[5] For example, the Polish part of *LINDSEI* has 114,862 words, while the French part has 143,044 words.

[6] However, see Sinclair (1991).

[7] <http://users.ox.ac.uk/~martinw/dlc/chapter2.htm> (7 December 2011).

Native Language (NL) vs. Interlanguage (IL) branch in particular (see Figure 1). Ellis (1994) stresses the importance of collecting comparable samples of learner language in interlanguage comparisons. However, the issue of comparability has not always been so obvious to scholars and, in fact, one of the reasons why many important aspects in applied linguistics have remained inconclusive is that researchers have not been "comparing like with like" (Granger 1996: 44). Granger also suggests deciding on genre sensitivity, rather than ease of access, when choosing the right reference corpus for our study. For this reason, the *PLINDSEI* and *LOCNEC* corpora were considered most appropriate for this study. Both were compiled according to the same criteria and can, therefore, be considered as comparable.

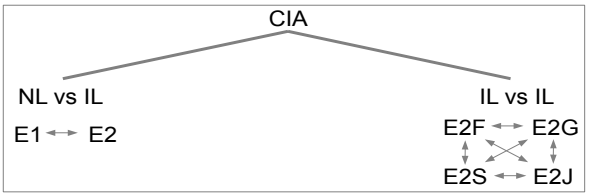


Figure 1. Contrastive interlanguage analysis (after Granger 1996: 44)

The methodology applied consisted in a partially automatic extraction and then manual filtering of PVs from the POS-tagged Polish spoken learner corpus of English and its reference NS corpus. The corpus data were used in their normalized frequencies of B-turns only. The rejected A-turns comprised the sections of the interviewers and they were of no interest for the research. Their only purpose was to keep the conversation going. B-turns, however, consisted of the interviewee part of the recordings, that is the students, selected along criteria of language and age level (for details see Gilquin et al. 2010). The overall number of words were 114,816 and 161,724 for the learner and the reference corpora, respectively. In order to arrive at a fully comparable corpus, a decision was taken to normalize the native corpus, which was done by eliminating 11 files randomly. Only the B turns analyzed which also limited the number of the number of words. The corpora produced 95,906 words in the learner corpus and 118,554 in the NS corpus.
words. Table 3 below provides the final word numbers after normalization in both corpora.

4. PHRASAL VERBS: WHAT, WHY, HOW?

It would be impossible to count all the PVs in the English language. One of the main reasons is that new ones are being constantly added, in some cases nouns or adjectives being the witnesses of the newest inventions, e.g., *I'm Christmased out* ('I'm sick of Christmas') in the 1996 movie *Elmo saves Christmas*, now widely used by the British and the American alike in the pre-Christmas period. It is an isolated example of a PV not to be found in dictionaries, but used by NSs. The number of such PVs to be "discovered" will remain unknown, and the number of coinages that will see the light of day will most probably remain a mystery. The other side of the coin is that some PVs are naturally getting out of use, and therefore dying out. There is also the factor of polysemous PVs (see, e.g., Table 5 below and the example of *pick up* in section 5.1). For all the above-mentioned reasons counting all PVs in the English language is not a straightforward and easy task. Still, there are researchers who have attempted to guess the exact number of PVs, which varies from as few as 700 (Bywater 1969: 97) up to 12,000 (Courtney 1983). It therefore seems that the attempt to count PVs is similar to that of calculating all vocabulary items in a given language.

For the purpose of this paper, PVs are selected in a two-stage procedure, after which their classification comprises another two stages. First, the basic structural definition is adapted from Quirk et al. (1972), where PVs are understood as constructions formed of lexical verbs followed by adverbial particles, e.g., *drink up.* They are to be distinguished from prepositional verbs, e.g., *dispose of*, phrasal-prepositional verbs, e.g., *get away with*, and other multi-word verbs. At this step PV candidates are automatically selected. What follows is a manual filtering stage of lexical verbs with particles from prepositional verbs and phrasal-prepositional verbs to arrive at a list of only true PVs ready for further semantic classification. The second step of the PV definition determines semantic classification and consists of two stages: classification of PVs according to the particle and further division of PVs into semantic categories along the lines of compositionality (adapted from Darwin and Gray 1999 and Armstrong 2004).

The rationale behind our research is to verify the existing research on the actual difficulty experienced by Polish speakers when learning PVs. The problem is that PVs are very common in native speech and therefore the difference between native and learner use strikes particularly hard (e.g., Celce-Murcia and Larsen-Freeman 1999; Darwin and Gray 1999; Armstrong 2004).

In terms of the procedure, PVs were extracted using the *WordSmith Tools* software, employing the search on tags. These were six different forms of a lexical verb (VV*), each followed by a particle as its context word (RP). All the C7 tags together with their explanations and examples are set together in Table 2 for clarity.

The outcome of this procedure was a list of potential PVs with all possible verb forms present in the corpus, e.g., while the particle *up* was searched for, together with the verb forms of *get*, the native corpus brought the following results: *gets up*, *getting up*, *get up* and any other string of signs that the data would show.

