Univerzita Karlova v Praze

Filozofická fakulta

Ústav anglického jazyka a didaktiky

Diplomová práce

Lucie Gillová

Tagging a spoken learner corpus

Značkování žákovského korpusu mluvené angličtiny

Praha, 2014 vedoucí práce: PhDr. Tomáš Gráf

Ráda bych poděkovala vedoucímu své diplomové práce PhDr. Tomáši Gráfovi za odborné vedení, cenné rady a obrovskou trpělivost.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 12.8. 2014

podpis

Abstract

The aim of the thesis is to propose a tagging system for a learner corpus of spoken English which would, apart from tagging errors, focus also on the features specific for spoken language. Theoretical part, therefore, introduces basic concepts including learner language, the development of learner corpora in the last 20 years and both classical and computer-aided error analysis. Features typical of spoken language are described in the theoretical part as well since these are the focus of the research part of the thesis. The Louvain tagging system used for error-tagging of a leaner corpus of written language is used as the basis for the tagging system proposed in this thesis. Based on the analysis of 20 transcriptions taken from the Czech part of spoken learner corpus LINDSEI, modifications of the categories taken from the Louvain error-tagging system are proposed and new categories necessary for a better description of spoken language are introduced. The tagging system proposed in this thesis should make further analysis of the tagged corpus easier.

Key words: spoken language, learner language, learner corpora, error analysis, error tagging

Abstrakt

Cílem této práce je navrhnout systém značkování žákovského korpusu mluvené angličtiny, který by se kromě chyb zaměřoval i na značkování specifik mluveného jazyka. V teoretické části proto práce stručně nastiňuje žákovský jazyk jako takový, vznik a vývoj žákovských korpusů v posledních 20 letech a jak klasickou, tak počítačem podporovanou chybovou analýzu. Kromě toho jsou v teoretické části popsána specifika mluveného jazyka, na která se pak soustřeďuje část praktická. Jako základ pro navrhovaný systém značkování je použit Lovaňský značkovací systém, který je ale určený pro žákovský korpus psaného jazyka. Na základě analýzy přepisů 20 nahrávek z české části žákovského korpusu LINDSEI jsou navrženy úpravy kategorií stávajících a kategorie nové, které by měly lépe zachytit prvky typické pro mluvený jazyk a tak usnadnit jeho analýzu po označkování celého korpusu.

Klíčová slova: mluvený jazyk, žákovský jazyk, žákovské korpusy, chybová analýza, značkování chyb

Table of Contents

List of Figures

List of Abbreviations

BASE	British Academic Spoken English corpus
BAWE	British Academic Written English corpus
BNC	British National Corpus
CEA	Computer-aided Error Analysis
CECL	Centre for English Corpus Linguistics
CIA	Contrastive Interlanguage Analysis
CIC	Cambridge International Corpus
CLC	Cambridge Learner Corpus
COBUILD	Collins Birmingham University International Language Database
CQL	Contextual Query Language
EA	Error Analysis
ELF	English as a Lingua Franca
ELFA	English as a Lingua Franca in Academic Settings
ELT	English Language Teaching
FRED	Freiburg English Dialect Corpus
FRIDA	French Interlanguage Database
ICE	International Corpus of English
ICLE	International Corpus of Learner English
JEFLL	Japanese English as a Foreign Language Learner corpus
L1	first language (mother tongue)
L1A	first language acquisition
L2A	second language acquisition

LAD	Language Acquisition Device
LANCAWE	Lancaster Corpus of Academic Written English
LINDSEI	Louvain International Database of Spoken English Interlanguage
LOCNEC	Louvain Corpus of Native English Conversation
LOCNESS	Louvain Corpus of Native English Essays
LONGDALE	Longitudinal Database of Learner English
MICASE	Michigan Corpus of Academic Spoken English
NICT JLE	National Institute of Information and Communication Technology - Japanese Learner English
OCR	Optical Character Recognition
POS tagging	Part-of-speech Tagging
SLA	Second Language Acquisition
UCL	Université catholique de Louvain
VESPA	The Varieties of English for Specific Purposes dAtabase
VOICE	Vienna Oxford International Corpus of English
XML	Extensible Markup Language

1. Introduction

The study of learner language has become easier with the possibilities brought by the learner corpora. Corpus linguistics enables the researchers to focus on learner language and to collect and process a large quantity of data, something that would have been impossible when learner language research focusing on learner language as a phenomenon worth studying on its own was established in the 1960s. To process a large amount of data it is important to be able to search it for various features without the need to go through it word by word. This is where tagging plays an important role. Primarily, most corpora are tagged for parts of speech. However, since one of the most important features of learner language are errors made by learners, errors are the feature that most of the learner corpora are tagged for and various tagging systems have been developed for error tagging of learner corpora.

Similar to non-learner corpora, most learner corpora deal with written language because processing spoken language is still much more difficult and time consuming than processing written language. Most tagging systems in use are thus designed for written corpora, not taking into consideration features typical of spoken language (an exception being for example the NICT JLE corpus – *National Institute of Information and Communication Technology – Japanese Learner English corpus*). The aim of this thesis is to analyze data from a spoken learner corpus and based on this analysis to propose changes in a tagging system necessary to capture the specific features of spoken language. A corpus recorded at the Department of English Language and ELT Methodology at Charles University will be used because besides transcripts of the Czech learners of English, recordings are available as well which should be essential during the error identification process. The corpus is a part of a big international corpus of learner English called LINDSEI (*Louvain International Database of Spoken English Interlanguage*) which contains data from speakers of various L1s (first or native languages) but recordings are not a part of the corpus and so only the Czech part will be used in this thesis. Another reason for using this corpus is the tagging system used as the basis for a new one proposed in this thesis. Louvain error-tagging system was developed by Louvain research group led by Granger for a written corpus of learner English (ICLE – *International Corpus of Learner English*). ICLE is a written counterpart of LINDSEI.

To introduce the topic of the thesis, its first chapter will deal with learner language, its definition and the historical development of its study. Error analysis will be briefly described in the next chapter, focussing on the definition of error as well. The third chapter will introduce learner corpora, their variability and possible applications and will focus on LINDSEI in greater detail. Linked to learner corpora, error tagging, with a special emphasis on the Louvain tagging manual, will be described. Since the focus of this thesis is spoken language, it will be discussed in the next chapter and its idiosyncrasies will be described in detail. Lastly, spoken corpora will be briefly described to show how they make spoken language research easier. Following the description of the methodology used in this thesis, the analysis of the data in connection to possible changes in the tagging manual will be the last part of the thesis. This part should show the areas in which the error tagging system should take into account specific features of the spoken language and it could also possibly introduce tags that do not indicate a mistake but rather a feature specific for spoken language (e.g. expressions such as pronouns which are used differently because of different processing of spoken language). Introducing these new tags should enable further research of the specific features of spoken language in order to identify what is typical of English language learners and what is the same for both English language learners and native speakers of English.

2. Learner Language

2.1.Historical Overview

The notion of learner language as something important on its own dates back to the 1960s and 1970s when some researchers concerned with language teaching started to note that describing spoken or written language produced by a language learner as an imperfect reproduction of the target language does not seem to capture all important aspects of this learner product. Behaviourist theory of L2A was the prevailing theory in the 1940s and 1950s. The theory assumed that language learning was, similar to any other kind of learning, only a habit-formation and in the process of learning a language, the old habits are replaced with new ones. Therefore, learner language was only compared with the target language because it was treated as an imperfect product which should be perfected by further teaching of the problematic pieces of the target language. As a result, the only important task in analysing learner language was to compare grammatical structures of the mother tongue to the grammatical structures of the target language and, based on this comparative analysis, predict what would be problematic for learners and adjust teaching of the target language accordingly. Analysis of learner language was done only to confirm problematic areas discovered in the comparative analysis of the two languages.

The shift of perspective was facilitated by Chomsky's conception of the way human beings learn their mother tongues. In 1965, Chomsky came with the idea of a language acquisition device (LAD) that enables all of us to learn languages. He claimed that this device was universal and only thing being learnt, or more accurately set to correct values, by small children, were the grammatical parameters of a given language (Chomsky, 1965: 25). Children are exposed to a language and they are building their own grammars or representations of the language from the input. While processing the input they have received from their parents or other people they have come in contact with, they start to use their internal grammars eventually to reach a stage in which they could produce a potentially infinite number of correct sentences. Chomsky's concept of language learning was developed to describe first language acquisition. However, it was adapted by researchers in second language acquisition (SLA)[1] to describe learning of basically any language and thus caused a shift from the study of language teaching to the study of

[1] For the distinction between SLA and L2A, see 2.2.

language learning and more importantly to the study of learner language as an important manifestation of the learning process in L2A.

Corder's article 'The significance of learners' errors' from 1967 (reprinted in 1981a) is considered to be the start of a shift in perspective on the language learning processes because he abandoned the behaviourist view that language learning should be based on repeated teaching of various aspects of a language until they are perfected by the learner and any mistakes made in the process are either signs of the imperfect process of teaching or just signs that human beings are imperfect and so is sometimes their language. In his seminal article, Corder focuses on errors and their significance in the study of SLA and he claims that they are important because they allow us to observe the dynamic language system the learner is using. Corder (1967/1981b: 10) claims that "[a] learner is using a definite system of language at every point in his development" and he compares L2A (second language acquisition) with L1A (first language acquisition), claiming that both these processes are very similar. However, he still maintains the distinction between acquisition (of a mother tongue) and learning (of a second language). The main distinction between these two terms lies in the inevitability of the L1A (at least under normal conditions) and also in a developmental stage at which L1 or L2 are acquired. He proposes that since we do not take mistakes made by a child acquiring his/her mother tongue as something condemnable, we should look at mistakes made by the learner as being significant for the learner's language system at a particular time similarly. Following his assertion of the significance of learners' errors, he concludes his article with stressing the importance of studying learner language for the improvement of current teaching practices.

Similarly to other researchers, Corder later came with a term for the description of learner language. In his article published in 1971 (and reprinted in 1981a), he proposes that learner language is a type of idiosyncratic dialect. He claims that unlike social dialects, an idiosyncratic dialect has a set of rules that are unique for the particular speaker and are never shared as a whole set by another speaker of the same language. He describes several types of idiosyncratic dialects: poetic language, language produced by people suffering from aphasia and language produced by children learning their mother tongue. The fourth class of idiosyncratic dialects is learner language. According to Corder (1971/1981c: 17),

> [i]t is regular, systematic, meaningful, i.e. it has a grammar, and is, in principle, describable in terms of a set of rules, some sub-set of which is a sub-set of the

> rules of the target social dialect. His [learner's] dialect is unstable (we hope) and is not, so far as we know, a 'langue' in that its conventions are not shared by a social group [...], and lastly, many of its sentences present problems of interpretation to any native speaker of the target dialect.

Corder (1976/1981d: 66–67) acknowledges that he is describing the same phenomenon described by Selinker (1972) as interlanguage and by Nemser (1971; quoted in Corder 1976/1981d: 66–67) as an approximative system, pointing out that each of the terms stresses a different aspect of learner language: Selinker is emphasizing the position of the learner language between L1 and L2: it is a mixed system; Nemser stresses the "goal-directed development of the learner's language towards the target language system" (Corder, 1976/1981d: 66); Corder's own term – idiosyncratic or transitional dialect – was later changed to transitional competence (to show the connection with competence described by Chomsky) which emphasizes that the learner has some knowledge of the language system (he is competent) and the knowledge is developing (transitional).

Although several terms for learner language were introduced, only Selinker's (1972) term interlanguage has gained acceptance and is still widely used. He observes that utterances produced by the L2 learners are not the same as utterances produced by the native speakers of the target language. Given this difference, it is logical to assume that there is a separate linguistic system which Selinker calls interlanguage.

He, similarly to Corder, focuses on the differences between interlanguage and target language. He also introduces processes responsible for those differences. It is important that not all of them are caused by the interference of L1. The first one is language transfer which means occurrence of "fossilizable items, rules, and subsystems" (Selinker, 1972: 216) that are part of the native language of the learner. The second process is transfer-of-training which describes the features of the interlanguage traceable back to the strategies used in language teaching. Strategies used by the learner to learn the material given to him can also influence the interlanguage and are called by Selinker strategies of second-language learning. The fourth process is again learner-centred and involves the strategies of second-language communication, meaning the strategies used by the learner to communicate with a native speaker of the target language. The last process described is the overgeneralization of the target language linguistic material. According to Selinker (1972: 217): [c]ombinations of these processes produce what we might term entirely fossilized linguistic competence." However, he does not provide any more detailed explanations as to the processes listed above and his term for learner language seems to impose the notion that it is something that is not developing or if so, there are always some structures that fossilize and it is not clear from his article whether the process of fossilization can be reversed.

All these problematic aspects of Selinker's article have only encouraged other researchers to focus on learner language in greater detail and many studies shedding light on interlanguage have been conducted since the publication of the original article, developing the notion of interlanguage, arguing with the irreversibility of the fossilization process and proposing alternative theories to the idea of LAD or similar inner systems.

To conclude, the theories of L2 acquisition have developed significantly over the past 60–70 years. Starting with behaviourist theories, the only aspect in which the study of learner language was important was to confirm that transfer from learner's L1 occurs and to find the problematic areas. This perspective changed with innatist theories claiming that LAD is responsible for L1A. These theories assumed that LAD was reactivated in the process of L2A and, more importantly, they did not necessarily attribute too much importance to L1 influence, thus making the study of learner language substantial for the discovery of the natural order of learning a particular language. Later on, the notion of LAD was abandoned by many researchers and cognitivist theories state that languages are learnt using the same cognitive processes involved in other kinds of learning (thus partly going back to behaviourist theories but assuming different learning processes, not habit-formation). Lastly, the importance of the social interactions have also been stressed in SLA research (for more information on learner language research, see Tarone and Swierzbin, 2009).

2.2.Defining learner language

For the purposes of this thesis, learner language will be defined as spoken or written language produced by learners. As Ellis and Barkhuizen (2005: 4) emphasize: "learner language is not a monolithic phenomenon but rather highly variable" and "it is [also] not the only type of data available to SLA researchers" (ibid.). Those statements are right, considering the experiments using various technology in SLA research and also the various ways of eliciting data for learner language research. However, the other possibilities of studying L2A will not be discussed here in detail since the thesis deals with learner language in general and does not aim at describing the field of SLA.

One more definition is needed – terms SLA and L2A are not used interchangeably in this thesis. The definition by Ellis and Barkhuizen (2005) is adopted; SLA thus refers to the field of study while L2A describes the process of L2 learning. It is, nevertheless, important to bear in mind that these terms are sometimes still used interchangeably in literature.

3. Error analysis

Learner language is most often studied using the description of the mistakes learners make in their spoken or written utterances. As noted in Chapter 2.1, mistakes made by learners were studied even before the change in perspective on the importance of learner language. However, with Corder's article in 1967, the study of learner language has become much more important because he emphasized the significance of the mistakes made by learners for the better understanding of the learning process. The article basically started what he later named error analysis, a field of study that has been widely criticized but brought some important findings and has been partly 'resurrected' with the development of learner corpus research. It is concerned with identifying the learners' errors and explaining them in terms of their possible origin.

This chapter will describe the history of error analysis and also focus on the way various authors define errors and the way the error will be defined for the purposes of this thesis.

3.1.History of error analysis

Prior to Corder (1967/1981b), errors were considered to be manifestations of language transfer or of the imperfection of the teaching method used. This was a typically behaviourist framework, as discussed above. However, Corder (1967/1981b) established a different way of seeing learners' errors. He claimed that errors were so important because they allowed us to study learner language since we do not have any other way of understanding the underlying system the learner uses when he writes or speaks. In this article, he establishes a distinction between systematic and non-systematic errors. He emphasizes that even in our own native speech, we often make errors due to various factors such as "memory lapses, physical states such as tiredness, and psychological conditions such as strong emotions" (Corder, 1967/1981b: 10). He claims that these errors occur in learner language as well and are not systematic because the learner knows the rule and usually uses it. He proposes to call them mistakes, as opposed to errors which are systematic and originates from the learner's lack of knowledge of the target language.

In his subsequent articles, Corder (1981a: 36) established a procedure of conducting an error analysis which consists of three steps. The first one is to identify the error. This is closely connected with the definition of error as outlined in 3.2. The second step in error analysis is the description of error. This step should lead to establishing error categories which would allow researchers to count error types frequencies (cf. section 4.1.2.1 on error tagging systems which are basically fulfilling this step). The last step in the error analysis should be error explanation. Since Corder no longer accepted the grounds of behaviourism, this step is really important because it should lead to decision whether the error is caused by L1 interference or whether it is caused by some other factor such as general learning strategies etc.

Starting with Corder's article, error analysis was really popular (see Spillner, 1991) but it was also widely criticized. Schachter and Celce-Murcia (1977) provide a systematic criticism of error analysis, showing also how the individual problems of error analysis are connected. One of the problems of error analysis is that it focuses on errors in isolation, researchers usually did not take into account the context. Since the errors are analyzed in isolation, there might be mistakes in their classification and following the wrong classification, wrong frequencies of different types of errors would be counted. Since the classification can be faulty and frequencies not counted correctly, any conclusions about the difficult areas of the target language are not very reliable. Another important concern is the way researchers have identified causes of systematic errors. Probably because Corder listed it as one of the three basic steps of error analysis, conclusions about the errors' origin were drawn much too easily. The last problem of error analysis, according to Schachter and Celce-Murcia (1977) is the sampling process used in majority of the studies. The authors usually worked with a very limited set of data which could cause biased results because they were not representative of the particular learner language or of learner language in general.

3.2.Definition of an error

One of the most difficult parts of error analysis is defining what error is and what is not. There are several reasons for this difficulty: first of all, languages are constantly changing and the norms develop and change as well so it is sometimes difficult to draw a line between what is and what is not acceptable, especially when considering a language such as English which does not have a codified rules such as for example Czech does (*Pravidla českého pravopisu*). Secondly, the lack of codified rules is even more complicated in the case of English language because of the number of speakers of English and consequently the number of varieties of English. Even if a grammar book is taken as a basis for the definition of a standard, most of the grammar books still focus on written language and do not provide sufficient description of spoken language (at least of spoken English).