C7 TAG	DEFINITION + EXAMPLE
VV0	base form of lexical verb (e.g., *give*, *work*)
VVD	past tense of lexical verb (e.g., *gave*, *worked*)
VVG	*–ing* participle of lexical verb (e.g., *giving*, *working*)
VVI	infinitive (e.g., *to give... It will work...*)
VVN	past participle of lexical verb (e.g., *given*, *worked*)
VVZ	*–s* form of lexical verb (e.g., *gives*, *works*)
RP[8]	prep. adverb, particle (e.g., *about*, *in*)

Table 2. C7 tags used for searching PVs[9]

It should be noticed at this point that not only particles immediately following verbs were taken into consideration in the analysis, but also words with further search horizons, that is one or more words separating the verb from the particle, e.g., *<VVG> showing* ***<PPIO2>us*** *<RP> round <VVI>,* and *put* ***<AT> the <NN1> paper <.>.*** *<RP> back.* In this way, it may be assumed that all possible "verb plus particle" cases were retrieved and analyzed. Table 6 below presents a full list of the particles found in both the native and the non-native corpora.

Apart from the fully automatic extraction based only on the criteria of CLAWS-implemented POS grammatical categories, manual filtering was necessary, as the data were not free from noise. The first filter consisted of the application of the transitivity criterion (Armstrong 2004) to both the POS tagged Polish spoken learner corpus of English and its reference corpus.[10] Whether a PV is intransitive or not is a topic for separate discussion (see, e.g., Quirk et al. 1972).

The transitivity criterion serves to decide if a candidate for a PV (based on all the criteria described above) is actually a PV (Armstrong 2004). This, in turn, is based on the following premise: particles must be intransitive, i.e., they must form a unit with the verb, not with their object. Based on this principle, the PV candidates *set in* and *get up,* although seemingly perfect PVs, because they are made of a verb and a particle, were not considered to be true PVs, because the context of the PV candidate was taken into consideration. To be precise, the particle, which belongs to the object in *set in an aerobics class*, *getting up these mountains*, gives away lack of affinity with the PV category at the same time.

5. DISCUSSION OF FINDINGS

Having complied with all the above-mentioned criteria, the following results were reached when it concerns the overall number of PVs.

	NNS	NS
PVs	227	875
Corpus size B turns, (NS normalized)	95,906	95,862
Chi-square		384.28

Table 3. PV general token counts

[8] Bolinger (1971: 26, 28) uses the term 'prepositional adverb' to refer to particles which can be used both as adverbs and as prepositions, e.g., *in*, *up*. 'Prepositional adverb/particle' is CLAWS terminology. However, not only are all PV candidates extracted from the corpora with the use of POS tags at step 1, but they also go through manual filtering where prepositional adverbs are excluded and only true particles are left.

[9] For other POS tags, see CLAWS manual <http://ucrel.lancs.ac.uk/bnc2sampler/ guide_c7.htm#m3prepadv-prep> (7 May 2012).

[10] However, in order to have data tagged in a systematic way, none of them was manually annotated. The 1.5% error is small enough to leave the resource and not mingle with data manually, since introducing a human factor means initiating uncertainty as to the quality of the data which a computer corpus is not normally associated with.

	NNS	NS
PVs	85	274
Corpus size in types	4,917	5,606
Chi-square value		79.86
TTR[11]	37.44 %	31.31 %

Table 4. PV general type counts + TTR

In compliance with the first hypothesis, the learners' token number of PVs is indeed significantly smaller than that of NSs, as shown by the chi-square value (see Table 3). NNs appear to use almost four times as many tokens of PVs as learners, hence the high chi-square value. In order to be able to count the TTR, the types for both speaker groups are also presented in Table 4. From the tables above one may conclude that the learner's speeches seem to be lexically more varied than those of NSs. However, this may be due to the presence of hapaxes, which may not necessarily reflect learners' real competence of the vocabulary. The TTR values will be compared against values of semantic PV groups later in the paper.

5.1. The semantic approach to PV grouping: stage one (the particle)

PVs are semantic units. There is a bond between the verb and the particle but the degree of attachment varies. "In one case, the main factor determining the unity between the verb and the particle is semantic, mainly lexical, in the other, formal syntactic" (Sroka 1972: 180). Semantic categories comprise three groups: transparent, semi-transparent and opaque (cf. Jackendoff 1997; Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004). The first aim of the research was to check if the distribution of learners' semantic categories of PVs was inversely proportional to the level of PV idiomaticity as observed in NSs' use (e.g., Howarth 1998: 178). As the first stage of the semantic approach, PVs were grouped according to the particle (e.g., Rudzka-Ostyn 2003), bearing in mind that the particle carries more semantic load than the verb. Therefore, it was more logical to list PVs not according to the verb, but under the particle. Such a classification was carried out for both speaker groups separately and, as a result of that, interesting observations came into light (see further down).

Naturally, it frequently happened that within a given PV form, different meanings occurred. Although structurally identical, e.g., *pick up*, in *pick up a girl*, *pick up a job* and *pick up a tent* were classified as different types, because they come from three different semantic domains (see Bentivogli et al. 2004) and as such they could hardly be classified under one semantic term. All the aforementioned examples come from the *LINDSEI* corpus, but similar examples also occurred in the *LOCNEC* corpus, showing how the two speaker groups diverged not only in the number of verbs, but also in their sense distribution. Some of the differences in meaning expressed by the two speaker groups in a group of several PVs from the *LINDSEI* and *LOCNEC* corpora are shown in Table 5.