Corder in 1967 (1967/1981b) did not try to define errors at all, he only focused on the significance of errors and later (Corder 1971/1981c) distinguished them from mistakes (see 3.1). He provided an algorithm which shows how to analyze learner language and how to find where it differs from the target language, see Fig. 1. The diagram shows that not all that appears as a normal sentence in the target language can be interpreted as such (covertly idiosyncratic sentence) which is important for the error identification process but the diagram does not include any attempt at defining an error. Even when Corder describes the three steps in error analysis, he does not discuss the identification of an error in greater detail. The lack of a definition of an error is one of the points error analysis was criticized for.

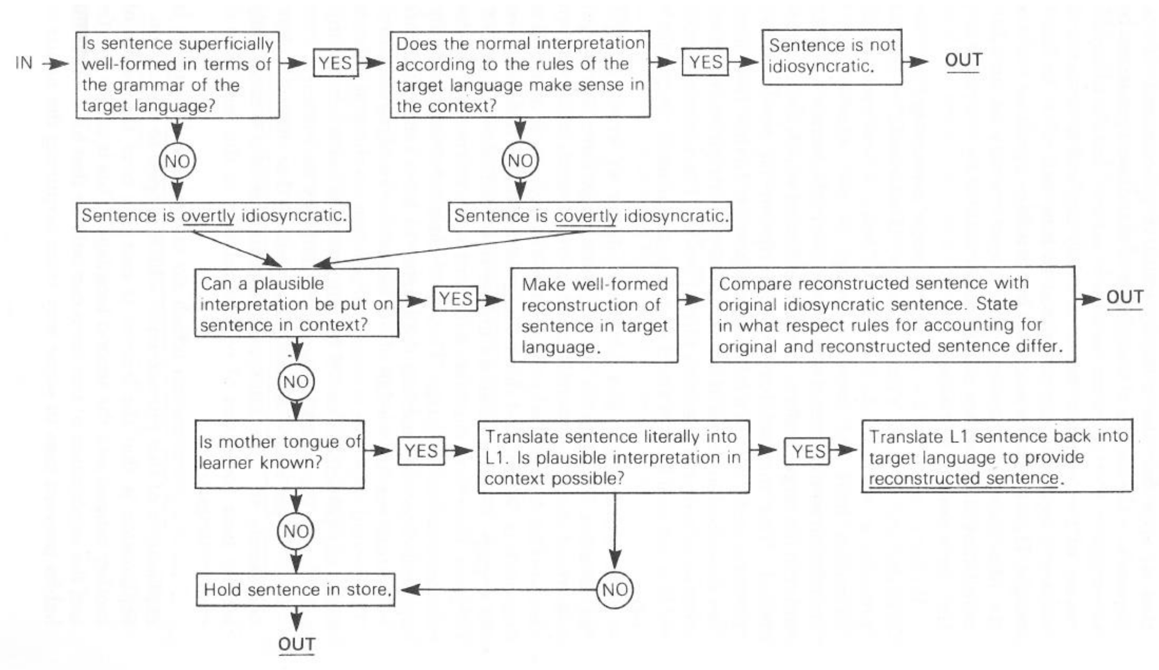
IN
Is sentence superficially well-formed in terms of the grammar of the target language?
YES
Does the normal interpretation according to the rules of the target language make sense in the context?
YES
Sentence is not idiosyncratic.
OUT
NO
NO
Sentence is overtly idiosyncratic.
Sentence is covertly idiosyncratic.
Can a plausible interpretation be put on sentence in context?
YES
Make well-formed reconstruction of sentence in target language.
Compare reconstructed sentence with original idiosyncratic sentence. State in what respect rules for accounting for original and reconstructed sentence differ.
OUT
NO
Is mother tongue of learner known?
YES
Translate sentence literally into L1. Is plausible interpretation in context possible?
YES
Translate L1 sentence back into target language to provide reconstructed sentence.
NO
Hold sentence in store.
NO
OUT

Fig. 1 Corder (1971/1981c) - Algorithm for providing data for description of idiosyncratic dialects

Later, there were numerous attempts to define what an error is and they usually differ according to the research aims of a particular researcher. There is still no single definition that would be accepted by the majority of researchers. Thornbury (2006: 75) defines error as "an instance of the learner’s language that does not conform to accepted norms of usage, and which is attributed to incomplete or faulty learning. These norms by which errors are judged are usually defined in terms of adult native speakers of Standard English." Lennon

(1991: 182) provides a broader definition of error, stating it is "[a] linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterparts." The second definition takes into account for example also differences between registers. For the purposes of this study, error should be understood as a deviation from the accepted norms of usage in Standard British English. These norms should be based on the description of the English language provided in *Mluvnice současné angličtina na pozadí češtiny* (Dušková et al., 2006) and will be supplemented by *Longman Grammar of Spoken and Written English* (Biber et al., 1999).

4. Learner Corpus Research

Learner corpus is a collection of authentic utterances produced by learners of some second language although authenticity of the utterances can be quite difficult to ensure since the situations in which the data are collected are often quite unnatural (unlike corpora of native speakers' utterances). Learner corpora can be either spoken or written (or combined), can be based on a single second language or can be multilingual; they can also include utterances in the target language produced by speakers of various L1s. Given the character of learner corpora, diverse types of variables need to be recorded and made available to the linguists using a particular corpus; these include age, sex, education in general, mother tongue(s), second language(s) and proficiency level as well as information about the task used to elicit the data such as type of task, information about the person eliciting the data, time limits etc.

Error analysis discussed in Chapter 3 has become once again popular with the development of learner corpora which brought new possibilities to the field of SLA research and to the study of learner language. Similarly to other branches of linguistics, corpus research allows linguists to study larger quantity of data and to search them for specific features which may be important for various purposes. Learner corpus research is quite new, it dates back to the 1980s but many learner corpora have been collected since and are used for the study of learner language which is one of the ways to study L2 acquisition. To show the development of this field, various learner corpora will be introduced in this chapter, with special attention to LINDSEI (see below). Learner corpora are usually tagged for the errors made by learners and so the second part of this chapter will deal with the error tagging systems and their connection to error analysis. The Louvain error-tagging system will be described in greater detail because it will be the basis for the tagging system proposed here.

4.1. Learner Corpora

As stated above, learner corpus research dates back to late 1980s when researchers started to realize possibilities learner corpora provide them in the research of second language acquisition and in the research of L1 (first or native language) and L2 (second or target language) interference. Since then, many learner corpora have been built, large part of them based on English as a second language but with different native languages of the speakers (others include for example German, French, Italian, Spanish or Finnish as L2). The proportion of the target languages among the learner corpora is presented in Fig. 2, confirming that English is the language studied most often (in 78 out of 130 learner corpora).

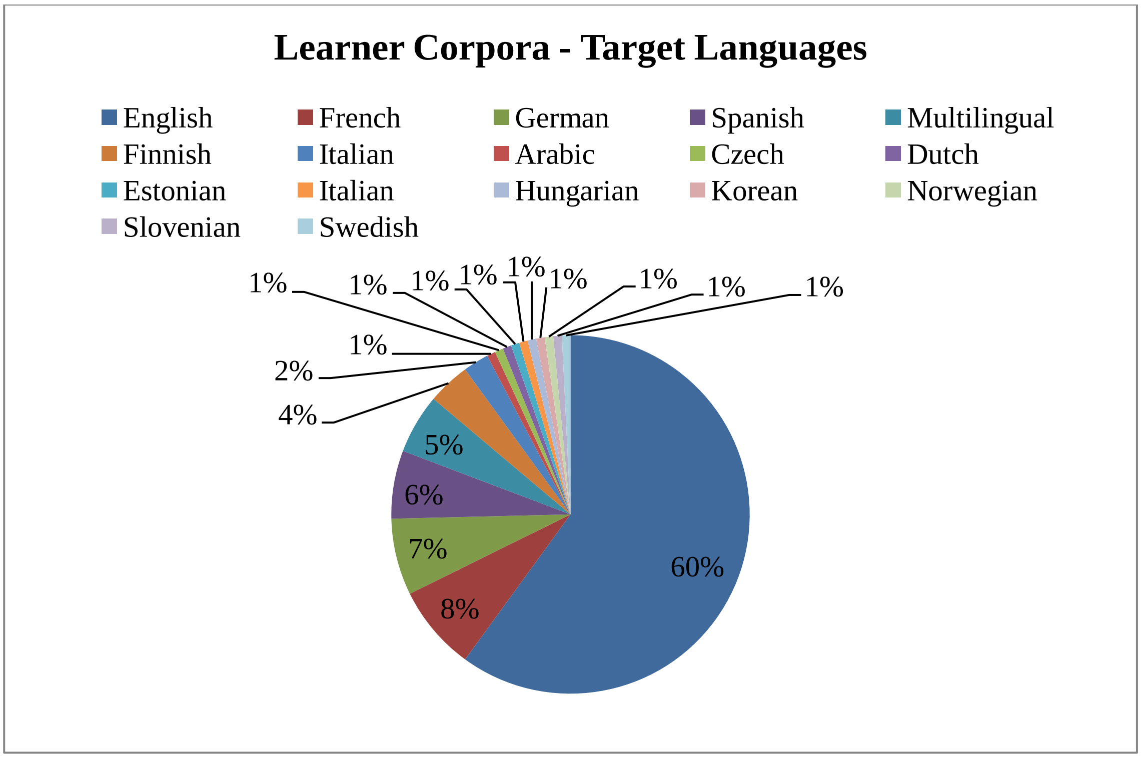

Fig. 2 Learner corpora according to the target language[2]

The majority of learner corpora is written, including corpora such as ICLE (*the International Corpus of Learner English*), JEFLL (*Japanese English as a Foreign Language Learner corpus*) or LANCAWE (*Lancaster Corpus of Academic Written English*). There are several corpora that include both written and spoken language (e.g. the *Barcelona English Language Corpus*). With the advancement of modern technologies that enable much easier sound recording and sound processing, the number of spoken corpora has increased although they are smaller than written corpora (OCR – optical character recognition – being much easier and less time-consuming than transcribing the recordings). Learner corpora have also followed that trend; there are for example *The ANGLISH*

[2] based on the list of corpora at http://www.uclouvain.be/en-cecl-lcworld.html which also includes corpora that contain L1 speakers of the target language (and the percentage of L1 and L2 target language speakers is usually not given)

corpus, The Eastern European English learner corpus or LINDSEI (*The Louvain International Database of Spoken English Interlanguage*).

Overall, there is a number of learner corpora varying in size and the type of data collected, suitable for different research aims which are defined prior to the data collection. For the purposes of this thesis, only Louvain learner corpora will be described in greater detail, with the main focus on LINDSEI.

4.1.1. Louvain Learner Corpora[3]

CECL (Centre for English Corpus Linguistics) at Université catholique de Louvain (UCL) is responsible for compilation of several corpora, besides learner corpora they also compile pedagogical corpora for the study of teaching materials and have also collected data for two corpora of native speakers of English (LOCNESS and LOCNEC) that will be described in detail later. The centre was founded in 1990 by Sylviane Granger, one of the leading figures in learner corpus research, and in the same year, a work on the first learner corpus compiled in Louvain began (De Cock, 2011). Since learner corpora in Louvain are one of the first compiled, many methodological issues were addressed by CECL researchers, including selection of the data for a learner corpus, error tagging and even compilation of comparative corpora of native speakers.

The first and probably best known learner corpora compiled in CECL is ICLE (*International Corpus of Learner English*) which contains argumentative essays written by higher intermediate or advanced learners of English. ICLE contains essays written by students with different mother tongues thanks to the collaboration of several partner universities of UCL: Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana and Turkish. Although the compilation of the corpus started in the early 1990s, the first version of this corpus was not published until 2002. According to the ICLE websites, the researchers are working on the third version of the corpus now. The current, second version of the corpus contains 3.7 million words and users can use learner variables including detailed language information (mother's mother tongue, father's mother tongue, language of instruction at various stages of education) and also information on age, sex and stays in English-speaking countries. The corpus contains two types of text – argumentative essays and literature examination

[3] All the information was retrieved from http://www.uclouvain.be/en-258636.html

papers (the latter should not amount to more than 25% of a national subcorpus). Both types of text can be written at home and students are allowed to use dictionaries and grammar books which is probably the biggest problem of this corpus. Many interesting features of learner language cannot be observed in ICLE because for example the richness of vocabulary or errors in some problematic structures can be influenced by the use of dictionaries and reference books.

CECL compiles other learner corpora, most of them focusing on English as a target language. There are two interesting corpora being built now, first of them is VESPA (*The Varieties of English for Specific Purposes dAtabase*). The researchers want to compile a corpus that would contain written texts by L2 English speakers and would include texts from various scientific disciplines, various types of texts and also texts from writers at different stages of study (from BA to PhD students). The second interesting project of CECL is LONGDALE (*Longitudinal Database of Learner English*) that aims to collect longitudinal data from university students. Like VESPA it started in 2008, and several universities participate in the data collection. The data collection process is quite simple, students are given four topics for an argumentative essay, they should write between 500 and 700 words and they write an argumentative essay (on different topic) every year while they are studying at university. The data in both of these corpora are treated in the same way as data in ICLE. The only non-English corpus in Louvain is FRIDA (*French Interlanguage Database*) which contains texts written by learners of French. The corpus is divided into three sections – texts by Dutch speakers, texts by English speakers and texts by speakers from various mother-tongue backgrounds. The corpus is error-tagged.

Lastly, Louvain also compiles two corpora by native speakers of English. LOCNESS (*The Louvain Corpus of Native English Essays*) is a corpus made up of British university students' essays, British pupils' A level essays and American university students' essays. It was compiled to have a set of data produced by native speakers of English, the data that could be compared with the findings from ICLE. LOCNEC (*The Louvain Corpus of Native English Conversation*) is the second corpus, containing spoken utterances by native speakers at British universities. The speakers were performing the same tasks as speakers in LINDSEI so LOCNEC provides comparable data for spoken English.

4.1.1.1. LINDSEI

LINDSEI (*Louvain International Database of Spoken English Interlanguage*) is the last of the learner corpora compiled in CECL. It will be described in detail because it is the corpus used in this thesis. The compilation of LINDSEI started in 1995 and it is still being built. There are currently eleven complete parts that are transcribed (complete means containing utterances by all 50 speakers of one mother tongue) but more are being processed. The corpus contains learner variables similar to those mentioned in ICLE description. Participants are asked to fill in their name (but the data are later made anonymous), age, sex, nationality, native language, father's and mother's mother tongues, language spoken at home, information about the languages used as media of instructions at all stages of education, current education an information about stays in English-speaking countries. It also includes information about other languages spoken by the participant and basic information on the interviewer (sex, native language, foreign language(s) and relation with learner). The learners are interviewed by an interviewer (there can be one or several for the national subcorpora, in the Czech subcorpus, there were two). There are three tasks: the first one is a discussion of a topic selected by the learner, the second is a free discussion, and in the last part the learner is asked to describe a picture.

The interviews are recorded and recordings transcribed according to the same conventions so the data are comparable across the whole corpus. The interview is preceded by a code indicating a learner by number and the country (CZ for the Czech Republic) and ends with a code marking the end. The three parts are also separated by specific codes (S is used for set topic, F for free discussion and P for picture description). All these codes are written as tags, e.g. <S> and </S>, and so are letters A and B marking the speaker turns. All the words that were not actually said by either learner or interviewer are marked in a similar way. These include sounds (laughter, coughing etc.), contextual comments (somebody enters the room), voice quality (for example when the speaker was speaking and laughing at the same time, or whispering etc.), foreign words or pronunciation (<foreign> </foreign>), unclear passages or passages where anonymisation was needed (such as passages containing the name of the interviewee). Phonetic features are not transcribed, the only exceptions are length (using a colon :) and strong forms of articles. No punctuation dividing clause and sentences is used and there are no capital letters marking the beginning of a sentence as well. Dots are used to mark pauses and their number indicates the length of a pause (one dot for a short pause < 1s, two dots for a medium pause 1–3 seconds and three dots for a long pause >3 seconds). Filled pauses are transcribed in brackets (e.g. (ehm), (mm), (erm) etc.). To record other important features of spoken language, overlaps are marked by tags that mark the beginning of an overlap in both turns but the end of an overlap is not marked. Another feature of spoken language recorded is the false starts. Only the actually pronounced part of a word is transcribed and followed by an equals sign (e.g. rep= repetition...).

To summarize, LINDSEI contains data that have been collected at several universities and are treated in the same way so the whole corpus is comparable across different mother tongues. The comparable corpus of English native speakers LOCNEC is also available which makes LINDSEI a very useful source for studying spoken interlanguage of English language learners with diverse mother-tongues background.

4.1.2. Learner Corpus Analysis

There are two ways in which data from learner corpus can be analyzed, each of them based on a different theoretical approach to learner language: Contrastive Interlanguage Analysis (CIA) and Computer-aided Error Analysis (CEA). CIA is based on the approach that sees learner language as an imperfect version of target language. CEA is based on the assumption that learner language should be studied on its own (see Chapter 3). However, the use of these two methods in learner corpus research is not necessarily connected with these assumptions because for example the use of CIA can show not only imperfections that the learner has to correct but also features that can be shared by various interlanguages and as such can help with explaining how languages are learnt. These methods can thus usefully complement each other.

According to Granger (2002: 8–10), CIA involves two types of comparison. Researchers can compare learner language with target language and thus show where these two differ. This approach has several advantages but probably the most important one is the fact that researchers do not focus on errors only, they can also detect overuse or underuse of some linguistic features in learner language. Since one of the most important applications of learner corpus research lies in improving the teaching process, CIA may show the areas in which the improvement is needed. The second type of comparison is between speakers of different mother tongues learning the same language. Firstly, this approach can show the interference of a mother tongue when only one language group shares a specific feature that is not typical of native speakers or is considered to be a mistake by native speakers. Secondly, it can also show developmental patterns in language learning when there are features that are atypical of native speakers but all groups of non-native speakers share them. In this way, it can also help researchers to understand the nature of L2 acquisition.