PV	NS SENSE	NNS SENSE	NS & NNS SENSE
take off	to achieve wide use or popularity	start	–
work out	to accomplish by work or effort		to prove successful, effective, or satisfactory
come over	approach		pass as somebody
go through		travel	experience, examine
come up			approach

Table 5. Polysemy across speaker groups: Common PVs with different senses

Subsequent analysis dividing the PVs into three semantic categories revealed that the polysemous PVs also crossed the idiomaticity line sometimes, not only across the speaker groups, as presented in Table 5, but also within one speaker group. Such a situation happened, for example, in the case of *come down*, which may be a literal, compositionally transparent PV, as in *prepare for the curtain to come down*, meaning 'to move downward', but which may also be a totally idiomatic, opaque PV, as in *she broke her hip and came down with cancer*, meaning 'to become sick with (an

[11] Type/Token Ratio (TTR): the number of types divided by the number of tokens. This indicates how rich or lexically varied the vocabulary in the text is. In the example of NNS, the TTR is 85 (types) ÷ 227 (tokens) x 100 = 37.44 %.

illness)'. All the form-sense differences do not mean that there are no common PVs between those two speaker groups, as may be seen in the fourth column of Table 5.

Another interesting observation pertaining to particle use by the two speaker groups is that, out of all of the particles present in the corpora, quite a few were absent from the learners' repertoire and one particle was used by the students only. The particle division into groups is shown in Table 6 below.

	PARTICLES
COMMON	*across*, *along*, *around*, *back*, *down*, *in*, *off*, *on*, *out*, *over*, *through*, *up*
SOLELY NS	*(a)round*, *after*, *away*, *by*, *for*, *to*
SOLELY NNS	*about*

Table 6. Particles in NS and NNS speakers

The division of PVs according to the particle was the first stage of the semantic analysis of PVs, because the basic meanings of the particles were to help out in deciphering the transparent and semi-transparent PVs. This, in turn, paved the way for the third category of PVs: opaque PVs, since the meaning of some of the opaque PVs is not fully opaque, but has its semantic roots in the meaning of the semi-transparent and transparent categories (see Rundell 2005).

Rudzka-Ostyn (2003), in her pedagogically-oriented book, employed the cognitive approach as a means for effective acquisition of PVs, and proposed 17 particle meanings, for which the leading meanings are presented below. Some of the particle senses are listed here in order to clarify further PV division.

- *on* stands for continuation of an action or situation, e.g., *walked on*, *rambled on.*
- *around* stands for location or motion (in different directions) often viewed from a central point, paths in all kinds of directions, e.g., *travel around*, *come (a)round*, *look around*, *bossing around*.
- *through* stands for motion inside an entity from end to end, activities viewed as complete(d) motions, e.g., *drive through*, *slept through*, *soaked through.*
- *over* is being or moving higher than and close to something or from one side to the other, examining thoroughly from all sides, e.g., *turning over*, *lingered over*.

In a similar fashion, other researchers have tried to group particles according to their meaning. By way of comparison to the particles above, the following definitions are worth quoting here:

- *on* means some more (Jackendoff 1997), expresses continuative action if used with activity verbs (Celce-Murcia and Larsen-Freeman 1999), i.e., verbs that express action not state, e.g., *carry on*, *keep on.*
- *around* stands for in a circle or with a circular motion, expresses absence of purpose (continuative) if used with activity verbs (Celce-Murcia and Larsen-Freeman 1999), e.g., *mess around*, *play around.*
- *through* means from beginning to end, e.g., *read through*, *think through.*
- *over* is again or re-, iterative if with activity verbs, e.g., *think over*, *do over (and over again)*, (Celce-Murcia and Larsen-Freeman 1999).

As can be noticed from the comparison of the two approaches, Jackendoff's (1997) and Celce-Murcia and Larsen-Freeman's (1999) definitions are simpler, which enables quicker grasping of the idea of particle, and thereby of PV transparency. Rudzka-Ostyn's (2003) definitions, on the other hand, are more elaborate, provide a thorough analysis within each particle and are accompanied by elaborate exercises, which is more suitable for self-conscious students.

5.2. The semantic approach to PV grouping: stage two (compositionality)

What follows from the grouping of PVs according to the particles is the semantic analysis of PVs, based on the idea of compositionality (Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004). Compositionality means that the inherent parts of a PV (verb and particle) either mean the same as they do when they are used on their own (as in *pulled up the anchor*), i.e. they are semantically transparent, or they are partially transparent (e.g., *locked up the office*, where *up* does not indicate upward position, but a redundancy aspect), or they cannot be semantically broken down at all (e.g., in *came across sth*, neither *came* nor *across* mean what they do when they are used on their own).

There are various labels for the same terminology in the literature. Table 7 sums up the major ones. Before going on to the analysis of the corpora, each of the semantic PV categories needs to be defined: opaque, semi-transparent and transparent.

Opaque PVs (Armstrong 2004) are also referred to as idiomatic (Jackendoff 1997; Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004) or noncompositional (Jackendoff 1997). These PV combinations consist of a verb and

a particle which are both opaque. Like other types of idiom, they are probably stored as whole units in the lexicon and as such they have to be 'learnt' as units, e.g. *come across*.