The second way of analyzing data from learner corpora is CEA which is based on the error analysis from the 1970s. However, it is different from error analysis because, according to Granger (2002: 10): "[the studies] are computer-aided and involve a higher degree of standardization and, even more importantly perhaps, because errors are presented in the full context of the text, alongside non-erroneous forms." In this way, CEA solves most of the problems that error analysis was criticized for but an error classification system must be still developed in order to analyze learner language.

CEA has two possible methods of data analysis (Granger, 2002). Firstly, a potentially problematic feature can be selected in advance and a corpus is searched for this particular feature only (e.g. Loke et al., 2013). This can be sometimes useful but it involves several problems connected with the feature selection. A researcher has to make an assumption about what would be problematic for the learners (although usually based on experience and a preliminary analysis of the data) and can thus miss problematic areas that he/she does not expect and that would be equally interesting. The second method, error tagging, is much more time-consuming because it requires a number of steps to be completed but once completed, it can reveal much more about learner language. As Granger says (2002: 10), a learner corpus can be tagged either for a selected set of features or all the errors made by learners can be tagged. It is a time-consuming process but once the corpus is tagged for errors, it can be used in various ways and it consequently saves time because researchers can search for a concrete type of error just with the error tags. This is more advantageous than the first method mentioned because researchers do not have to come up with complicated queries using CQL (contextual query language) and can simply search for an error category using an error tag. A learner corpus can be also tagged for parts of speech (POS tagging) which can be helpful in some aspects. However, POS tagging has not been used widely in studying learner language because in order to search for a specific word, POS tags are not needed and searching for a part of speech is usually too unrestricted to discover something typical of learner language. Therefore, there have been only several studies using POS tagging to study categories that are problematic for learners.

4.1.2.1. Error Tagging

To be able to tag a learner corpus for errors, a set of error tags using some sort of error classification system has to be developed first. Researchers compiling a learner corpus usually develop a system of their own so the error-tagging systems differ considerably among individual learner corpora. This is sometimes criticized, Díaz-Negrillo and Fernández-Domínguez (2006: 86) claim: "one aspect that current EA [error analysis] is said to be in need of further work is standardisation of error typologies. Unlike other areas where more standardisation might exist, such as learner corpus design, corpus researchers have yet to agree on a general scheme of error annotation." It is of course valid criticism since the results of studies based on investigating different error-tagged learner corpora may not be easily comparable. However, the differences between error-tagging systems stems from the difficulty of defining an error. Looking at the POS tagging, the information about individual parts of speech is easily available from dictionaries and grammar books so by using these resources, it is quite easy to define a part of speech based on the dictionary information (word classes that a particular lexical item can belong to) and on the position in a sentence (syntactic information). The same does not apply to the definition of an error (see 3.2) and so the systems of error-tagging will necessarily differ in some aspects.

However, there are some features that are necessary for an error-tagging system in order to be effective. Granger (2003: 467) provides four characteristics that a system should have:

1) It should be informative but manageable.

2) It should be reusable (meaning that it should be possible to use it also for different languages).

3) It should be flexible (Granger uses the word flexible to describe that tags should be easy to add or delete in the annotation stage and should be easily retrievable later).

4) It should be consistent (i.e. an error tagging manual should be provided to prevent inconsistencies between different annotators).

An ideal error tagging system should be as close to these characteristics as possible and it should also be easy to expand when a new feature appears in learner language or the system needs to be adapted for another language with different linguistics categories etc. Such a system could be described as flexible or expandable which is implied in the second characteristic by Granger (2003) because tagging errors in another language would probably include adding some linguistic categories (of course depending on the language the system was developed for and on the language the system is being adapted for). It would be probably better to use the word flexible for this characteristic instead of the meaning described by Granger (2003) because what she describes is mostly connected with the annotation process, not with the tagging system itself.

As explained above, error-tagging systems vary. Since many of them are intended for the internal use of researchers tagging a particular corpus only and are thus not publicly available, the claim about the differences between different error-tagging systems is based mainly on the description of the systems in Díaz-Negrillo and Fernández-Domínguez (2006) and also on the information obtainable from the studies describing error-tagging system for four different learner corpora (Nicholls, 2003; Izumi et al., 2005; Granger, 2003; and Dagneaux et al., 2008). According to Díaz-Negrillo and Fernández-Domínguez (2006: 87), there are 12 learner corpora associated with error-tagging systems. Although they vary in the way the tags are coded, corrections made and in the amount of information that is included, they share some features as well. Most of the error tagging systems are based on some sort of linguistic classification although the level at which this classification is applied differs (for example the Louvain tagging system tags start with domains such as grammatical or lexical while the tags in the tagging system for the Cambridge Learner Corpus (CLC) start with a type of error – e.g. omission or a wrong form used). Most of the systems described divide the errors according to a word class, although on different levels of error classification and they also work with a different number of word classes. Lastly, most of the systems also include a correct form, usually inserted after the error.

Louvain tagging system is the basis for the system proposed in this thesis and will be thus described in detail. To illustrate how it differs from other error-tagging systems, CLC, *FreeText Project* (partly developed in Louvain) and the corpus of the National Institute of Information and Communication Technology – Japanese Learner English (NICT JLE), these will be described and compared with the Louvain error-tagging system. The last one mentioned is a spoken learner corpus but the error-tagging system does not reflect differences between spoken and written language, it contains only one tag which marks unintelligible utterances. However, the transcription process include adding discourse tags (see below).

Louvain tagging system is incremental (meaning it is not restricted to a certain number of levels) and individual tags contain several levels of information which is organized hierarchically – each tag has several positions that narrow down the error that is marked. This is similar to POS tagging but in most POS tagging systems, one letter in a tag means one position while in Louvain tagging system, some distinctions are marked by more than one letter (e.g. ADJ for adjective). This does not pose a problem at the current stage of the Louvain system development, however, using only one letter for one position in a tag could be more transparent than using more than one letter and, most importantly, useful for a computer analysis, especially searching the corpus with the use of CQL. All three systems mentioned above work similarly and individual categories are sometimes described with more than one letter. This could be caused by the relatively low level of complexity of the error-tagging systems used, most of which need only two or three positions in a tag. Only the tags used in NICT JLE corpus are based on XML (extensible markup language) so the structure of the tags is easily identifiable for computer analysis.

The Louvain error-tagging system uses only one tag[4] which contains all the information about the error (e.g. (LS) hospitalize $put up$). This is probably the best solution because in case of uncertainty about the right tag, it allows the use of more than one tag which immediately implies that there are more possible interpretations of the error. The error tagging system for NICT JLE corpus works similarly, using only one tag for one error, the tag containing all the information (e.g. <v_lxc crr= "put up">hospitalize</v_lxc>). However, this system is not very detailed and the tags described in Izumi et al. (2005) have all only two levels so there is no need to divide them into several tags. The error-tagging system for CLC uses usually also only one tag with two positions but it also uses embedded tags so when the wrong lexical item is used and this item is wrongly spelled, it is first tagged for wrong spelling and than for the wrong lexical item used (e.g. <#RV><#S>hospitalize| hospitalise</#S>|put up</#RV>; Nicholls, 2003: 575). In *FreeText Project*, tags are combined to provide all levels of information; thus the tag

[4] A tag is defined as a string of letters marked by brackets (their type differs in individual systems) for the purposes of this thesis. Thus (GVV) meaning Grammar Verb Voice in the Louvain system is one tag (Dagneaux et al., 2008: 25). Similarly <#RV>word</#RV> in CLC stands for R – word or phrase needs replacing, V defines a word class (verb) (Nicholls, 2003) . In *FreeText Project*, the tags are organized according to the level of classification but each level is represented by a separate tag.

marking an error in word order will be <X> <ORD> (Granger, 2003: 478) where X marks a syntactic domain and ORD marks that it is an error in word order.

As for the classification systems, three out of the four systems discussed start with a linguistic classification of errors, the Louvain error-tagging system is very much similar to the *FreeText Project* which is logical considering their origin (although *FreeText Project* is based in Louvain only partly). The Louvain system divides errors into 8 categories: formal errors, grammatical errors, lexico-grammatical errors, lexical errors, word order errors and words missing/redundant, punctuation errors, stylistic errors, infelicities (Dagneaux et al., 2008: 4–5). The FreeText tagging system is very similar, it divides errors into 9 categories: form, morphology, grammar, lexis, syntax, register, style, punctuation, typo (Granger, 2003: 468). The Louvain tagging system than specifies the category further and, being incremental, the specification can be as detailed as necessary. Most of the tags in the Louvain tagging system have three positions (e.g. GADJCS – Grammar, ADJective, Comparative/Superlative; this example illustrates also the unrestricted number of letters for individual positions). Only two tags have four positions. Nevertheless, the system can be easily expanded and adapted for different types of learner corpora and that is the most important advantage of this system when compared to other systems of error-tagging.

The error-tagging system for the NICT JLE corpus is slightly different because the classification of errors is based mainly on word classes. The errors are divided into 12 categories: noun, verb, modal verb, adjective, adverb, preposition, article, pronoun, conjunction, relative pronoun, interrogative and others (Izumi et al., 2005: 76). Since the tags have only two positions (more exactly three since the correct form is a part of the tag as well), specifications of word classes are presented as separate categories in the first position of a tag. Such a classification system is less useful than the Louvain incremental system because when more categories are added, the system can easily become confusing for the user. The Louvain tagging system would simply add a position specifying a subcategory such as pronoun – personal on the following position while NICT JLE corpus needs to create a new category in the first position which is than specified and thus the number of possible categories in the first position would increase too much. The second position in the tag specifies what type of error it is, e.g. in inflection or tense for verbs etc. The last category of errors, others, is a category for errors that do not fit anywhere else. There are errors such as "Japanese English" or "misordering of words" which are the inevitable result of the classification according to word classes and also a particular weakness of this system in that it mixes errors types with sources of errors. The system also contains a category of "unknown type errors" which could be useful but probably more so when used alongside other categories as a category of certainty[5], expressing annotator's doubt about the type of error tagged and added to an otherwise complete tag. Izumi et al. (2005) do not provide any example of this type of error so it is difficult to imagine what would fall into this category since their tagging system requires a correct form as a part of the tag. Overall, this classification system is a nice example of the problems connected with the decision to base the whole system on word classes. Izumi et al. (2005: 79) claim that they intend to develop their tagging system further: mainly to add a linguistic level at which the error occurs to add information on the gravity of error (whether it interferes with understanding) and also to differentiate errors from "unnatural" expressions.

Apart from error annotation, the NICT JLE corpus uses discourse tags to encode important information in the transcriptions of recordings. Similarly to LINDSEI, there are various features of spoken language recorded but, unlike LINDSEI, it is tagged for errors. In LINDSEI, some of the features specific for spoken language are given in round brackets (e.g. filled pauses), some are not marked in brackets at all (unfilled pauses marked by full stops) and there are pointy brackets used for marking foreign words, turns in conversation (for both foreign words and turns, a tag marking the beginning and a tag marking the end are used: <foreing>Liberec</foreing>) or sounds such as sighs (some of them can be marked by only one tag, cf. <laughs> versus <starts laughing> </stopt laughing>) etc. This list shows that marking these features is not very systematic in LINDSEI. The NICT JLE corpus, on the other hand, uses standardized way of coding them. All the tags are marked by pointy brackets, their beginning and ending is clearly marked as well (using <tag> for the beginning and </tag> for the ending). The system of discourse tags contains tags for marking filled pauses, repetitions, self-corrections, incomplete utterances, non-verbal sounds, utterance with a laugh, unclear utterance or use of Japanese words etc. (Izumi et al., 2004: 34). The description of this system is given in this section as a basis for the system proposed in the research part of the thesis because it should combine features specific for spoken language with learner errors.

[5] This category is not used in any of the systems described here but proposed later in the research part.

The CLC error-tagging system is the one that differs greatly from the other three systems described above because it does not start the error classification with linguistic categories (Nicholls, 2003). The tags in CLC have two positions, the first describing a general type of error, the second describing a word class. However, these two positions are not necessarily hierarchical as in the Louvain error-tagging system because there are never more than two positions in a tag and the order can be easily switched without any consequences. The first position contains tags describing "wrong form used, something missing, word or phrase needs replacing, word or phrase is unnecessary (redundant) and word is wrongly derived" (Nicholls, 2003: 573–574). Besides these general types of error, countability and agreement errors can also occur in the first position. The second position is used for coding a word class (pronoun – anaphoric, conjunction, determiner, adjective, noun, quantifier, preposition, verb, adverb; besides word classes, punctuation can occur on the second position as well). Apart from this classification of errors, the CLC error-tagging system also contains a set of special tags used for coding spelling errors, American spelling used instead of British, idiom and collocation errors, incorrect word order or wrong tense of verb or inappropriate register. Overall, the CLC error-tagging system is easy to use and can mark some of the errors made by learners precisely but since not all the errors fit into the categories described by the two position tags, there is a need to devise other, "special" categories.

To conclude, all the systems described in this part have both advantages and disadvantages. The greatest advantage of the Louvain error-tagging system seems to be its flexibility. It is incremental and, although the classification of errors is at some levels problematic, it allows addition of potentially infinite number of specifications and is thus very easy to use and to adapt for other research purposes. This classification system is, despite its limitations, still the most sophisticated one. The other systems include categories with errors that could be easily subsumed under some other, not "special" category (e.g. wrong tense of verb in CLC). The greatest weakness of the FreeText Project system is the fact that every category is expressed by a separate tag, otherwise, it is very similar to the Louvain error-tagging system. The error-tagging system for the NICT JLE corpus consists of a quite limited set of tags. The correction of the error is included in the tag which can be useful when looking for possible errors in a particular lexical item. However, the same is true for corrections in the Louvain system (inserted as $correction$) so the insertion of this information into a tag seems unnecessary. The CLC error-tagging system uses too many special tags which could be included into an error taxonomy that have tags with more than two positions.

5. Spoken Language

Spoken language differs from written language in many aspects, however, it was not studied extensively until relatively recently because the technology allowing easier study of spoken language is relatively new. More precisely, some aspects of spoken language were studied quite early, mainly pronunciation and prosody of spoken language, because to study these features, a limited amount of data is needed and can be thus made without the use of recording technologies (even though corpus research can provide a new perspective). However, a corpus-based research of spoken language is a relatively new field of linguistics and spoken corpora develop much slower because of the time-consuming character of processing data for such corpora. Moreover, written language was taken as a norm for quite a long time and features specific for spoken language were not described in grammar books in greater detail. As Carter and McCarthy (2006: 167) put it: "[t]he term 'standard grammar' is most typically associated with written language, and is usually considered to be characteristic of recurrent usage of adult, educated native speakers of a language."

However, since the assumption that spoken language is the same as written language and does not need to be studied separately is no longer valid and researchers have technological options that allow them to study spoken language more easily, there have been many studies focusing on the nature of spoken language and many spoken language corpora have been built (see 5.2). Based on the corpus research, description of spoken English is provided in two important grammar books: Biber et al. (1999) base their description on the *Longman Spoken and Written English Corpus* and the description by Carter and McCarthy (2006) is based on the *Cambridge International Corpus* (CIC). Both these descriptions will be used as the basis for the description of features specific for spoken language.

5.1. The Specific Features of Spoken Language

Generally, spoken language happens in real time and is usually not prepared or planned. It is usually used in some sort of interaction and, therefore, happens face to face. A dialogue is considered to be a prototypical case of spoken language although it may contain a large portion of monologues (Halliday, 1989: 46). Biber et al. (1999) describe conversation which is in its nature also a dialogue. Being a dialogue, spoken language is used in a shared context and, as Carter and McCarthy (2006: 164) claim, it "reflects the immediate social and interpersonal situation". Therefore, deictic expressions and hedging are important means of expressing these relations. Taking as an ideal example of spoken language conversation, it is important to note that it is interactional and some of the features that are considered to be typical of spoken language are a direct consequence of this fact. Considering, for example, turn taking in conversation, filled pauses can be seen as a device used by speakers to indicate that their turn has not ended yet. Since conversation is usually not prepared and speakers react to each other, repetitions and self-corrections occur as well. Although there are features specific for spoken language only, Carter and McCarthy (2006: 164) emphasise that spoken and written language are not separate entities, they form a continuum. Moreover, spoken language is quite difficult to write down because sentences, the units used in written language, are not easily identifiable in spoken language[6]. Some of the features of spoken language do not have any equivalent way of transcription and, therefore, new means of expressing these features in writing need to be developed.

In spoken interactions, the speaker cannot think too much ahead and can be interrupted, so there are strategies employed to solve this limitation and also features that are a consequence of it. The lexical structure of spoken language is usually much simpler than that of written language; Biber et al. (1999: 1044) claim that "conversation has a strikingly low lexical density" and Carter and McCarthy (2006: 169) stress this feature when they talk about simple phrasal structure. Similarly, the sentence structure is different than in written language. Two features typical of spoken language, according to Biber et al. (1999: 1052), are dysfluency and errors. They describe them as performance phenomena and see dysfluency as hesitations mainly.

Hesitations in spoken language can be expressed in several ways. First of them is by pauses that can be either silent (unfilled) or filled. Unfilled pauses are more frequent than filled ones (Biber et al., 1999: 1054). They appear quite logically at important syntactic boundaries but they also occur in the places where the speaker is not sure how to continue and may occur when the speaker corrects himself/herself. According to Carter and McCarthy (2006: 172), long unfilled pauses can be perceived as problematic by other participants although he does not call them errors. Biber et al. (1999: 1054) claim that

[6] Term *sentence* will be used for the description of spoken language structure because of the need of comparing it with written language.

unfilled pauses occur at major transition points. Filled pauses, on the other hand, tend to occur at places where the speaker wants to signal that he/she has not finished yet. They are used at major planning points. They can be also used when the topic has changed or an important word is used because they are used as time-gaining devices. Secondly, repeats occur in spoken language, they are another way of buying time for thought and are not considered to be a sign of "sloppy or lazy performance" (Carter and McCarthy, 2006: 173). Biber et al. (1999: 1055–1062) focus on different types of repeats in the corpus, showing that single words are repeated most often and number of repetitions decreases with the number of words. Their research also shows that the same applies to the words with high frequency which are repeated most often and that the number of repetitions depends on grammatical category of a word, pronouns being repeated most often. Repeats are sometimes called false starts together with reformulations. Biber et al. (1999) refer to reformulations as retrace-and-repair sequences while Carter and McCarthy (2006) describe them as recasts. They occur when the speaker goes back and reformulates something he/she has already said. Unlike repeats, reformulations are often accompanied by other types of dysfluencies such as filled pauses because the speaker has to think about the way of reformulating what has just been said, which might take some time. The part of utterance that is repeated by the speaker remains grammatically incomplete. According to Biber et al. (1999: 1063–1064), there are other examples in which the utterance remains grammatically incomplete but these are not used to mark hesitation: a speaker is interrupted or corrected by the interlocutor. Reformulations are important for learner language research because they can show linguistic areas where the learner is not sure yet.

Besides hesitations in spoken language, the fact that language is planned in a particular moment can be illustrated by the sentence structure that is not the same as in written language. First of all, there are sentences inserted into sentences, described as "parenthetical structures" by Biber et al. (1999: 1067). They are inserted into another sentence but are not integrated and can be easily omitted without a change in meaning. Similarly to reformulations, a sentence can start with one structure and continue with a structure that is really not connected to the first one. Carter and McCarthy (2006: 171) call it "clausal blends".

Apart from these, general tendencies in structuring sentences or clausal units in spoken language are observed by both Carter and McCarthy (2006) and Biber et al. (1999). Biber et al. (1999: 1072) divide clausal unit/units into three parts. The main part is the body of the speaker's message which can contain any number of clauses (clausal units in words of Biber et al. 1999). This part is similar to written language although it can be structured syntactically in the way described above. It can be preceded by a preface which is a type of utterance launcher typical of spoken language (others being fronting, discourse markers and overtures). Biber et al. give only noun phrase prefaces as an example. Unlike fronting, they are often used with a co-referential pronoun that is a part of the body of the message (e.g. Anna's parents, do you think they are coming?). The third part described by Biber et al. (1999) is a tag. Tags are described as "afterthoughts to a grammatical unit, especially a clausal unit" and "a retrospective qualifications loosely attached to the preceding clausal material" (Biber et al., 1999: 1080–1081). They can be divided into several categories: retrospective comment clauses, retrospective vagueness hedges, question tags, noun phrase tags, other non-clausal units retrospectively added, self-supplied answers and vocatives. The number of tags is not limited to one for one clausal unit.

Carter and McCarthy (2006) introduce similar categories, they talk about headers and tails. Headers are described as "a particular type of structure [...] where an item within the clause structure is placed before the clause and repeated (usually as a pronoun) in the clause itself" (Carter and McCarthy, 2006: 193). The authors contrast them with other types of fronting where the elements still remain in the clause structure. This corresponds with the distinction provided by Biber et al. (1999) where utterance launchers are further subdivided and only noun phrase prefaces correspond to headers. The structures corresponding to tags in Biber et al. (1999) are called tails in Carter and McCarthy (2006). "Tails are typically noun phrases. They clarify or make explicit something in the main clause. Most commonly a tail consists of a full noun phrase which clarifies or repeats the referent of a pronoun in the clause that comes before it" (Carter and McCarthy, 2006: 194). This illustrates that their definition of tails does not fully overlap with tags described by Biber et al. (1999) because Carter and McCarthy describe tags as a separate category and tails have a structure similar to headers.

To conclude, spoken language shares some structural features with written language. Given the nature of planning in spoken language, there are, however, necessarily features that are typical of spoken language only. They include different structure of sentences, both in the main message and in the structures preceding and following the message itself; and also means for expressing hesitation and gaining more time to think about what the speaker wants to say and how he would say it (pauses, repetitions and reformulations).

5.2.Spoken Corpora

Spoken corpora have made the study of spoken language much easier. The first corpus of spoken English, *The London-Lund Corpus of Spoken English* was based on two projects: the *Survey of English Usage* by University College London and the *Survey of Spoken English* conducted at Lund University. The data had been collected since 1959 and the corpus was released in 1990[7]. In the 1980s and 1990s, some of the large corpora of English added a section of spoken language as well, e.g. BNC, COBUILD corpus or CIC (Luzón et al., 2007: 4). These corpora contain a large amount of data that was recorded in various situations and locations and they provide enough data for various types of analyses, including analyses of differences between dialects, registers and also for the study of variables such as gender, age or social background in spoken language.

Since the 1990s, the compilation of spoken corpora have become easier even though still quite time-consuming and various corpora for various purposes have been compiled. There are a lot of specialized spoken corpora nowadays, including corpora of learner language (see 4.1), academic language (e.g. MICASE – *Michigan Corpus of Academic Spoken English* and BASE – *British Academic Spoken English*) or dialect corpora such as FRED (*Freiburg English Dialect Corpus*). Generally, spoken corpora can be divided into several categories. Firstly, there are corpora focusing on varieties of English which contain either one (corpora of British, American or Australian English) or several varieties of English (ICE – *the International Corpus of English*, a corpus of both spoken and written languages that contains several varieties of spoken English, including such varieties as Singaporean or Nigerian English). Besides varieties, there are also dialectal corpora such as FRED but spoken dialectal varieties are most often a part of larger corpora such as BNC. Diachronic corpora of spoken English are not so common, probably the only example being *The Diachronic Corpus of Present-Day Spoken English* which uses the data from the *Survey of English Usage*. Lastly, there are also corpora of non-native speakers, apart from learner corpora, corpora of English as a lingua franca (ELF) are compiled (VOICE – *Vienna Oxford International Corpus of English* and ELFA – *English as a Lingua Franca in Academic Settings*).

In general, spoken corpora can be used for various research objectives and the research is limited only by the method of data transcription and by the variables recorded in a

[7] http://www.helsinki.fi/varieng/CoRD/corpora/LLC/

particular spoken corpus. Transcription practices vary but usually, pronunciation and prosodic features are not recorded in spoken corpora, these features are recorded only in specialised corpora for phonetic research. The features typical of spoken language described in 5.1 are usually recorded in transcription (although practices vary and pauses are sometimes measured and an exact duration is given and sometimes only short, long and medium pauses are distinguished). Similarly, most transcriptions include extralinguistic information such as arrivals of new participants, interruptions of the conversation etc.; and also the information about the voice quality (laughter or whispering) and non-verbal features such as smiling, pointing at something etc. Sentence boundaries are transcribed in two ways: they are either marked by normal punctuation although pauses etc. are included as well, or there is no punctuation because using it would mean interpreting the data in some way. For more information on the data transcription in LINDSEI, see 4.1.1.1 and also 4.1.2.1 for a brief overview of the transcription practice in the NICT JLE corpus.

Based on the type of data recorded and on the transcription methods used, various kinds or research can be conducted (for more information, see Luzón et al., 2007). Most of the spoken language corpora (mainly the large corpora that include spoken language as well, e.g. BNC) are POS-tagged which simplifies the search for certain linguistic features. Overall, spoken corpus research is an important field of linguistics, with more corpora being built and new methods being used for data recording and transcription (including aligning a recording with its transcription).

6. Material and Method

The research part of this thesis is the analysis of the spoken English learner corpus called LINDSEI, more specifically its Czech part (English produced by Czech native speakers). The analysis is based on the hypothesis that there will be differences that distinguish a spoken learner corpus from a written learner corpus and these differences will have to be taken into account when developing or adapting a tagging system for such a corpus. Therefore, the aim of this thesis is to look at features specific for spoken English and propose necessary changes in the error-tagging system used in Louvain (Dagneaux et al., 2008) that would reflect these features (as described in 5.1). It will also briefly mention features typical of both written and spoken language that do not need any modifications.

Transcriptions of most of the recordings in the Czech part of LINDSEI (46 out of 50) are already available so it is not necessary to transcribe them again. The Czech part of LINDSEI has been selected because the recordings are available as well and when there are some ambiguities and unclear points in the transcriptions, the recordings are used to solve them.

The first step of the analysis is an error annotation of the corpus. Firstly, the possible errors in the transcriptions and features that may be specific for spoken language are marked in the first 20 transcriptions (out of 50) and, in case of uncertainty, BNC is consulted. When the marking is completed, the errors or the features typical of spoken language are tagged (or tagged later when an appropriate tag was not available in the Louvain error-tagging manual). To ensure that the error-annotation is correct, the transcriptions are compared with tagged transcriptions which were error-tagged by another annotator. However, inter-annotator agreement is not calculated for the purposes of this thesis and error-tagging is not done by a native speaker of English. For the purposes of this thesis which aims at devising a system of tags more suitable for a spoken learner corpus, these steps were not necessary but when the whole corpus is being tagged, a further research of inter-annotator agreement will be needed.

The second step is the analysis itself: errors and features typical of spoken language[8] were analysed and divided into categories based on the *Louvain Error-Tagging Manual*

[8] The examples from the corpora differ in length but since there are no sentence boundaries in LINDSEI, examples are excerpted so that the errors would be clear and no standard length of an example is defined.

and on the description of spoken language provided in 5.1. 200 hundred examples were selected from the features marked in the error-annotation process in order to show the whole range of errors and other features investigated in the thesis. The scope of the thesis does not allow to present all the errors and spoken language specific features collected so the Appendix provides only the sample of 200 examples (and not a random sample because the aim of this thesis was not to investigate the distribution of the different categories but to show which errors and which specific features do occur in spoken language). A detailed description of this sample is then provided and necessary changes in the error-tagging system proposed.

7. Research Part

7.1. Error classification based on the Louvain error-tagging system

The main purpose of this thesis consists in identifying features which require special attention in the annotation of a spoken corpus. However, a brief description of some of the features shared by both spoken and written English is necessary as well. The examples of tagged errors from the spoken corpus are used to illustrate the taxonomy of errors developed by CELC in Louvain and also to show where this taxonomy could be expanded or changed in order to describe spoken learner language more precisely. The taxonomy of errors of the Louvain error-tagging system will be, however, changed only when the changes are required because of the nature of spoken language, the hierarchy of errors will not be altered because it is out of the scope of this thesis.

7.1.1. Form

The first category of errors devised by CELC are errors concerning the form of a word. They are further subdivided into morphological and spelling errors. Morphological errors are a feature that will necessarily occur both in spoken and written English because they include derivational (1) and inflectional (2) errors. Both types of errors can be expected to appear quite often but they appear only twice in the sample. They are not spoken language specific and as such will not be discussed in greater detail (the same approach will be adopted in the whole thesis).

1) has quite a **(eFM)** <u>pragmatical</u> **$pragmatic$** approach
2) I mean the youngest is . ten years old and the **(eFM)** <u>olders</u> **$older ones$** are . eleven

Spelling errors are a feature typical of written English, in a spoken corpus, they could be replaced by pronunciation errors. However, since only the pronunciation of the strong form of articles (transcribed as the[i:] and [ei]) is recorded in the corpus because they are variants that express emphasis and otherwise, pronunciation is not included in the transcriptions in LINDSEI, it cannot be investigated in this thesis because transcribing the recordings phonetically would exceed the time constraints of the current thesis. The pronunciation variant of articles will be subsumed under the category of articles discussed in Articles 7.1.2.3 and pronunciation will not be discussed further.

7.1.2. Grammar

Grammatical errors are the largest category of errors which is subdivided according to parts of speech. These errors occur both in written and in spoken language although their frequency may slightly differ and a subdivision that would reflect different features of spoken language may be needed in some cases. Together with wrong lexical choices, grammatical errors are the most frequent errors in the whole sample (not taking into account dysfluency discussed in 7.2.2.).

7.1.2.1. Nouns

Errors affecting nominal categories can be either number or case errors. Number errors occur several times in the sample. As shown by examples 3 and 4, they can be identified as errors without any doubt although their origin may differ, e.g. example 3 is probably caused by the interference of Czech because Czech speaker would say 'v neděli' in singular, for a repeated action. Example 4 is, on the other hand, probably just a slip of the tongue.

3) especially on **(eGNN)** <u>Sunday</u> **$Sundays$**

4) those were just . few words some . family members some **(eGNN)** <u>animal</u> **$animals$** . colours

Errors in noun case are less common, there is only one example (5) in the whole sample. In example 5, the correct form is a noun in nominative functioning as a modifier of the head of a noun phrase, not the *'s*-genitive used by the speaker.

5) your **(eGNC)** <u>bachelor's</u> **$bachelor$** thesis

None of the examples from the corpus is ambiguous and, therefore, there is no need to adapt the category for spoken language corpus.

7.1.2.2. Determiners

Determiners are another category that can pose a problem for language learners. The examples extracted from the corpus are errors in usage of both demonstrative and indefinite determiners, some of the indefinite determiners can be further subclassified as quantifiers (although the Louvain tagging system does not distinguish this category). Errors affecting demonstrative determines are illustrated by examples 6 and 7. Example 7 is a part of a description of a movie plot. There are more than two couples in the movie so correcting 'other' to 'the other' couple would not work in this context because it would mean that there are only two couples.

6) he is just so clever **(eGDD)** so **$such a$** clever guy
7) there is (eh) **(eGDD)** other **$another$** couple

Indefinite determiners are illustrated by examples 8 and 9. Example 8 is a quantifier error. Quantifier errors occur several times as can be seen in the Appendix 2. However, they are not a separate category in the Louvain tagging system and since they are not typical of spoken language, no changes are made in their classification. In example 9, the indefinite determiner is replaced by an indefinite article because the phrase 'nice haircut' is used to simply describe a picture and the use of 'some' would add emphasis where it is not intended by the speaker (for this example, the recording was examined as well).

8) it started **(eGDI)** few **$a few$** years ago
9) she is smiling . and: she has . har<?> (eh) hairdress (eh) her= hairstyle some haircut **(eGDI)** some **a** nice haircut

These four examples illustrate that there were no ambiguous instances of determiners in the sample that would require a separate category specific for spoken language, all the examples extracted from the corpus can be corrected as errors.

7.1.2.3. Articles

Since Czech does not express definiteness, articles tend to be problematic for Czech English learners and this tendency holds in the spoken corpus of advanced learners of English as well. The Louvain tagging system does not divide this grammatical category any further, however, based on the analysis of the data, I propose a classification that could be useful for error-tagging LINDSEI because it takes into account that some errors in article usage may be more significant for spoken English.

Majority of the examples analysed are clearly errors. These are represented by examples 10 and 11. The correct zero article in example 10 expresses generic reference. An indefinite article in example 11 expresses indefinite reference of the noun phrase 'better word'.

10) you can . spot in **(eGA)** the **0** . todays' magazines
11) for want of **(eGA)** the **a** better word

However, there are also instances where it is difficult to decide whether the article is used in a correct form and a correct place or not because of the structure of spoken language. In example 12, the incorrect use of an article is tagged as a clear error because although we may speculate that the speaker used it because he originally meant to finish with a noun phrase, there is no evidence for it. Apart from the indefinite article, the unit is a normal sentence. The same cannot be said about example 13 which is not tagged for article error because 'a quite' can be interpreted as a false start or reformulation. This interpretation is reinforced by the presence of the filled and unfilled pauses that follow the expression. However, since it is only an interpretation and example 13 can be theoretically tagged for both reformulation and article error, it would be probably useful to include the information about other possible interpretations of a certain feature into the tag itself. In example 13, therefore, probably the best solution is to mark it not as an error but as a feature specific for spoken language (discussed in 7.2.2 in greater detail) but at the same time add a category (or suffix) of uncertainty on the last position in the tag because similar problems are likely to occur again in other examples and a systematic way of marking them is thus useful for further analysis. The category would express the annotator's uncertainty about the status of such a form and also the possibility of multiple interpretations.[9] This type of information can be especially useful in spoken learner corpora because when the recordings either are not available or even listening to them does not provide one correct solution and the interpretation is still difficult, it would allow the search for features difficult to interpret and, consequently, allow researchers to interpret them later.

12) and it was in the dark and we couldn't . we: it was **(eGA)** a **0** really difficult because we almost missed the ship

13) that was **(sDRu)**[10] a quite (eh) . an . advantage for me

To mark that some of the expressions are difficult to categorize, the category of uncertainty is thus introduced, adding suffix *u* to the last position of a tag. The main function of this category is for an annotator to mark features that are not easily tagged so

[9] The latter at least until the possibility of multiple tags for one feature will be resolved because in the current system, multiple tags are used only for errors that need several steps to be corrected, similarly to the embedded tags in the CLC although in the CLC, all the correct forms are provided at every step of the correction process. (Nicholls, 2003).

[10] The use of lower case *s* in the first position of the tag is explained in 7.1.2.4

that the researchers could search for them specifically. To distinguish this category from the rest of the tag, since it does not categorize the error as such, a lower case *u* is used.

Looking at examples 14 and 15, there is an indefinite article used correctly but the form is incorrect. Both these examples are clearly errors but since they present a category that differs from a simple article error, it may be useful to include a further specification of the tag. Addition of a category that would show that the type of article is used correctly (definite or indefinite) but the form is incorrect seems to be the most plausible solution.

14) which has **(eGAF)** an . **a** (eh) strong (LS) impact $effect$ on my life

15) for me it's **(eGAF)** a **an** important part of the . of the movie

Similarly, the places where the pronunciation of an article is given could be marked as formal features as well (illustrated by examples 16 and 17), as mentioned in section 7.1.1. However, since these features can be searched for by simply using the square brackets and they have mainly the emphasizing function, it is not necessary to tag them specifically at this stage of the system development. Moreover, the transcription of the pronunciation of articles is not entirely unified in LINDSEI.