Semi-transparent PVs (Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004), also referred to as aspectual (Jackendoff 1997; Dehé 2002) or semi-transparent (Celce-Murcia and Larsen-Freeman 1999), are formed by a verb which retains its lexical meaning and a particle which does not (Armstrong 2004). Other researchers claim that what the particles express is not aspect, but *aktionsart* (e.g., Brinton 1988). Jackendoff (1997) and Celce-Murcia and Larsen-Freeman (1999) talk of aspectual PVs when the particle conveys its (aspectual) meaning. According to Jackendoff (1997: 541), "[i]n these cases, the particle does not satisfy an argument position of the verb; rather it contributes an aspectual sense, often paraphrased by some sort of adjunct PP. *Run/sing on*, for instance, means roughly 'run/sing some more'".

Fully transparent PVs (Armstrong 2004), otherwise called directional, literal (Celce-Murcia and Larsen-Freeman 1999) or fully compositional (Jackendoff 1997), are PVs in which the particle has a directional meaning and the verb is a verb of motion. The subject therefore moves in the direction specified by the particle in the manner specified. Here we can expect relatively little difficulty with the use of those PVs on the part of the learner (see, e.g., Celce-Murcia and Larsen-Freeman 1999). Some particles lend themselves easily to the transparent and fully compositional category, e.g., *back*, which is almost always directional, e.g., *come back*, *give back.* Another feature of compositional PVs is that they are self explanatory in terms of their semantics, hence there is no need for them to be listed as separate items in the lexicon. Jackendoff (1997: 541) stresses in this respect that "there is no need to list the verb-particle combinations in the lexicon, since the particle satisfies one of the verb's argument positions, and the meaning is fully compositional".

AUTHORS(S)	DEGREE OF IDIOMATICITY FROM HIGH TO LOW		
Quirk et al. 1972	highly idiomatic	semi-idiomatic	nonidiomatic
Jackendoff 1997	idiomatic = noncompositional meaning	aspectual	directional = compositional meaning
Celce-Murcia and Larsen-Freeman 1999	idiomatic	semi-transparent	literal
Armstrong 2004	idiomatic = opaque	aspectual = semi-transparent	directional = transparent

Table 7. PV compositionality

Both in Table 7 and in the description of the categories the PVs have been presented by introducing the degree of idiomaticity from absolute to none for clarity. However, it needs to be stressed that there is no direct endpoint to either of the categories, as is pointed out by Downing and Locke (2006: 343), who claim that "[i]t is by no means easy to establish boundaries between what is idiomatic and what is not", and by Biber et al. (1999), who maintain that verbs should rather be graded according to relative fixedness rather than to binary categories, given the categorization difficulties. Figure 2 below demonstrates the linear scale of idiomatic compositionality. It needs to be remembered that the categories are not points on the scale of compositionality. It is relative boundaries between them that need to be borne in mind, not so much the categories themselves. All of the PVs within each of the definitions of the compositional categories have been classified along the definitional criteria, and the outcome is summed up in Tables 8-10 below.

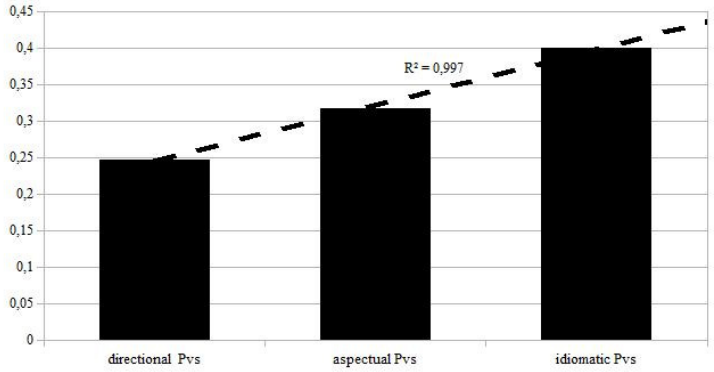


Figure 2. Incidence of PV use in the *LOCNEC* corpus expressed in PV TTR according to the categories of compositionality

The three tables below present different ways of looking at the data. Table 8 presents the types of PVs compared against the number of word types for the calculations to be valid. Chi-square values are presented in the last column to demonstrate statistical significance. Even assuming the strictest significance values, for $p<0.0001$, where the critical

value is 15.13, the chi-square values are still higher and therefore show significant differences between the two corpora in all three cases, with the opaque PVs showing the greatest discrepancy. One can obviously spot the gross differences between the corpora, but what strikes us immediately the most is that learners have a rather equal distribution of PVs, which might imply their having relatively few problems with idiomaticity.

Types account for only part of the picture, though. Table 9, in turn, concentrates on PV tokens derived from each corpus, and the chi-square values are also calculated here, this time against the number of word tokens, demonstrating again significant differences. What is striking in this comparison is that as idiomaticity grows, the significance diminishes. A quick look at the raw numbers reflects this tendency in both corpora. However, only a TTR comparison offers a deeper insight into the actual tendency of use when those two groups are compared.