16) was a[ei] a[ei] experience also . very very powerful

17) falls in love with the with the[i:] oldest . daughter . Jane . and his friend Mr Darcy (eh) falls in love with . the[i:] . second . oldest . second oldest

7.1.2.4. Pronouns

Another category of grammatical errors are errors in the usage of pronouns. This category is further subdivided according to the type of pronoun that is used incorrectly. However, not all the pronoun categories were found in the sample analysed and so the description of this category will focus mainly on the analysis of personal pronouns which are typically the most frequent type of pronouns in spoken language, together with demonstrative pronouns (Biber et al., 1999: 1042). There are instances that are clearly not correct, see examples 18–20 where the type of error can be easily identified. In example 18, a singular pronoun 'it' is used to refer to a noun in plural. In example 19, the referent of the third person singular pronoun is a man, thus the co-referential pronoun is 'he', not 'it'.[11] Example 20 shows a different type of error, the personal pronoun is missing which is more

[11] However, this can be a feature typical of informal spoken language and further analysis is needed.

likely an omission of the obligatory indirect object of a ditransitive verb but that is not tagged as a separate category in the Louvain tagging system.

18) clothes and **(eGPP)** it's **$they are$** used very much
19) but if if **(eGPP)** it **he** was . an artist . then: he shouldn't have done it
20) could you tell us some: (ehm) give **(eGPP)** 0 **us** a tip for a . good . German TV show

Besides these clearly identifiable errors, there are examples that would be probably considered erroneous in written language but are entirely acceptable in spoken language (although the analysis of the data from LOCNEC confirming this assumption is yet to be done). Examples 21–23 illustrate this issue. Although all of them contain a personal pronoun that is not and probably should be co-referential with the preceding expression, it would be too strict to classify them as errors. In example 21, both pronouns refer to books mentioned earlier in the conversation that the interviewee has to read. It is possible that the second pronoun refers to 'reading' in general but a more plausible interpretation is that both personal pronouns have the same referent. Example 22 shows similar problem – 'it' can be co-referential with 'them' but it can also refer simply to music. Example 23 shows an instance where 'it' is used to refer to the whole situation and thus is not problematic at all.

21) some of them were . plays like drama . some of **(sGPP)** it was poems
22) I started listening to the Beatles my dad loved them and . so I liked **(sGPP)** it too so I listened to it as well
23) these girls are probably not very . (er) honest . honest people yeah that these are . quite (em) . let's say . <lip sound> (eh) <starts laughing> yeah <stops laughing> I wouldn't judge it yeah . they can

In order to mark the category needed for the description of examples 21 and 22, an addition of a category to the first position and, consequently, a shift of the other positions to the right, is proposed. A prefix is used to differentiate between features specific for spoken English and errors. To distinguish it from the other parts of the tag, it is written as *e* for errors and *s* for spoken language, in lower case letters because thus, it will be obvious that it is a category different from the other categories included in any tag in the system. The marking is similar to the one proposed for the uncertainty category. Therefore, examples 18–20 are tagged as eGPP and examples 21 and 22 as sGPP. Also, for the second category, no corrections are included because it is not an error.

Apart from personal pronouns, errors in other types of pronouns occur as well but they are not very numerous and they are not spoken language specific. The only error involving a possessive pronoun (an omission) is shown in example 24. The noun phrase in this example lacks a determiner but since the determiner is a pronoun, it is not tagged as eGDD.

24) I got **(eGPO)** 0 **my** bachelor (eXNPR) title

Example 25 illustrates the only error involving an indefinite pronoun found in the sample. Although the speaker uses 'any' correctly, he uses 'something' instead of 'anything' in the very same sentence.

25) we didn't have any mobile phones or **(eGPI)** something **$anything$** like that .

Reflexive pronouns are represented in the sample by only one error, shown in example 26. 'Them' cannot be used in this sentence because it would have a different referent than the noun 'people'.

26) people .. don't want to see **(eGPF)** them **$themselves$** as they are

Errors in the usage of relative pronouns are somewhat more frequent. Examples 27 and 28 show that the speakers have occasional problems with the difference between 'who' and 'which'. Example 29 presents problems for the analyst, it could be interpreted as a self correction because the speaker uses 'which' after 'children' but the rest of the example is not connected to the beginning syntactically. Therefore, the example is not a relative pronoun error but a self-correction (see 7.2.2.3)

27) authors **(eGPR)** which **who** are not really taught here very much
28) the actors . (eh) **(eGPR)** which **who** are really good
29) their children **(sDC)** which . you know you would think okay maybe there isn't a connection

7.1.2.5. Adverbs

Errors in the usage of adverbs are of two types – either a wrong adverb is used or there is an error in the position of the adverb. Examples 30 and 31 illustrate the latter, wrong position of 'also' being one of the most frequent errors in this category. Nevertheless, the possible modifications of this category are discussed in 7.1.5. Examples 32–33 illustrate the problem with deictic adverbs. Although they are put together with other incorrect uses of adverbs, distinguishing them by adding D for deixis in the last position should be useful because deixis is very important in spoken language (cf. Carter and McCarthy, 2006). However, there are not enough data to prove that it is a feature really typical of spoken language and, consequently, the category has not been added.

30) there was **(eGADVO)** a band playing also **$also a band playing$**
31) always went **(eGADVO)** a little back **$back a little$**
32) this city it's . it's London . (eh) I've <laughs> I've been **(eGADV)** here **$there$**
33) we were **(eGADV)** here **$there$**

7.1.2.6. Verbs

The last word class not yet discussed is verb. It is a complex category further subdivided according to the grammatical properties of the English verb. Examples 34–36 are errors in verbal morphology, in all of them, an incorrect form of the verb is used. In example 34, there are even several incorrect forms used and it is the only case where the speaker probably really did not know the correct form.

34) with her eyebrows roused (em) rised <overlap /> **(eGVM)** risen **$raised$**
35) I **(eGVM)** no study **$don't study$** English language
36) we: had **(eGVM)** went **$gone$** there

Although advanced learners of English definitely know that the suffix -s is added to a verb in the third person singular, there are several number errors, illustrated by examples 37 and 38.

37) was like five . five parts and this also . on= only **(eGVN)** have **has** . two . hours
38) where her problems . **(eGVN)** starts **$start$**

The auxiliary verb category is quite frequent in the data analysed. Learners tend to make errors both in the selection of an auxiliary verb in general (examples 39 and 40) and in the auxiliary verb used in the subordinate clause of conditional (examples 41 and 42). It would be useful to mark the errors in conditional separately because by marking them, it would be easier to find both the errors in auxiliaries and tense errors (as illustrated by example 42 where both the auxiliary and the verb in the subordinate clause are wrong). However, since they are not errors typical of spoken language, the change in the tagging system is not made in this thesis.

39) we **(eGVAUX)** should have read **$were supposed to read$**

40) you **(eGVAUX)** are able to **can** understand everything

41) if . (er) the woman . (er) **(eGVAUX)** would be **$were$** . (er) a really good friend of mine (er) . I think I would lie

42) if if I **(eGVT)** didn't have **$hadn't had$** this experience I would probably **(eGVAUX)** fire **$have fired$** it up

The error category that slightly differs when tagging spoken English is verbal tense. There are, similarly to other grammatical categories discussed so far, examples that are clearly incorrect. These can be illustrated by examples 43 and 44. In 43 there is a tense error in the indirect speech, example 44 is simply a tense error, the past simple tense is used instead of the present perfect tense (despite the fact that the speaker was repeating the instructions).

43) they actually asked the lady . whether we **(eGVT)** are coming **$were coming$** again someday

44) so my favourite . movie or . the movie I've . I **(eGVT)** saw **$have seen$** and I th= . I think that is really good

Apart from clear errors where nothing spoken language specific needs to be marked there are examples such as 45. The verb 'describe' is used in the present simple tense although it is probably a part of the indirect speech introduced by 'he told us that...'. This is one of the examples where it is difficult to decide whether it is an error or not because 'describe' can be either connected with 'loved' and then it should be in the past tense (two coordinated predicates), or it can be interpreted as a general statement about the description of the forest (something people always do) and then the present tense could be used. The first interpretation seems more plausible but the uncertainty category suffix is added to mark that there are two possible interpretation.

45) he told us that: . people loved the forest part but that they . also **(eGVTu)** describe **$described$** it as similar to the Amazon forest or something like that so

Putting aside the examples discussed above, there are examples in which there are tense inconsistencies across larger segments of the speech but which should not be corrected as errors. This is again a feature typical of spoken language. It is caused by the nature of spoken language which happens in real time and the speaker does not always feels the need to correct himself/herself. Similarly to inconsistencies in the use of personal pronouns, the use of prefix marking this as a feature specific for spoken language is recommended in cases illustrated by the examples below.

46) their kids **(sGVT)** got into a fight and one hurts the[i:] other . (em) and they start talking about this

47) well was doing his best but she wasn't satisfied she **(sGVT)** seems to be criticising her portrait . so she yeah she **(sGVT)** is very upset obviously <laughs> with something so: . maybe she asked him to: . try another one just second attempt and: . the second one . with better hair and which is more . feminine or more more fashionable I don't know . possibly . (eh) was all right for her so . then she: . she bought the picture and she invited her friends to see it

48) (uhu) (eh) maybe that here she doesn't like . she doesn't like the painting . so she she she **(sGVT)** told the painter to draw it . to draw her differently . and now when (eh) . he . (eh) . **(sGVT)** changed the the picture of her . the[i:] . her appearance she she's happy she's satisfied even though it's not really her so . <sniffles> it's the . hypocrisy and (erm) superficiality of of people . probably <laughs>

Examples 46–48 show that in spoken language, speakers tend to be inconsistent in the use of tenses when narrating something but since it is probably a feature typical of spoken language (not necessarily English only, at least in Czech, native speakers tend to switch tenses as well in longer narrations), it will not be tagged as an error. Example 46 is a part of conversation where the speaker narrates the plot of one of his favourite plays and apart from 'got', the speaker uses the historical present. Examples 47 and 48 are parts of longer descriptions of four pictures, the first one is mostly in the past tense which is the reason why the present tense is tagged (although it is not an error, it is probably specific for spoken language), the second one is predominantly in the present tense (historical present again) so the past tense is tagged as a feature specific for spoken language.

7.1.2.7. Word class

Last grammatical category included in the Louvain error-tagging system is the inappropriate use of a word class. Examples 49–51 of word class error provided here show that adjectives modifying a noun phrase appear to be a problem for language learners. However, at least example 51 can be also seen as an error affecting word order (or, more specifically, the position of an adverb) because it can be corrected as 'to speak German fluently and know...', without the change in word class and since the possibility of using two error tags and two corrections is still not established for the corpus, the error was marked as uncertain. Example 52 illustrates the use of an adjective instead of a noun. Although there may be some problems with distinguishing types of errors, there is nothing spoken language specific in the examples extracted from the corpus.

49) his mother . didn't speak very **(eGWC)** well **$good$** English as well but

50) when I was at home . I think for four months because (eh) . of the . health= **(eGWC)** healthy **$health$** reason

51) Germany . is (eh) . is (er) much closer to us so to: (mm) . to speak **(eGWCu)** fluently **$fluent$** . German and know (eGADV) a lot of $a lot$ about (eh) history

52) and I (eGVT) was (er) in $have been to$ (eh) **(eGWC)** German **$Germany$** twice

7.1.2.8. Summary

To summarize the similarities and differences in tagging grammatical features of learner corpora, only the categories introduced because of the specific features of spoken language will be briefly repeated. Based on the analysis of articles, pronouns and verbs, a prefix dividing certain features into two categories: error *e* and features specific for spoken language *s*, have been introduced. To distinguish it from other parts of a tag, it is written in lower case and is used as the basic distinction for the features and errors tagged in the spoken corpus. It is useful not only when errors and spoken language features need to be distinguished in grammar, it is used as an overall distinction of the two important domains tagged in the Czech part of LINDSEI. The features specific for spoken language are analysed in the second section of the research part.

The second important innovation in the error-tagging system based on the analysis of grammatical features in the sample is the introduction of a category that can be again added to virtually any tag and that expresses uncertainty about the tag assigned to a particular error. Ideally, there would be a clear definition of an error and every error identified in a learner corpus would be easily assigned a tag. However, since defining an error is not unproblematic and analysing spoken language only amplifies this difficulty (see 3.2 and 5.1), there is a need to somehow express that some of the tags may not have been assigned to an error or other feature if the transcription was annotated by another researcher. To enable the research of ambiguous features, the category of uncertainty (in the form of a suffix *u*) is added to some of the tags when necessary.

Otherwise, the categorization of errors follows the hierarchy defined in the Louvain error-tagging system and only minor changes are made in the system (addition of the category that specifies that the error or spoken language variation is only in the form of a article – F added to the GA tag).

7.1.3. Lexico–Grammar

The third category used in the Louvain error-tagging system is the category that contains errors where lexico-grammatical rules have been violated in some way. It contains complex errors. The complexity is understood as a combination of general grammatical rules violations and also of violations of morpho-syntactic properties of a certain lexical item. This category is further subdivided into several sections.

The first section contains complementation errors. Only two examples (53 and 54) were extracted from the corpus, both of which illustrate errors in the complementation of adjectives.

53) **(eXADJCO)** worth to say **$worth saying$**
54) here I am **(eXADJCO)** used to work **$used to working$**

The second subcategory includes dependent prepositions. It is similar to the first subcategory but there is a shift in perspective. The first subcategory focuses on the element that is being complemented while the second one focuses on the element that is used as complement of any part of speech. Examples of prepositional complementation of adjective (55), noun (56) and verb (57 and 58) have been retrieved from the corpus.

55) woman . was . (erm) . **(eXADJPR)** blind on **$blind in$** one eye
56) one of the **(eXNPR)** books . from **$books by$** . Stephen King
57) . she's **(eXVPR)** pointing to **$pointing at$** something
58) she could **(eXVPR)** boast with **$boast about$** . boast with it

The third subcategory that belongs into this section contains errors affecting countable and uncountable nouns. Quite surprisingly, only one example (59) was extracted from the sample, an uncountable noun used as countable with plural suffix -s.

59) waiting for the[i:] **(eXNUC)** outcomes **$outcome$**

None of the examples extracted from the corpus can be seen as spoken language feature rather than an error and thus no changes in classification are proposed.

7.1.4. Lexis

The category that includes errors concerning lexical choices is very similar to the category of lexico-grammatical items although the errors described there are less complex. However, they can occur in both written and spoken language and are fairly common. The data collected from the corpus confirm this assumption because all the examples below could as well have occurred in written language.

Firstly, there are wrong lexical choices that involve a single word (60, and 61 where the adjective is used in order to describe the appearance not the character of the girl). Errors in single words can be further specified if the wrong lexical choice is a false friend in Czech (62 'akce' or 63 'gymnázium').

60) **(eLS)** <u>cease</u> **$fade$**

61) not very **(eLS)** <u>nice</u> **$pretty$** girl

62) there are some **(eLSF)** <u>actions</u> **$special offers$**

63) during **(eLSF)** <u>gymnasium</u> **$grammar school$**

Wrong lexical choices in the case of whole phrases are classified in the same way (64–68). False friends occur less frequently among phrases than among single words in the sample but there are several examples of phrasal false friends: 'akční úterky' in 66, 'v ideálním případě' in 67, and 'na dobré cestě' in 68.

64) or **(eLP)** <u>just after school</u> **$freshly graduated$** . (em) teacher

65) . the German . (em) **(eLP)** <u>the German language</u> **$German$**

66) **(eLPF)** <u>action Tuesday</u> **$Tuesday's offers$**

67) **(eLPF)** <u>in the ideal case</u> **$ideally$** people . or the students should have . read a lot of books

68) that you are **(eLPF)** <u>on a good way</u> **$heading in the right direction$**

The last category of lexical errors are errors in connectors which include coordinating and subordinating conjunctions and logical connectors (category that corresponds to conjuncts in Quirk et al., 1985). There are several occurrences of this type of errors and all the examples collected are subordinating conjunctions (examples 69 and 70).

69) you should do whatever he wants . **(eLCS)** even when **$even if$** he wants to . paint a completely different picture

70) I think that . **(eLCS)** as **$once$** (eGVT) you're $you've been$ there for a longer and longer time . you get used to it

7.1.5. Word redundant, word missing and word order

The use of a word that is redundant in a sentence, omission of a word without which the sentence is not complete and errors in word order (other than wrong word order that include adjectives and adverbs, as described in 7.1.2.5) all occur in the corpus although tagging a word or a phrase as redundant is not so easy in spoken English because of its syntactic structure. There are only two examples of redundancy in the sample:

71) so (er) we . **(eWRS)** usually **0** . used to (er) fire up something . which we found (er) in the street

72) who invited a lady . to be painted to sit . (er) **(eWRM)** a model **0** for him

In example 71, 'usually' is redundant because the habitual character of the action is already expressed be the verb 'used to'. In example 72, there is a redundant nominal phrase because the fact that the woman was being painted and pose for the painter is already expressed by the verb phrase 'sit for somebody'.

The occurrence of missing words, on the other hand, seems to be quite a regular feature of spoken English. The most frequently omitted part of speech is the verb, here illustrated by examples 73–75 and in the Appendix. All the verb forms omitted in the sample were forms of the verb 'to be'. In example 75, the whole existential *there*-construction is missing. Omission of other word classes is not so common, for example in 76, a noun is missing.

73) . it might **(eWM)** 0 **be** his girlfriend or so

74) well . <laughs> this **(eWM)** 0 **is** really . (em) . childish perhaps

75) . I . don't . like the[i:] atmosphere which is . on at (eGA) the concerts 0 <overlap /> usually . because **(eWM)** 0 **$there are$** too many people .