COMPOSITIONALITY	NNS	NS	CHI-SQUARE
transparent PVs	27	85	30.06
semi-transparent PVs	31	89	28.06
opaque PVs	27	100	41.99

Table 8. Categories of PV compositionality in NS and NNS corpora, expressed in PV types

COMPOSITIONALITY	NNS	NS	CHI-SQUARE
transparent PVs	84	344	158.06
semi-transparent PVs	70	281	126.94
opaque PVs	73	250	97.08

Table 9. Categories of PV compositionality in NS and NNS corpora, expressed in PV tokens

COMPOSITIONALITY	NNS	NS
transparent PVs	32.1%	24.7%
semi-transparent PVs	44.2%	31.67%
opaque PVs	36.9%	40%

Table 10. Categories of PV compositionality in NS and NNS corpora, expressed in PV TTR

Table 10 sums up the TTR for each of the groups within the NNS and NS corpora, respectively. What can be concluded from this juxtaposition is that Polish speakers overuse the transparent category (32.1% vs. 24.7%), but their tendency of use is not linear as it is in the case of NSs. Natives exhibit more types than tokens of PVs as the level of idiomaticity grows (24.7% > 31.67% > 40%), while Polish speakers break the linearity at the level of the semi-transparent category. If the tendency was linear, there should be more than 44.2% of semi-transparent PVs used. At this point it seems valuable to compare the TTR values to the overall TTR for the general PV use, calculated at 37.44% and 31.31% for the NNS and the NS corpora, respectively (see Table 4). Such a comparison would enable us to conclude that learners' lexical variability probably stemmed from the use of transparent and semi-transparent PVs rather than the use of idiomatic PVs.

It is important to notice that the learner distribution of compositional categories turns out to be unequal only after comparison to the reference corpus. When looked at in isolation, learners employ all compositional categories of PVs with comparable 'ease' (see Tables 8 and 9). However, when compared to the reference corpus, where the distribution of the categories is not equal, the picture is different. In this respect, Celce-Murcia and Larsen-Freeman (1999) appear to be correct in their prediction that learners will be afraid or reluctant to use aspectual PVs, and even more so with idiomatic ones. It had been expected, however, that the tendency would be inversely proportional to the growth of idiomaticity, an expectation which did not come true.

Hypothesis number two can be verified at this stage. Table 10 above demonstrates that the tendency in PV use in the native group is directly proportional to idiomaticity. Just as idiomaticity grows, so does the use of PVs increase. However, in the non-native corpus, the tendency is not preserved. It is, however, also not inversely proportional, as had been predicted, so that the linearity of PV compositionality in the case of the non-native corpus is broken.

In trying to explain this unexpected distribution, one important observation must be made. Although PVs in general have been shown to be underused by learners, not many unusual, understood as non-dictionary PVs, can be noticed within the native corpus. This might be attributed to the conditions in which the data were collected. It was interview type of data rather than surreptitious recordings, so that the interviewees were fully aware that they were being recorded. Despite the advantages of such data, speakers could potentially have refrained from using vocabulary items they judged colloquial or inappropriate. After all, the material was being recorded in an academic environment and was intended to be made available for scientific use. Marks (2005: 12) stresses the fact that although the reality is more complicated, there is some basis for at least the first four of the beliefs that are still shared by teachers and students alike: that PVs are "colloquial, casual, informal, characteristic of speech rather than writing (...) and perhaps even a bit

sloppy or slovenly, uneducated, not quite proper". He calls them widespread popular wisdom about PVs among learners and teachers. The learners' proportionally higher semi-transparent PV use, on the other hand, may, interestingly, also stem from the same conditions. In the case of the Polish corpus, though, the recordings, although resembling an exam situation, were taking place in front of the interviewees' colleagues (which is reflected in the transcripts). On the one hand, this situation could have relaxed the subjects, by making them feel at ease. On the other, learners could have felt the pressure of being recorded, and therefore felt as if in an exam situation, which, in turn, could have made them recall and employ all grammar and vocabulary normally reserved for such occasions.

6. INDIVIDUAL LEARNER TENDENCIES

Apart from the verification of the major hypotheses, there are other observations to be made from the data analyzed. It was assumed that there would be differences in PV use resulting from the variation in foreign language experience of the learners. Despite the language level criterion being externally estimated in all *LINDSEI* language learners as advanced,[12] their exposure to the English language naturally varied. By "exposure to language" two factors are meant, namely, time spent in an English-speaking country and years of English language studied in a classroom situation prior to university.

The first look at the distribution of PVs among learners (see Figure 3) in relation to their length of stay in an English-speaking country does not clear up the picture. The lowest number of PVs used (counted in tokens) is 0 (learners nos. 25 and 48), the highest 27 (learner no. 24). When looking at the time spent in English-speaking countries, it is 2 and 4 months for the people who did not use PVs in their speech at all, and 0 months for the person who used 27 PVs. The list of the 8 highest values of PVs used, set against the length of stay in an English-speaking country, is presented in Table 11.

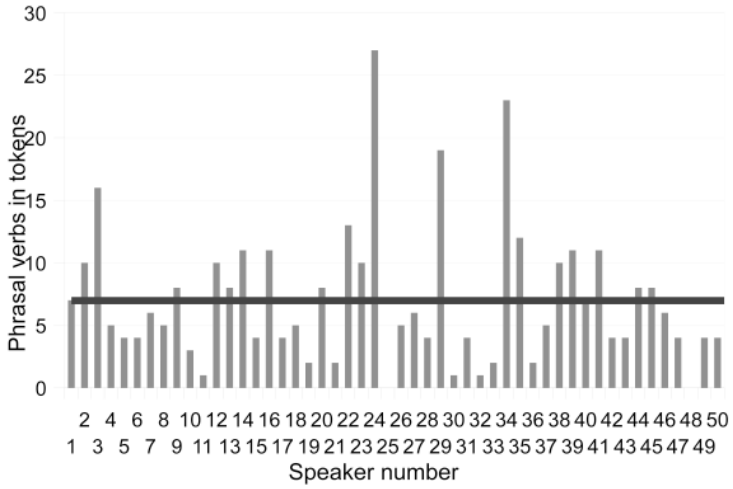


Figure 3. General PV distribution in individual users – *PLINDSEI*

Another factor worth taking into consideration is the number of years spent learning English in a classroom situation. One would expect that the students who had the longest exposure to English at school would have a good command of English PVs. One striking observation, however, is that the majority of students who had 8-11 years of English at school (longest periods) used 0-1 PVs in their conversations and only 1 student used 23 PVs. When we have a look from the perspective of learners with the highest token number of PVs in the corpus, the relationship between the token number and years of English at school is not clear either: these learners (tokens of PVs being 18-27) learnt English in a classroom situation from merely 4 years to as many as 10. In order to make sure if there is no correlation between the years of English at school and the number of PVs used, the Pearson correlation coefficient for the two variables was counted. The result, taking all 50 learners into account, is –0.02. We can thus safely confirm that it is difficult to say that it is the exposure to classroom English that predisposed any learners towards using PVs.

Most of the students also spent some time in an English speaking country so the correlation of this factor and the number of PVs used was calculated using Pearson product-moment correlation coefficient. The length of stay varies as much as 0-7 months, and the PV use within this group varies from 1 to 27. One speaker who did not travel to an

[12] All *LINDSEI* participants were third- or fourth-year 'English Language and Linguistics' University students.

English-speaking country happened to use 27 PVs; another used only 1 PV. Similarly, a 3-month stay brought about a result of 18 PVs in the interview of speaker no. 29, and only 1 PV of speaker no. 30. The Pearson correlation coefficient for the length of stay in an English-speaking country is 0.1. The results therefore suggest that no correlation between the number of phrasal verbs used and either of the ratio variables was found.

SPEAKER NO.	PVS USED	LENGTH OF STAY IN UK (IN MONTHS)	YEARS OF ENGLISH AT SCHOOL
24	27	0	6
29	18	3	4
34	23	1	10
25	0	2	11
48	0	4	5
11	1	0	8
30	1	3	8
32	1	7	8

Table 11. Number of PVs in relation to length of stay in UK and number of years of English at school

Thus, the third hypothesis, namely, assuming that language learning experience, defined as "years spent learning English in natural environment and in a classroom",[13] bears a noticeable influence on the quantity of learner PV use, cannot be confirmed. It turned out that the length of stay in an English-speaking country does not unravel the causes of PV underuse by learners, and neither does the length of English learning in school conditions, expressed in years. My supposition is that it is the quality of learning and teaching, rather than the number of years, that influences the use of PVs. As far as exposure to the English language in natural conditions is concerned, it involves more than months of passive living in a country for a foreign language learner to progress to a higher level.

As mentioned before, NSs also display different levels of command of vocabulary, even if we look at their PV distribution (Figure 4). Their PV use is, however, much more balanced than the learners', with the lowest PV token being 7, the highest 44. What is also characteristic of the NS use is high token and low type values in the most of them, which means that speakers have the tendency to repeat what is already in their repertoire, and so the PVs *go down* (9), *pick up* (a skill) (9) and *go over* (8) belong to the most commonly used PVs. On the one hand, *spread around*, *turn around*, *get away*, *look after*, *pay back*, *get by*, *put down*, *feel for*, *put on*, *leave out*, *take over*, *fall through* and *split up* are hapaxes. Had the corpus been larger, such hapaxes might have had greater occurrence and would have therefore exerted a stronger influence on the overall interpretation of the data.

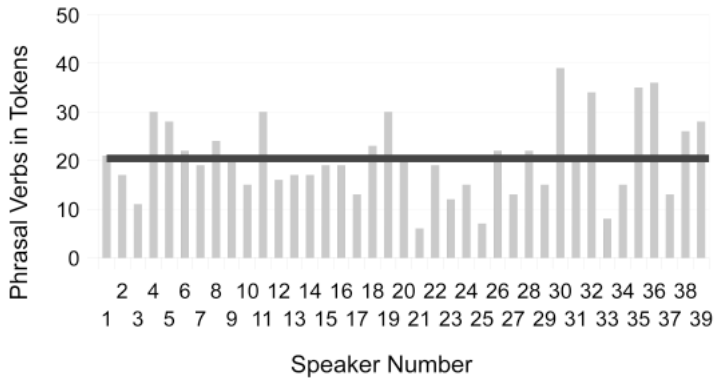


Figure 4. General PV distribution in individual users – *LOCNEC*

Finally, there is also a group of PVs common to both groups (native vs. learner) in terms of their tokens used. These are: *come back* (41 vs. 28), *go back* (56 vs. 13), *get back* (14 vs. 5), *come in* (13 vs. 2), *go out* 'leave' (26 vs. 15), *sit down* (15 vs. 3), *go on* 'continue' (32 vs. 10), *come on* (2 vs. 8), *wake up* (10 vs. 1), *go out* (socially) (51 vs. 7), *find out* (7 vs. 10), *make up* 'invent' (5 vs. 13) and *show off* (5 vs. 5). Apparently, out of the group of thirteen PVs in common, five belong to the transparent category of PVs, and four PVs belong to semi-transparent and idiomatically opaque categories. The sample is, however, too tiny to attempt any comparison, and the distributions of single PVs are not