76) and I'm on the other **(eWM)** 0 **$side$** of the . of the fence

Based on the sample investigated, a further analysis of words missing and redundant may be needed because there may be some tendencies and in such case, specifying the word class of a missing word can be useful for further analysis. Overall, no spoken language specific features have been found in these categories.

Probably the most complicated of these three categories is the category of the errors in word order (which are tagged either as eWO or eGADVO indicating that the problematic word is an adverb according to the Louvain system). These errors are especially difficult to analyze considering the syntactic structure of spoken language and dysfluencies typical of spoken language such as false starts or self-corrections. Violations of the standard word order principles are here defined as otherwise complete units that do not contain reformulations (described in 7.2.2.2).

Based on the analysis of the data from LINDSEI, two categories of word order errors or word order problems that need to be marked can be observed. First of all, there are deviations from the standard word order that can be identified as errors even in spoken language. In example 77, there is no possible interpretation that would justify the deviation. Similarly, example 78 cannot be interpreted as a deviation from the standard word order in order to emphasise a different part of the sentence because, unlike in the examples 79–81 below, 'really' cannot be used to intensify the noun in this example.

77) **(eWO)** here this in **$here in this$** seminar . there . are fifteen people so there's . discussion and so on

78) these **(eGADVO)**[12] really (er) . actions can **$actions can really$** affect our (eGNN) life $lives$

The second subcategory includes instances where adverb can be used as an intensifier and it can be intensifying different parts of the unit analysed. Thus, in example 79, 'really' could be used either to intensify the whole predicate part of the clausal unit or it can be used to intensify only the adjective which would be considered to be a standard word order. In order to decide which of these interpretations is correct, the recording was used for verification and consequently, the position of 'really' was marked as an error because there is no emphasis on this word when the speaker is talking. 'Really' in example 80 can be analysed in a similar way but even on the basis of the recording, it is difficult to decide whether it is an error or not. Sine seeing 'really' as an intensifier of the noun phrase 'that

[12] For the purposes of this thesis, the structure of Louvain error-tagging system was not modified so the examples concerning deviations from the standard word order are tagged as either wrong word order or wrong word order that includes adverb.

much experience' seems more plausible in this instance, the feature is not tagged at all. The same applies to example 81. However, further analysis of this subcategory is needed in order to determine whether there are any unclear instances where the use of the uncertainty suffix could be necessary.

79) the same time I **(eGADVO)** really was **$was really$** happy that I was finally there
80) I don't have really that much experience as my friends
81) yeah still I I really love Oscar Wilde

7.1.6. Infelicities

Infelicities are defined as units that do not contain any errors but are still considered to sound foreign and non-natural (examples 82 and 83). However, since they are not specific for spoken language, they will not be studied further.

82) and there is the . **(eZ)** this possibility is also available **$this is also possible$**
83) mother **(eZ)** had . (er) **$gave birth to$** my little sister only

7.1.7. Use of mother tongue

Although there are not many examples of the use of mother tongue in the data analyzed for the purposes of this thesis, the instances found in the sample show that some of them can be considered to be errors because a speaker automatically switches to his/her mother tongue when he/she is not sure and uses expressions that are not part of the target language. However, based on the examples 84, 85 and 86, it is difficult to decide whether this feature should be considered to be an error. Examples 84 and 85 contain interjections in Czech and this particular speaker makes a rather frequent use of them. Example 86 illustrates that even the inability to recall a word can be considered an error. In other instances, speakers are usually trying to find the word they are looking for or ask the interviewer but they do not switch into Czech.

84) <B> <foreign>**(eM)** jo **$yeah$** </foreign> airlines (eh) I don't remember I'm sorry
85) it . but: he (eGADVO) also . can $can also$ speak Hebrew he can speak Arabic he can speak <foreign> **(eM)** no **$well$** </foreign> English and stuff of course
86) Museum I (eGVT) I've seen $I saw$. (em) .. <foreign> **(eM)** sfinga **$sphinx$** </foreign> I'm not sure how to

Example 87 is used to illustrate why searching such instances only with the use of tags <foreign> is not so easy. Most of the instances where this tag is used are proper names of various origin, most often Czech but also names of places in foreign countries etc.

87) I went with a with a couple of my colleagues (eh) to <foreign> Čino= Činoherní klub </foreign>

Error-tagging of the use of the mother tongue in situations illustrated by examples 84–86 can be later used for the study of native language usage in non-native language situations. Therefore, the category M in the second position of a tag is introduced. It stands for mother tongue usage and the first position in all three examples is reserved for the prefix marking errors.

7.2. Features specific for spoken language

Apart from errors already described in the Louvain error-tagging system, there are other features typical of spoken language that can be tagged and analyzed. Some of the categories proposed in this part of the thesis are already analysed and described in detail in 7.1 because they were identified on the basis of detailed analysis of examples originally tagged (with some degree of uncertainty) as errors and also of examples that are in many aspects similar to them (but were not tagged as errors). Leaving aside the features already analyzed, several other categories are described in this part and all of them are again supported by examples extracted from the sample.

7.2.1. Grammar

The features typical of spoken language that can be classified in terms of grammatical categories are the ones already discussed. Nevertheless, a brief overview is given in this section. Only two subcategories of this category have been identified so far but the tagging system can be easily adapted when a new feature that is not really an error appears when the rest of the Czech part of LINDSEI is tagged. Since both the subcategories are similar to those of errors, almost no change in tags is made, only *s* is added to indicate that the tag does not describe an error.

The first feature is a spoken language specific usage of personal pronouns. There are examples (e.g. nos. 21, 22 and 23) where the reference is sometimes somewhat complicated and not always correct, different personal pronouns are used to refer to the same referent etc. Such instances are typical of spoken language, are tagged as such and are not considered to be errors in this thesis.

The second feature identified in the previous section are tense inconsistencies illustrated by examples 46–48. These examples show that in spoken language, there is variation in tense during longer narrations. Although such inconsistency would be probably corrected in writing, it does not appear unnatural in spoken language. Therefore, similarly to personal pronouns, a tag used for errors was adapted for this category as well.

7.2.2. Dysfluency

The term dysfluency is based on the description of grammar of conversation by Biber et al. (1999: 1066). It is used as a name for this section or domain for several reasons. Firstly, Biber et al. (1999: 1066) divide performance phenomena into errors and dysfluency. Since the aim of our tagging system is to take into account both errors and features specific for spoken language, adopting this term for most of the spoken language specific features seems to be appropriate. Secondly, the characteristics of spoken language described in this category are all more or less disrupting the fluency[13] of spoken language and this term seems to describe them all. Thirdly, adding letter D to the second position of a tag does not overlap with any other domain already established in this position.

Apart from the categories established below, there are several other features that could have been included in this category. Filled and unfilled pauses are, however, already part of the transcriptions (see example 88). Unfilled pauses are transcribed by full stops and filled pauses such as 'er' or 'em' are given in round brackets.

88) . (er) because (eh) it waas girls who a= who accompanied me . (eh) in the end .. (em) . there had to be plenty of water

Similarly, overlaps, sounds other than words and unintelligible utterances are already marked in the existing transcriptions.

7.2.2.1. Repetitions

Repetitions are one of the typical features of spoken language and they are very frequent in the sample analysed for the purposes of this thesis. Some of them are easily

[13] Fluency is not used as a linguistic term in this thesis, it simply refers to a subjective notion of uninterrupted flow of words in speech.

retrievable with software such as WordSmith but when a repetition is interrupted by a filled or unfilled pause or more than one word is repeated, searching the corpus for them is problematic. Repetitions are generally very numerous and most of them seem to be repetitions of a single word (examples 89, 90 and partly 91 – personal pronoun is repeated but then, the whole phrase 'they were' is repeated as well) which would confirm findings by Biber et al. (1999) discussed in 5.1. Repetitions of two words are also quite common in the corpus (examples 91 and 92).

89) also (eh) {one one} more (er) thing which was pretty important for me

90) . {we we} went by bus which is . a little annoying because it was a long long way .. but . it was definitely worth it (erm) {we we} went to

91) {they {they} were . they were} pink

92) a really good experience to . {think of think of} this novel

Considering the abundance of repetitions in spoken language, tagging spoken corpus for repetitions would result in introducing such a considerable quantity of tags that would render further work with the corpus almost unmanageable. Therefore, marking the repetitions with curly brackets is proposed since curly brackets have not been used for any other purpose in the corpus.

7.2.2.2. Reformulations

Although repetitions need not be assigned a special tag, the second category of dysfluencies are reformulations and for them, a set of categories for tagging will be proposed. Reformulation is defined as a feature of spoken language produced when speaker goes back to what he has already said and then reformulates it, thus creating one unfinished unit and one that continues in some way. For the purposes of this thesis, reformulations do not include corrections, therefore, neither the part that is being reformulated nor the reformulation itself contains an error of any kind. If they do, they are categorized as self-corrections.

Examples 93 and 94 illustrate what is taken as reformulations. Example 93 contain a long segment which is then reformulated. Example 94 illustrates that even a short unit in spoken language can be considered a reformulation.

93) which was very nice because **(sDR)** there were a lot of (erm) <lip sound> (er) . it was very international there were students from France Austria

94) = I was living in a family **(sDR)** so I had a I had keys\because I think that (er) . (erm) . unless you are not gonna work (er) in a . <lip sound> . (er) unless you are gonna work in a . in . <lip sound> . (er) let's say . have some job where you can . work with your English

Example 14 discussed in connection with articles ('that was **(sDRu)** a quite (eh) . an . advantage for me') is another possible example of reformulation but with uncertainty expressed by the suffix *u* in the last position of the tag.

7.2.2.3. Self-corrections

Self-corrections or corrections are not described as a separate category by Biber et al. (1999). For the purposes of tagging learner corpus, corrections are important and they differ from reformulations because they contain an error. Based on where the error occurs and what is being corrected, a further subcategorization is proposed and applied for self-correction tagging.

First of all, a wrong form can be corrected by the speaker. Examples 95–98 illustrate the first subcategory (marked as D – dysfluency, C – correction, C – Correct) where an error is corrected.

95) **(sDCC)** when we talking **when we were talking**

96) then it got **(sDCC)** worst **worse** because the teachers got worse

97) she wasn't really satisfied because . (er) for her . the portrait **(sDCC)** doesn't . (eh) **didn't** look . like her

98) I was afraid that I . (er) **(sDCC)** I'm going I was . I was afraid that **I was going** to hate it

Second category are the errors that are not corrected and very often a different type of error is made. Example 100 shows that there can be several errors made without reaching the correct form. Whether to use a separate tag for each correction is an issue for further discussion but since there are not so many similar examples, it is not so important at this stage how such cases will be tagged. To express that the result is again wrong, W is used to mark the wrong category in the fourth position of the tag.

99) I wish I had **(sDCW)** some . I wish I had (eGA) 0 a chance to . (er) work with English

100) she's looking with her eyebrows **(sDCW)** roused (em) rised <overlap /> (GVM) risen $raised$

The last type of self-corrections is an instance when the speaker starts with a correct form and in the attempt to correct himself/herself he/she actually makes an error. Examples 101 and 102 illustrate this type of dysfluency. The letter E for marking this category in the tag was chosen randomly.

101) I I am sorry I'm . **(sDCE)** my imagination . or (LSF) **fantasy** $imagination$

102) the movie **(sDCE)** I've . I (GVT) **saw** $have seen$ and I th= . I think that is really good . is . Pride and Prejudice

8. Conclusion

The analysis of the data extracted from the Czech part of LINDSEI shows that most of the error categories used for tagging ICLE can also be used when tagging a spoken learner corpus. Only several categories that cannot appear in spoken language are excluded. Spelling errors and punctuation errors are thus not part of the tagging system proposed. Similarly, errors concerning incomplete sentences are not included because the structure of spoken language, described in 5.1, cannot be described as incorrect because sentence is not a unit used to describe spoken language. The analysis confirms that there are features of spoken language that should be marked by special tags apart from errors because they are key features for description of spoken language and tagging them will make future research easier. Therefore, based on the analysis, several new categories are proposed.

First of all, the distinction between errors and features typical of spoken language is made. Since this division should be immediately obvious from a tag, this category is added as a prefix to the whole tag. This category is furthermore distinguished from the other positions of a tag by the use of lower case letters. Errors are marked by letter *e* and features specific for spoken language are marked by letter *s*. Apart from using *s* for newly introduced categories typical of spoken language only, it is also demonstrated (on examples that cannot be marked as errors in spoken language but would be corrected in written language) that some of the categories used for tagging errors in ICLE can be adopted for marking features specific for spoken language such as tense inconsistencies in narration etc. Introducing this category also enables researchers to potentially mark any category as spoken language specific when further research confirms such assumptions because the prefix can be added to any tag without restrictions.

The second new category proposed is the category of uncertainty. Based on the analysis, ambiguous examples were identified and since the current tagging system does not allow the use of more tags in such situations, adding suffix *u* at the end of a tag was proposed in instances where it is not clear whether something is or is not an error or does or does not belong into some category. It also enables marking the examples where more than one interpretation is possible. Even if a possibility of adding more than one tag was permitted, the category of uncertainty would distinguish places where more than one correction of an error is needed (and which can be marked by more than one tag even now) and where there is more than one interpretation (where currently only one interpretation has to be chosen and one tag added). This category is especially important when tagging spoken language because it allows researchers to mark features where it is difficult to tell whether they are errors or spoken language specific features and decide later after further analysis of data from a native language corpus such as LOCNEC is conducted.

Apart from the errors described in ICLE, a new category of errors concerning the use of the mother tongue in foreign language situations is added. Although foreign language features are already marked in LINDSEI, the majority of them are proper place or personal names and as such are of no interest for this kind of analysis. Therefore, marking a new category with the tag M is proposed in order to enable searching the corpus for such examples.

Lastly, new categories taking into account specific features of spoken language are proposed. There are categories that are based on the category of grammar used for marking errors in ICLE. The features that would not be considered errors in spoken language are still marked but instead of adding error prefix, they are marked with the prefix *s*. Apart from these, a new category of features specific for spoken language is introduced. Based on the term used by Biber et al. (1999: 1066), these features are subsumed under superordinate term dysfluencies and all of them are concerned with the structure of spoken language. The analysis shows that dividing reformulations into two categories is more suitable for learner language since corrections can show some interesting tendencies and areas where a speaker is not sure about the correct form of the chosen structure. Tagging self-corrections and reformulations separately can be also later used to compare the data from LINDSEI with the data from LOCNEC. Therefore, reformulations are defined as returning in speech and repeating a part of utterance in different words, without errors in either the part that is reformulated or the reformulation itself. This definition is used to differentiate between reformulations and self-corrections. Repetitions of the same word or the same phrase are described but are not included into the taxonomy of errors and features typical of spoken language. Instead, they are marked by curly brackets.

To conclude, the thesis has attempted to identify features of spoken language that appear in the learner corpus of spoken English and propose categories that should be tagged in addition to categories described in the Louvain error-tagging system. Based on the analysis of the data from LINDSEI, several such categories are described and modifications of the tagging system are proposed.

References

Biber, D. et al. (1999) *Longman Grammar of Spoken and Written English*. Harlow; New York: Pearson Education ESL.

Carter, R., & McCarthy, M. (2006) *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge: Cambridge University Press.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Corder, S. P. (1981a) *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

Corder, S. P. (1981b) 'The significance of learners' erorrs', in *Error Analysis and Interlanguage*, pp. 5–13. Oxford: Oxford University Press. (Reprinted from *IRAL,* 4 (1967): 161–170.)

Corder, S. P. (1981c) 'Idiosyncratic dialects and error analysis, in *Error Analysis and Interlanguage*, pp. 14–25. Oxford: Oxford University Press. (Reprinted from *IRAL,* 9 (1971): 147–160.)

Corder, S. P. (1981d) 'The study of interlanguage', in *Error Analysis and Interlanguage*, pp. 65–78. Oxford: Oxford University Press. (Reprinted from *Proceedings of the Fourth International Congress of Applied Linguistics* 2, by G. Nickel, ed., 1976, Stuttgart: HochschulVerlag.)

Costea, E. (2014, April 22) *The London-Lund Corpus of Spoken English (LLC).* Retrieved August 2, 2014, from http://www.helsinki.fi/varieng/CoRD/corpora/LLC/.

Dagneaux, E. et al. (2008) *The Louvain Error Tagging Manual*. Louvain: Centre for English Corpus Linguistics, Université catholique de Louvain.

De Cock, S. et al. (2011) 'Putting corpora to good uses: A guided tour', in S. De Cock et al. (eds.) *A Taste for Corpora*, pp. 1–6. Amsterdam: John Benjamins.

Díaz-Negrillo, A. and J. Férnandez-Domínguez (2006) 'Error tagging systems for learner corpora'. *Revista Española de Linguística Aplicada* 19, 83–102.

Dumont, A. (2014, April 8) *Learner corpora around the world*. Retrieved August 10, 2014, from http://www.uclouvain.be/en-cecl-lcworld.html.

Dušková, L. et al. (2006) *Mluvnice současné angličtiny na pozadí češtiny*. Praha: Academia.

Ellis, R. and G.P. Barkhuizen (2005) *Analysing learner language*. Oxford: Oxford University Press.

Granger, S. (2002) 'A Bird's-eye view of learner corpus research', in S. Granger et al. (eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 3–33. Amsterdam: John Benjamins.

Granger, S. (2003) 'Error-tagged Learner Corpora and CALL: A Promising Synergy' *CALICO Journal 20*(3), 465–480.

Gilquin, G. (2014, July 23). *UCL - Corpora*. Retrieved August 10, 2014, from

http://www.uclouvain.be/en-258636.html

Halliday, M. A. K. (1989) *Spoken and written language*. Oxford: Oxford University Press.

Izumi, E. et al. (2004) 'SST speech corpus of Japanese learners' English and automatic detection of learners' errors'. *ICAME Journal 28*, 31–48.

Izumi, E. et al. (2005) 'Error annotation for corpus of Japanese learner English', in *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora*, 71–80. Jeju Island: Association for Computational Linguistics.

Lennon, P. (1991) 'Error: Some Problems of Definition, Identification, and Distinction'. *Applied Linguistics 12*(2), 180–196.

Loke, D. L. et al. (2013) 'A corpus based study on the use of preposition of time "on" and "at" in argumentative essays of form 4 and form 5 Malaysian Students'. *English Language Teaching 6*(9), 128–135.

Luzón, M. J. et al. (2007) 'Spoken corpora: New Perspectives in Oral Language Use and Teaching', in M. C. Campoy & M. J. Luzón (eds.) *Spoken Corpora in Applied Linguistics*, 3–32. Bern: Peter Lang.

Nemser, W. (1971) 'Approximative systems of foreign language learners'. *IRAL 9*(2), 115–124.

Nicholls, D. (2003) 'The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT', in D. Archer et al. (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*, 572–581. Lancaster: Lancaster University.

Schachter, J. and M. Celce-Murcia (1977) 'Some Reservations concerning Error Analysis'. *TESOL Quarterly 11*(4), 441.

Selinker, L. (1972) 'Interlanguage'. *IRAL 10*(1-4), 209–232.

Spillner, B. (1991) *Error Analysis: A Comprehensive Bibliography*. Amsterdam: John Benjamins.

Tarone, E., and B. Swierzbin (2009) *Exploring Learner Language*. Oxford: Oxford University Press.

Thornbury, S. (2006) *An A–Z of ELT*. Oxford: Macmillan Education.

Sources

Gilquin, G. et al. (eds) (1995–) *The Louvain International Database of Spoken English Interlanguage.* Centre for English Corpus Linguistics at UCL in Louvain.[14]

[14] Only the Czech part that is not currently not available on the CD ROM published in 2010 was used.

Resumé

Zkoumání žákovského jazyka se v poslední době těší velké oblibě především díky možnostem, které badatelům poskytuje korpusová lingvistika. V začátcích výzkumu žákovského jazyka bylo možné zkoumat jen omezené množství dat. Žákovské korpusy v současnosti umožňují výzkum založený na velkém objemu dat. Pro takový výzkum je ale vhodné mít korpus nějakým způsobem označkovaný pro snazší vyhledávání. Protože je žákovský jazyk specifický a výzkumníky na něm zajímají jiné prvky než v korpusech nežákovských, používá se kromě značkování slovních druhů také značkování chyb. Většina značkovacích systémů pro žákovské korpusy je ale v současnosti určena pro korpusy psané a nezohledňuje specifika mluveného jazyka. Tato práce si klade za cíl tato specifika na základě analýzy dat z české části mluveného žákovského korpusu LINDSEI určit a navrhnout úpravy v již používaném lovaňském značkovacím systému (Daugneaux a kol., 2008), který je určený pro psaný protějšek LINDSEI, korpus ICLE.

Žákovský jazyk je pro účely práce definován jako mluvený nebo psaný jazyk produkovaný žáky, pro které není jazykem mateřským. Začátky zkoumání žákovského jazyka spadají do 60.–70. let 20. století, kdy začal být chápan jako něco, co může mimo jiné osvětlit osvojování druhého jazyka. Do té doby převládající behavioristické teorie jazykové akvizice totiž pracovaly s žákovským jazykem jako s nedokonalou verzí jazyka cílového, která může výzkumníkům pouze ukázat, v kterých oblastech je výuka jazyka nedostatečná. To se změnilo, když v roce 1965 představil Chomsky (1965: 25) svůj koncept *LAD* (= language acquisition device – 'zařízení pro osvojování jazyka'). Tento koncept předpokládá, že děti se rodí s mechanismem, který již obsahuje určité gramatické vzorce a ty jsou pouze nastavovány na správné hodnoty jazyka, kterému je dítě v dětství vystaveno. Ačkoliv byl tento koncept primárně určen pro popis osvojování mateřského jazyka, byl později přejat i pro popis osvojování jazyka cizího.

Corder (1967/1981b) zdůraznil význam chyb v žákovském jazyce a především toho, co nám o žákovském jazyce, který chápe jako dynamický systém, mohou říct. Sám později navrhl pro žákovský jazyk termín idiosynkratický dialekt (Corder 1971/1981c). Termínů pro žákovský jazyk bylo navrženo více, v současnosti se používá Selinkerův (1972) termín mezijazyk (*interlanguage*), který zdůrazňuje především jeho pozici mezi jazykem mateřským a jazykem cílovým. Kromě interference popisuje Selinker i další procesy, které mohou být za odlišnosti mezi cílovým jazykem a mezijazykem zodpovědné. Mimo jiné tvrdí, že u určitých struktur může docházet k fosilizaci, ale už neupřesňuje, zda může být zvratná. Od 60. let se výzkum žákovského mezijazyka rozvíjí a vědci se soustřeďují na nejrůznější aspekty, včetně toho, jak je jazyk vůbec osvojován. Ve výzkumu žákovského jazyka se tak pracuje i s kognitivními teoriemi a sociolingvistikou.

Jednou z metod, jak studovat žákovský jazyk, je chybová analýza, která byla velmi populární po publikování Corderovy studie (1967/1981b). Tato metoda chápe chyby jako důležité projevy, které mohou pomoci popsat jazykový systém na jednotlivých úrovních osvojování cizího jazyka. Chyby tedy nejsou brány jako něco špatného. Corder navíc rozlišuje mezi chybami, které jsou systematické a ukazují na neznalost pravidel, a chybami, které mluvčí obvykle nedělá. Corder (1981a: 36) také popisuje jednotlivé kroky chybové analýzy. Začíná identifikací chyby. Druhým krokem je popis chyby, který by měl vést k systému klasifikace chyb. Posledním krokem je vysvětlení chyby. Na rozdíl od behavioristických teorií už zde Corder nepředpokládá jako jediný zdroj chyby interferenci mateřského jazyka. Chybová analýza byla sice velmi populární, ale zároveň velmi kritizovaná. Hlavní výtky shrnují Schachter a Celce-Murcia (1977). Jsou to: analýza chyby v izolaci, nepřesná klasifikace a tím pádem chybný výpočet frekvencí, a přílišná snaha vysvětlit chybu.

Definovat, co je to chyba, je obtížné. Sám Corder (1981a) se chybu nijak definovat nesnaží, pracuje s tím, že žáci dělají chyby a ty je třeba popsat. Později se chybu snažilo definovat mnoho lingvistů, ale neexistuje jedna obecně přijímaná definice. Pro účely této práce je chyba definována jako odchylka od pravidel standardní britské angličtiny. Tato pravidla jsou založena na popisu anglického jazyka v *Mluvnici spisovné angličtiny na pozadí češtiny* (Dušková a kol., 2006) a jsou doplněna popisem mluveného jazyka v *Longman Grammar of Spoken and Written English* (Biber a kol., 1999).

V současnosti je studium žákovského jazyka velmi usnadněno existencí žákovských korpusů, což jsou sbírky počítačově uloženého žákovského jazyka. Mohou být jak mluvené, tak psané, a kromě jazyka samotného obsahují i metadata o žácích. Velká část žákovských korpusů zkoumá angličtinu jako cílový jazyk. Vzhledem k technické náročnosti zpracování dat pro mluvený korpus převažují korpusy psané. Jedním z míst, kde žákovské korpusy vznikají, je CECL - Centre for English Corpus Linguistics v Lovani. Největším psaným žákovským korpusem v Lovani je ICLE, který v současnosti obsahuje 3,7 milionů slov. Pro tuto práci je nejdůležitější korpus LINDSEI, který obsahuje přepisy nahrávek studentů angličtiny. Studenti hovoří na tři témata: rozhovor na vybrané téma, volná diskuze a popis obrázku. Rozhovory jsou nahrávány a přepisovány podle jasně daných pravidel, takže výsledné přepisy obsahují informace o pauzách (a jejich relativní délce), cizích slovech použitých v konverzaci i o dalších zvucích jako jsou smích nebo kašel. Nahrávání probíhá na několika spolupracujících univerzitách, takže korpus obsahuje angličtinu mluvčích různých mateřských jazyků. V Lovani také vznikají dva srovnatelné korpusy rodilých mluvčích angličtiny (psaný LOCNESS a mluvený LOCNEC).

Žákovské korpusy je možné zkoumat dvěma různými způsoby (Granger 2002). Prvním způsobem kontrastivní analýza, kdy je srovnáván buď jazyk žákovský s jazykem cílovým, anebo žákovské jazyky mluvčích různých mateřských jazyků mezi sebou. Tento přístup může poskytnout zajímavé informace o nadužívání nebo naopak nízkém užívání některých jazykových struktur v rámci žákovského jazyka. Při srovnání napříč žákovskými jazyky je také možné odhalit univerzálnější tendence osvojování cizího jazyka. Druhým způsobem je pak chybová analýza, která se soustřeďuje na chyby v žákovských jazycích, snaží se sestavit jejich taxonomii a případně je v rámci korpusu označkovat. Oba tyto způsoby vychází z odlišných teoretických východisek (viz výše), ale navzájem se nevylučují.

Počítačem podporovaná chybová analýza opět umožňuje dva přístupy ke zkoumání dat. Za prvé je možné podívat se na nějaký předem určený problém v žákovském jazyce a vyhledat si pouze ten. To může být někdy výhodné, ale nese to s sebou riziko opominutí chyb, které zrovna badatel neočekává. Druhý způsob je mnohem pracnější, protože je nutné v korpusu označkovat všechny chyby. Na druhou stranu jakmile je značkování jednou hotovo, práce s korpusem je mnohem snazší a vyhledávání chyb rychlejší. Úskalím chybového značkování korpusu je nutnost vytvořit si pro chyby klasifikační systém. Jak se ukazuje při rozboru existujících systémů chybového značkování, tento krok není vůbec jednoduchý. V teoretické části práce jsou detailněji popsány 4 systémy chybové anotace: Lovaňský systém chybového značkování, značkování v Cambridgeském korpusu CLC, v japonském korpusu NICT JLE a značkování v rámci korpusu FreeText Project. Ačkoliv se klasifikace chyb v rámci těchto systémů liší, všechny jsou do určité míry založené na lingvistické klasifikaci, chyby jsou na určité úrovni klasifikovány podle slovních druhů, a přepisy také většinou obsahují opravu chyby. Všechny systémy pracují s různými úrovněmi v rámci klasifikace, ale pouze systém pro FreeText Project používá pro každou úroveň samostatný tag. Lovaňský systém pracuje s několika většími kategoriemi, které dále dělí (jsou to chyby v tvaru slova, gramatické chyby, lexiko-gramatické chyby, lexikální chyby, chyby v slovosledu a přebývající nebo chybějící slova, stylistické chyby, chyby v interpunkci a nepřesnosti). Značkování CLC probíhá podobně, většina tagů má ale jen dvě úrovně, první popisuje typ chyb (špatně zvolené slovo nebo chybějící slovo atd.), druhá většinou upřesňuje slovní druh. Kromě toho ale obsahuje velké množství "zvláštních značek". Značkování japonského korpusu je založené primárně na slovních druzích, ale obsahuje i kategorie, které už identifikují zdroj chyby (japonská angličtina) a sběrné kategorie pro chyby, které se jinam nevešly (neznámý původ apod.) NICT JLE je korpus mluvený, takže kromě značek pro chyby obsahuje i značky pro prvky typické pro mluvený jazyk, jako jsou opakování, opravy a zvuky jako např. smích. Největší výhodou tohoto systému je to, že je konzistentní, všechny značky jsou v jazyce XML a značí se začátek i konec značeného jevu, např. <laughter> </laughter>.

Nakonec jsou v teoretické části práce popsána specifika mluveného jazyka. Mluvený jazyk se od psaného nutně liší, protože je vždy produkován v reálném čase a většinou je nepřipravený. Jako prototypický příklad mluveného jazyka je brána konverzace (tak jak je popisována v mluvnicích od Bibera a kol. (1999) a Cartera a McCarthyho (2006)). Konverzace má většinou formu dialogu a je interaktivní. Charakter mluvené řeči je tedy dán i tím, že mluvčí může plánovat dopředu jen omezeně a může být přerušen. Mezi prvky typické pro mluvený jazyk patří váhání (sem patří nevyplněné a vyplněné pauzy, opakování a přeformulování fráze nebo věty). Kromě výrazů vyjadřujících váhání má mluvený jazyk také specifickou strukturu. V rámci mluveného jazyka neexistují věty tak, jak jsou definovány v psaném jazyce. Větné jednotky (jak je označují Biber a kol. (1999)) mají navíc tři typické části , kromě hlavního sdělení jsou zde ještě 2 části, "tag" a "preface", které Carter a McCarthy (2006) označují jako "headers" a "tails". Kromě specifik mluveného jazyka jsou ještě ve stručnosti popsány mluvené korpusy, jejich rozdělení i výhody korpusové lingvistiky pro studium mluveného jazyka.

Praktická část práce je založena na označkování 20 přepisů z české části korpusu LINDSEI a jejich rozboru. Na základě této analýzy jsou pak navrženy nové kategorie nutné pro lepší popis mluveného jazyka. V kategorii formy jsou zachovány morfologické chyby a naopak vypuštěny chyby v pravopisu, které pro mluvený jazyk nemají smysl. Gramatická oblast je dále členěna podle slovních druhů, většina chyb, které se vyskytují v psaném jazyce, se vyskytuje i v jazyce mluveném. U substantiv jsou to chyby v genitivu a v čísle. V kategorii determinátorů jsou nejčastější chyby v užití kvantifikátorů. I pro pokročilé nerodilé mluvčí angličtiny jsou problematické členy, kromě jejich špatného užití se v korpusu vyskytují i případy, kdy je použit špatný tvar správného členu ('a' místo 'an' apod.). Pro tyto případy je navržena podkategorie. Chyby v užití zájmen se týkají především zájmen osobních, ale v přepisech se objevují i chyby v zájmenech zvratných, vztažných a neurčitých. Vyskytují se i chybně užitá adverbia a místa, kde má být použit jiný slovní druh. Asi nejkomplikovanější kategorií jsou potom slovesa, kde se nejčastěji objevují chyby v pomocných slovesech a ve slovesném čase, oboje poměrně často u kondicionálů. Kromě toho se vyskytují i chyby morfologické a chybějící '-s' ve třetí osobě.

Na základě analýzy gramatických chyb je navržena předpona, která by v rámci tagu na první pohled odlišila, zda se jedná o chybu, nebo o prvek typický pro mluvený jazyk. Při rozboru se totiž objevily případy, kdy je těžké určit, zda je brát jako chybu. Týká se to nekonzistence při používání slovesných časů a při užívání zájmen se shodnou referencí. Další nově navržená kategorie je přípona, která vyjadřuje nejistotu. Současný systém nedovoluje mít u jedné chyby více než jednu značku a v případě, že je těžké přesně určit kategorii, může anotátor ponechat uživateli korpusu možnost vlastní interpretace.

Ostatní kategorie užívané pro psaný jazyk se v LINDSEI vyskytují také. Jsou to chyby lexikální, lexiko-gramatické (převážně chybné komplementace slov), chyby ve slovosledu (ačkoliv u nich se zdá, že přinejmenším část bude opět typická pro mluvený jazyk a nebude se jednat přímo o chyby) a nepřesnosti.

Kromě kategorií, které jsou obsažené v lovaňském značkovacím systému, jsou ještě navrženy kategorie typické čistě pro mluvený jazyk. První z nich jsou gramatické jevy, které se nedají v mluveném jazyce považovat za chyby. Druhou velkou kategorií jsou pak neplynulosti (dysfluencies), pod které se řadí opakování, reformulace a opravy. Opakování se v korpusu vyskytují velice často a v rámci toho, aby zůstal korpus přehledný, práce navrhuje značit opakování pouze složenými závorkami, nikoliv jim přidělovat zvláštní tag. Reformulace jsou pro účely žákovského korpusu definovány jako místa, kde se mluvčí v promluvě vrací a vybírá jiné slovo, frázi nebo strukturu. Na rozdíl od oprav ale ani jedna z částí reformulace neobsahuje chybu. Opravy se pak dělí na 3 podkategorie (zda se mluvčí opraví dobře nebo špatně a zda opravuje chybu nebo něco, co je správně).

Předkládaná práce potvrzuje hypotézu, že v mluveném žákovském korpusu je třeba značit i prvky specifické pro mluvený jazyk. Lovaňský systém je tedy pro potřeby LINDSEI upraven a jsou do něj přidány nové kategorie. Zároveň je ale třeba další výzkum, který otestuje navrhované kategorie a zanalyzuje data z korpusu LOCNEC.