[13] The choice of the given variables was motivated by the availability of metadata for *LINDSEI* sound files. Each of them is linked to a profile which contains information about the learner, the interviewer and the interview itself. This information makes it possible to study the potential influence of certain factors on learner language.

equal (e.g., single occurrence vs. 56 occurrences at times). Thus, the group of PVs common to both learners and natives cannot be compared so easily along the compositional category lines.

7. Conclusions and directions for further research

In this paper the problem of learner underuse of PVs has been presented using the example of Polish advanced speakers of English. General substantial underuse has been verified thanks to the employment of a POS tagged corpus, without which this research would not have been possible. PVs were divided along the lines of the semantic compositionality criterion, preceded by their classification according to the particle. What was found out in the analysis of PV compositionality is that while the native use of PVs is linear, learners do not appear to follow this tendency. They underuse PVs within all of the compositional categories, but the idiomatically opaque PVs are neglected the most.

In an attempt to find the key to this underuse, proficiency in English was called up. Neither of the two variables checked (length of stay in an English-speaking country and years of English at school) brought meaningful results, however. From the angle of language proficiency, it remains an open question where the observed differences stem from, as participants with similar education and language experience displayed varying degrees of PV use.

Another possibility is that learners simply avoided using PVs and tried using one-word equivalents instead. It is however debatable if the one-word equivalents truly reflect the meaning of the PVs. *Dress up* and *disguise* are approximate synonyms, where *disguise* suggests an intention to deceive while *dress up* does not. The PV *sail through something* means 'to succeed' and is roughly the equivalent of 'to pass' when referring to an exam. However, only in the case of PV use is there the connotation of effortlessness (see Marks 2005). Checking whether learners do consciously avoid PVs would naturally require a systematic study in order to find out what vocabulary items were used instead of PVs and if they were effective replacements.

Finally, as regards further research, it would be necessary to investigate into the reasons why certain learners underuse PVs more than others. As the available learner metadata did not provide an answer to this question, it might perhaps be more worthwhile to examine the way in which PVs are taught and learned in language course books. Further research into learner underuse of PVs might be to design an observation exercise or a questionnaire for teachers, and to observe which and how many PVs are used by teachers in their communication with students.

References

Aijmer, Karin (ed.). 2009. *Corpora and language teaching*. Amsterdam: John Benjamins.

Armstrong, Kevin. 2004. Sexing up the dossier: a semantic analysis of phrasal verbs for language teachers. *Language Awareness* 13: 213–224.

Bentivogli, Luisa Pamela Forner, Bernardo Magnini and Emanuele Pianta. 2004. Revising WordNet Domains hierarchy: semantics, coverage, and balancing. In *COLING 2004 Workshop on Multilingual Linguistic Resources, Geneva, Switzerland, August 28, 2004*, 101–108. <http://www.aclweb.org/ anthology-new/W/W04/W04-2200.pdf> and <http://wndomains.fbk.eu/publications/Coling-04-ws-WDH.pdf> (12 July 2013).

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Bolinger, Dwight. 1971. *The phrasal verb in English*. Cambridge, Mass.: Harvard University Press.

Brinton, Laurel J. 1988. *The development of English aspectual systems: aspectualizers and post-verbal particles.* Cambridge: Cambridge University Press.

Bywater, F.V. 1969. *A proficiency course in English*. London: University of London Press.

Cappelle, Bert. 2005. *Particle patterns in English: a comprehensive coverage*. PhD dissertation. Leuven: Katholieke Universiteit Leuven.

Celce-Murcia, Marianne and Diane Larsen-Freeman. 1999. *The grammar book*: *an ESL/EFL teacher's course*. Second edition. Boston, MA: Heinle and Heinle.

Courtney, Rosemary. 1983. *Longman dictionary of phrasal verbs*. Harlow: Longman.

Dagut, Menachem and Batia Laufer. 1985. Avoidance of phrasal verbs – a case for contrastive analysis. *Studies in Second Language Acquisition* 7/1: 73–79.

Darwin, Clayton M. and Loretta S. Gray. 1999. Going after the phrasal verb: an alternative approach to classification. *TESOL Quarterly* 33/1: 65–83.

Dehé, Nicole. 2002. *Particle verbs in English. Syntax, information structure and intonation*. Amsterdam: John Benjamins.

Downing, Angela and Phillip Locke. 2006. *English grammar. A university course*. Second edition. London: Routledge.

Ellis, Rod. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.

Garside, Roger. 1995. Grammatical tagging of the spoken part of the British National Corpus: a progress report. In Geoffrey Leech, Greg Myers and Jenny Thomas (eds.), *Spoken English on computer: transcription, mark-up and application*. Harlow: Longman, 161–167.

Gilquin, Gaëtanelle. 2012. The ups and downs of phrasal verbs in spoken and written learner Englishes. Paper presented at ICAME 33. University of Leuven, 30 May–3 June 2012.

Gilquin, Gaëtanelle, Sylvie De Cock and Sylviane Granger (comp.). 2010. *Louvain International Database of Spoken English Interlanguage* (*LINDSEI*). Louvain-la-Neuve: UCL Presses universitaires de Louvain.

Granger, Sylviane. 1996. From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, Bengt Altenberg and Mats Johansson (eds.), *Languages in contrast. Textbased cross-linguistic studies*. Lund: Lund University Press, 37–51.

Hampe, Beate. 2002. *Superlative verbs. A corpus-based study of semantic redundancy in English verb-particle constructions*. Tübingen: Gunter Narr Verlag.

Howarth, Peter Andrew. 1998. The phraseology of learners' academic writing. In Anthony Paul Cowie (ed.), *Phraseology*. Oxford: Clarendon Press, 161–186.

Jackendoff, Ray. 1997. Twistin' the night away. *Language* 73/3: 534–559.

Leech, Geoffrey. 2004. Adding linguistic annotation. In Martin Wynne (ed.), *Developing linguistic corpora*: *a guide to good practice*. <http://users.ox.ac.uk/~martinw/dlc/chapter2.htm> (7 December 2011).

Marks, Jonathan. 2005. Phrasal verbs and other 'phrasal' vocabulary. In Michael Rundell (ed.), *Macmillan phrasal verbs plus.* Oxford: Macmillan.

Mukherjee, Joybrato. 2007. Exploring and annotating a spoken English learner corpus: a work-in-progress report. In Sabine Volk-Birke and Julia Lippert (eds.), *Proceedings of Anglistentag 2006*. Trier: Wissenschaftlicher Verlag Trier, 365–375.

Pelli, Mario G. 1976. *Verb-particle constructions in American English*. Bern: Francke.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1972. *A grammar of contemporary English.* Harlow: Longman.

Rudzka-Ostyn, Brygida. 2003. *Word power: phrasal verbs and compounds. A cognitive approach*. Berlin: Mouton de Gruyter.

Rundell, Michael (ed.). 2005. *Macmillan phrasal verbs plus*. Oxford: Macmillan.

Scott, Mike. 2008. *WordSmith Tools* version 5. Liverpool: Lexical Analysis Software.

Sinclair, John. 1991. *Corpus, concordance, collocation: describing English language.* Oxford: Oxford University Press.

Sroka, Kazimierz A. 1972. *The syntax of English phrasal verbs*. The Hague: Mouton de Gruyter.

Call for Papers and Special Issues

Aims and Scope

Research in Corpus Linguistics (RiCL, ISSN 2243-4712) is a scholarly peer-reviewed international scientific journal published annually, aiming at the publication of contributions which contain empirical analyses of data from different languages and from different theoretical perspectives and frameworks, with the goal of improving our knowledge about the grammar and the linguistic theoretical background of a language, a language family or any type of cross-linguistic phenomena/constructions/assumptions.

RiCL invites original, previously unpublished research articles and book reviews in the field of Corpus Linguistics. The journal also considers the publication of special issues on specific topics, whose edition can be offered to leading scholars in the field. These areas include, but are not limited to, the following topics:

- Corpus design, compilation and typology
- Discourse, literary analysis and corpora
- Corpus-based grammatical studies
- Corpus-based lexicology and lexicography
- Corpora, contrastive studies and translation
- Corpus and linguistic variation
- Corpus-based computational linguistics
- Corpora, language acquisition and teaching
- Special uses of corpus linguistics

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference A special issue is typically made of 6 to 10 papers, with each paper 12 to 20 pages of length.

A special issue can also be proposed for selected top papers of a conference/workshop. In this case, the special issue is usually released in association with the committee members of the conference/workshop like general chairs and/or program chairs who are appointed as the guest editors of the special issue.

The following information should be included as part of the proposal:

- Proposed title for the special issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers
- Name, contact, position, affiliation, and biography of the guest editor(s)
- List of potential reviewers if available
- Potential authors to the issue if available
- Estimated number of papers to accept in the special issue
- Tentative time-table for the call for papers and reviews, including
    - Notification of acceptance and revision guidelines
    - Final submission due
    - Time to deliver final package to the publisher

If the proposal is for selected papers of a conference/workshop, the following information should be included as part of the proposal as well:

- The name of the conference/workshop, and the URL of the event.
- A brief description of the technical issues that the conference/workshop addresses, highlighting the relevance for the journal.
- A brief description of the event, including: number of submitted and accepted papers, and number of attendees. If these numbers are not yet available, please refer to previous events. First time conference/workshops, please report the estimated figures.
- Publisher and indexing of the conference proceedings

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the “Call for Papers”.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should read the Author Guide.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing an introductory editorial to be published in the Special Issue.

Please send your proposal to **Editor-in-Chief, Prof. Javier Pérez-Guerra**, (Email: jperez@uvigo.es).

More information is available on the web site at http://www.academypublisher.com/ricl/.