Appendix 1:

List of tags

eFM	Error Form Morphology
eGNN	Error Grammar Noun Number
eGNC	Error Grammar Noun Case
eGDD	Error Grammar Determiner Demonstrative
eGDI	Error Grammar Determiner Indefinite
eGA	Error Grammar Article
eGAF	Error Grammar Article Form
eGPP	Error Grammar Pronoun Personal
eGPO	Error Grammar Pronoun Possessive
eGPI	Error Grammar Pronoun Indefinite
eGPF	Error Grammar Pronoun Reflexive
eGPR	Error Grammar Pronoun Relative
eGPD	Error Grammar Pronoun Demonstrative
eGADVO	Error Grammar Adverb Order
eGADV	Error Grammar Adverb
eGVM	Error Grammar Verb Morphology
eGVN	Error Grammar Verb Number
eGVAUX	Error Grammar Verb Auxiliary
eGVT	Error Grammar Verb Tense
eGWC	Error Grammar Word Class
eXADJCO	Error Lexico-Grammar Adjective Complementation
eXADJPR	Error Lexico-Grammar Adjective Dependent preposition
eXNPR	Error Lexico-Grammar Noun Dependent preposition
eXVPR	Error Lexico-Grammar Verb Dependent preposition
eXNUC	Error Lexico-Grammar Noun Uncountable/Countable
eLS	Error Lexis Single
eLSF	Error Lexis Single False friends
eLP	Error Lexis Phrase
eLPF	Error Lexis Phrase False friends
eLCS	Error Lexis Conjunction Subordinating

eWRM	Error Word Redundant Multiple
eWRS	Error Word Redundant Single
eWM	Error Word Missing
eWO	Error Word Order
eZ	Error Infelicity
eM	Error Mother tongue
sGPP	Spoken language Grammar Pronoun Personal
sGVT	Spoken language Grammar Verb Tense
sDR	Spoken language Dysfluency Reformulation
sDC	Spoken language Dysfluency Self-correction
sDCC	Spoken language Dysfluency Self-correction Correct
sDCW	Spoken language Dysfluency Self-correction Wrong
sDCE	Spoken language Dysfluency Self-correction from Correct form to a wrong one

Appendix 2

Categories				Examples
e	**F**	**M**		has quite a **(eFM)** pragmatical **$pragmatic$** approach
				I mean the youngest is . ten years old and the **(eFM)** olders **$older ones$** are . eleven
	G	**N**	**N**	especially on **(eGNN)** Sunday **$Sundays$**
				also her **(eGNN)** expressions **$expression$**
				one of my **(eGNN)** visit **$visits$**
				those were just . few words some . family members some **(eGNN)** animal **$animals$** . colours
				fact Delhi is a city but just in some **(eGNN)** part **$parts$**
	G	**N**	**C**	your **(eGNC)** bachelor's **$bachelor$** thesis
	G	**D**	**D**	he is just so clever **(eGDD)** so **$such a$** clever guy
				there is (eh) **(eGDD)** other **$another$** couple
	G	**D**	**I**	it started **(eGDI)** few **$a few$** years ago
				she is smiling . and: she has . har<?> (eh) hairdress (eh) her= hairstyle some haircut **(eGDI)** some **a** nice haircut
				we spent . **(eGDI)** much **$a lot of$** time on the Hebrides
				orientation is (er) . <lip sound> is **(eGDI)** some **a** discriminating factor
				there were also **(eGDI)** many **$a lot of$** people there
				. so **(eGDI)** much . of . **$much$** literature
				there's **(eGDI)** a lot **$a lot of$** unusual things
	G	**A**		you can . spot in **(eGA)** the **0** . todays' magazines
				for want of **(eGA)** the **a** better word
				(eGA) the **0** some some town by the sea because I like the . the: ... fishing-town look
				and it was in the dark and we couldn't . we: it was **(eGA)** a **0** really difficult because we almost missed the ship
				(eGA) the **0** Lochness lake
				I've chosen **(eGA)** the **0** topic three
				I wish I had some . I wish I had **(eGA)** 0 **a** chance to . (er) work

				with English
				if you have time and **(eGAu)** the **0** resources you can actually travel the whole length of it and and see everything
	G	**A**	**F**	for me it's **(eGAF)** a **an** important part of the . of the movie
				which has **(eGAF)** an . **a** (eh) strong (LS) impact $effect$ on my life
	Not tagged			falls in love with the with the[i:] oldest . daughter . Jane . and his friend Mr Darcy (eh) falls in love with . the[i:] . second . oldest . second oldest
				was a[ei] a[ei] experience also . very very powerful
	G	**P**	**P**	clothes and **(eGPP)** it's **$they are$** used very much
				but if if **(eGPP)** it **he** was . an artist . then: he shouldn't have done it
				could you tell us some: (ehm) give **(eGPP)** 0 **us** a tip for a . good . German TV show
				the weather got quite terrible and **(eGPP)** 0 **it** started raining so it was
				Canadian dollars they have (erm) (Z) the picture of the queen $the queen's portrait$ on **(eGPP)** it **$them$**
	G	**P**	**O**	I got **(eGPO)** 0 **my** bachelor (eXNPR) title
	G	**P**	**I**	we didn't have any mobile phones or **(eGPI)** something **$anything$** like that
	G	**P**	**F**	people .. don't want to see **(eGPF)** them **$themselves$** as they are
	G	**P**	**R**	authors **(eGPR)** which **who** are not really taught here very much
				the actors . (eh) **(eGPR)** which **who** are really good
				this system of colleges (eh) . **(eGPR)** where **$which$** is not as prominent as . in Oxford
				that was the reason . **(eGPR)** that **why** I am here
				it is about (em) . a couple of men who work there . **(eGPR)** that **who** are . <giggles> taking . part
	G	**P**	**D**	the dubbed movies they do here . so I prefer watching **(eGPD)** that **$them$** in English
				I love this . <lip sound> novel . and therefore . I really wanted to see

				(eGPD) that **it**
	G	ADV		this city it's . it's London . (eh) I've <laughs> I've been **(eGADV)** here **$there$**
				we were **(eGADV)** here **$there$**
				I'm not really sure if . it's my imagination or (WM) 0 if it's really **(eGADV)** here **$there$**
				in= **(eGADV)** nearby **$near$** <foreign> Liberec </foreign>
				the food . which . she (eh) gave us (eh) wasn't . good . **(eGADV)** too **$either$**
	G	ADV	O	there was **(eGADVO)** a band playing also **$also a band playing$**
				always went **(eGADVO)** a little back **$back a little$**
				: he **(eGADVO)** also . can **$can also$** speak Hebrew
				how (erm) . these **(eGADVO)** really (er) . actions can **$actions can really$** affect our (GNN) life $lives$
				you can **(eGADVO)** see there . (er) (mm) Al Pacino **$see Al Pacino there$**
				literature would **(eGADVO)** be also **$also be$** nice
				the same time I **(eGADVO)** really was **$was really$** happy that I was finally there
				would **(eGADVO)** communicate often **$often communicate$** in English
	Not tagged			I don't have really that much experience as my friends
				yeah still I I really love Oscar Wilde
	G	V	M	with her eyebrows roused (em) rised <overlap /> **(eGVM)** risen **$raised$**
				I **(eGVM)** no study **$don't study$** English language
				we: had **(eGVM)** went **$gone$** there
				he made her **(eGVM)** to look **$look$** better
				I managed to both read (er) the written version . and **(eGVM)** seen **see** the movie
				interesting for me to **(eGVM)** found **$find$**
	G	V	N	was like five . five parts and this also . on= only **(eGVN)** have **has** . two . hours

				where her problems . **(eGVN)** starts **$start$**
				Busan dialect it's . kind of sometimes it **(eGVN)** sound **$sounds$**
				. almost everyone in (erm) (eh) . in my surrounding<?> . around me . **(eGVN)** know **$knows$** English
				her hair **(eGVN)** are **is** different
				an experience that **(eGVN)** have **has** . taught me
	G	**V**	**AUX**	we **(eGVAUX)** should have read **$were supposed to read$**
				you **(eGVAUX)** are able to **can** understand everything
				if . (er) the woman . (er) **(eGVAUX)** would be **$were$** . (er) a really good friend of mine (er) . I think I would lie
				if if I **(eGVT)** didn't have **$hadn't had$** this experience I would probably **(eGVAUX)** fire **$have fired$** it up
				one would be quite lost (eLCS) when if he . when he he **(eGVAUX)** would get **got** a topic
				. if I **(eGVAUX)** would be **$were$** kind of rude I would say . okay
				it would be . very convenient (LCS) when if he . **(eGVAUX)** would marry **$married$** one of . her daughters
				I **(eGVAUX)** should **$am going to$** describe the film or play
				it **(eGVAUX)** was **$would be$** something impossible here in Prague
	G	**V**	**T**	they actually asked the lady . whether we **(eGVT)** are coming **$were coming$** again someday
				so my favourite . movie or . the movie I've . I **(eGVT)** saw **$have seen$** and I th= . I think that is really good
				he told us that: . people loved the forest part but that they . also **(eGVTu)** describe **$described$** it as similar to the Amazon forest or something like that so
				I wanted to know what they **(eGVT)** sing **$were singing$** about
				was really surprised that we **(eGVT)** have **had** everything dubbed
				he was (eLS) unable $incapable$ you know of thinking that he **(eGVT)** can **$could$** prepare two teas (eLP) at one time $once$
				and I **(eGVT)** was (er) in **$have been to$** (eh) (eGWC) German $Germany$ twice
				(eGVT) I've seen **$I saw$** it (er) on my birthday

				ever since we started snorkelling **(eGVT)** I always had **$I've always had$** this urge
	G	**WC**		his mother . didn't speak very **(eGWC)** well **$good$** English as well but
				when I was at home . I think for four months because (eh) . of the . health= **(eGWC)** healthy **$health$** reason
				Germany . is (eh) . is (er) much closer to us so to: (mm) . to speak **(eGWCu)** fluently **$fluent$** . German and know (eGADV) a lot of $a lot$ about (eh) history
				and I (eGVT) was (er) in $have been to$ (eh) **(eGWC)** German **$Germany$** twice
				that don't really sound all that **(eGWC)** well **$good$** in Czech
				everything turns out . turns out to be very . **(eGWC)** well **$good$**
	X	**ADJ**	**CO**	**(eXADJCO)** worth to say **$worth saying$**
				here I am **(eXADJCO)** used to work **$used to working$**
	X	**ADJ**	**PR**	woman . was . (erm) . **(eXADJPR)** blind on **$blind in$** one eye
	X	**N**	**PR**	one of the **(eXNPR)** books . from **$books by$** . Stephen King
				. make **(eXNPR)** contact to **$contact with$**
				I got (GPO) 0 my bachelor **(eXNPR)** title from **$title for$** that
				final **(eXNPR)** exams from **$exams in$** it
				people share some . <lip sound> **(eXNPR)** interest for **$interest in$**
	X	**V**	**PR**	. she's **(eXVPR)** pointing to **$pointing at$** something
				she could **(eXVPR)** boast with **$boast about$** . boast with it
				it **(eXVPR)** depends . on **$depends 0$** if he just . got the money for the portrait
				I don't want . (GVNF) **(eXVPR)** listen 0 $to listen$ **$listen to$** these theoretical things
				. it **(eXVPR)** reminds me **$reminds me of$** Oscar Wilde
				I **(eXVPR)** dropped out from **$dropped out of$** the other school
				this movie **(eXVPR)** provides you 0 **$provides you with$** some realistic (er) . <lip sound> (erm) . realistic view on America
	X	**N**	**UC**	waiting for the[i:] **(eXNUC)** outcomes **$outcome$**
	L	**S**		**(eLS)** cease **$fade$**

				not very **(eLS)** nice **$pretty$** girl
				I can't **(eLS)** recall **$remember$**
				which apparently doesn't sui= **(eLS)** suit **$please$** her
				I logged in a a Czech movie database (er) to: . **(eLS)** evaluate **$review$** the movie
	L	**S**	**F**	there are some **(eLSF)** actions **$special offers$**
				during **(eLSF)** gymnasium **$grammar school$**
				quite an **(eLSF)** affair **$big thing$**
				so she **(eLSF)** specialized **$planned$** the lessons according to the topics that were to be . (eh) discussed during the final exam
				(eLSF) linguistic **$linguistics$** which is not
	L	**P**		or **(eLP)** just after school **$freshly graduated$** . (em) teacher
				I think I keep . forgetting . the German . (em) **(eLP)** the German language **$German$**
				because . it's really not . **(eLP)** very possible **$possible$**
				it's also **(eLP)** wanted from **$expected of$** us to: (er) (Z) to get (er)
				in . **(eLP)** in a comparison to **$in comparison with$**
				he **(eLP)** enters the **$goes to$** university as well
				one of the women **(eLP)** has her head on the side **$tilts her head$**
				I would **(eLP)** say . truth **$tell the truth$**
	L	**P**	**F**	**(eLPF)** action Tuesday **$Tuesday's offers$**
				(eLPF) in the ideal case **$ideally$** people . or the students should have . read a lot of books
				that you are **(eLPF)** on a good way **$heading in the right direction$**
	L	**C**	**S**	you should do whatever he wants . **(eLCS)** even when **$even if$** he wants to . paint a completely different picture
				I think that . **(eLCS)** as **$once$** (eGVT) you're $you've been$ there for a longer and longer time . you get used to it
				it would be . very convenient **(eLCS)** when **if** he . (eGVAUX) would marry $married$ one of . her daughters
	W	**R**	**S**	so (er) we . **(eWRS)** usually **0** . used to (er) fire up something . which we found (er) in the street
	W	**R**	**M**	who invited a lady . to be painted to sit . (er) **(eWRM)** a model **0**

				for him
	W	**M**		. it might **(eWM)** 0 **be** his girlfriend or
				well . <laughs> this **(eWM)** 0 **is** really . (em) . childish perhaps
				. I . don't . like the[i:] atmosphere which is . on at (eGA) the concerts 0 <overlap /> usually . because **(eWM)** 0 **$there are$** too many people .
				and I'm on the other **(eWM)** 0 **$side$** of the . of the fence
				so . I **(eWM)** really looking **$was really looking$** forward . to it
				because . it **(eWM)** 0 **is** very interesting
				that was in Czech translation <overlap /> presumably is it right </A> <B> <overlap /> yeah . yeah yeah **(eWM)** 0 **it** was </B>
				we can see also a woman . it might **(eWM)** 0 **be** his girlfriend
	W	**O**		**(eWO)** here this in **$here in this$** seminar . there . are fifteen people so there's . discussion and so on
				of how long **(eWO)** can you **$you can$** stay down there how deep you can go . and we always wanted to have (er) wanted to have (er) some some depth gauge or something
				the play is . complex . **(eWO)** too much **$much too$** complex for for just
				what **(eWO)** you call **$do you call$** it
				I **(eWO)** found very interesting the comparison **$the comparison very interesting$**
	Z			and there is the . **(eZ)** this possibility is also available **$this is also possible$**
				mother **(eZ)** had . (er) **$gave birth to$** my little sister only
				so **(eZ)** he had he had the only idea **$the only thing he could suggest was$** to go . (er) . to friends
				(eZ) look into the depth (er) of the language **$explore language in depth$**
	M			<B> <foreign>**(eM)** jo **$yeah$** </foreign> airlines (eh) I don't remember I'm sorry
				it . but: he (eGADVO) also . can $can also$ speak Hebrew he can speak Arabic he can speak <foreign> **(eM)** no **$well$** </foreign>

				English and stuff of course
				Museum I (eGVT) I've seen $I saw$. (em) .. <foreign> **(eM)** sfinga **$sphinx$** </foreign> I'm not sure how to
		Not tagged		I went with a with a couple of my colleagues (eh) to <foreign> Čino= Činoherní klub </foreign>
				<foreign> Liberec </foreign> where I live
				and . (em) . <foreign> docent Čermák </foreign> was in the in the committee
				it's called (eh) <foreign> hrdý budžes </foreign>
s	**G**	**P**	**P**	some of them were . plays like drama . some of **(sGPP)** it was poems
				I started listening to the Beatles my dad loved them and . so I liked **(sGPP)** it too so I listened to it as well
				I . approached a person . and **(sGPP)** they were just okay I can't do anything about it go to a different person and the different person told me
				but then I forget (er) I mean all these Welsh names **(sGPP)** it's <overlap /> it's hard to remember
		Not tagged		these girls are probably not very . (er) honest . honest people yeah that these are . quite (em) . let's say . <lip sound> (eh) <starts laughing> yeah <stops laughing> I wouldn't judge it yeah . they can
	G	**V**	**T**	their kids **(sGVT)** got into a fight and one hurts the[i:] other . (em) and they start talking about this
				well was doing his best but she wasn't satisfied she **(sGVT)** seems to be criticising her portrait . so she yeah she **(sGVT)** is very upset obviously <laughs> with something so: . maybe she asked him to: . try another one just second attempt and: . the second one . with better hair and which is more . feminine or more more fashionable I don't know . possibly . (eh) was all right for her so . then she: . she bought the picture and she invited her friends to see it
				(uhu) (eh) maybe that here she doesn't like . she doesn't like the painting . so she she she **(sGVT)** told the painter to draw it . to draw her differently . and now when (eh) . he . (eh) . **(sGVT)** changed the the picture of her . the[i:] . her appearance she she's happy she's

				satisfied even though it's not really her so . <sniffles> it's the . hypocrisy and (erm) superficiality of of people . probably <laughs>
				I really liked it and I was finally . (er) in a group . (er) . where people **(sGVT)** share some . <lip sound> (XNPR) interest in for language as I do so
	D	**R**		which was very nice because **(sDR)** there were a lot of (erm) <lip sound> (er) . it was very international there were students from France Austria
				= I was living in a family **(sDR)** so I had a I had keys\because I think that (er) . (erm) . unless you are not gonna work (er) in a . <lip sound> . (er) unless you are gonna work in a . in . <lip sound> . (er) let's say . have some job where you can . work with your English
				that was **(sDRu)** a quite (eh) . an . advantage for me
	D	**C**		their children **(sDC)** which . you know you would think okay maybe there isn't a connection
	D	**C**	**C**	**(sDCC)** when we talking **when we were talking**
				then it got **(sDCC)** worst **worse** because the teachers got worse
				she wasn't really satisfied because . (er) for her . the portrait **(sDCC)** doesn't . (eh) **didn't** look . like her
				I was afraid that I . (er) **(sDCC)** I'm going I was . I was afraid that **I was going** to hate it
				. if you **(sDCC)** can bury **can be buried** in this river
				both **(sDCC)** this game and the movie . (eh) sorry **the play** and the movie was
				it's very very complex **(sDCC)** game . **a play**
				but I really **(sDCC)** like it . I really **liked it** .
				who (eh) . hears a a poem . (eh) **(sDCC)** on the school (er) ... (erm) yeah in **at school**
				she's very upset and cries **(sDCC)** at **on** the stage yeah
				I know that **(sDCC)** I have learnt . learnt everything them . **I have th= . taught them everything**
	D	**C**	**W**	I wish I had **(sDCW)** some . I wish I had (eGA) 0 a chance to . (er) work with English

				she's looking with her eyebrows **(sDCW)** roused (em) rised <overlap /> (GVM) risen $raised$
	D	**C**	**E**	I I am sorry I'm . **(sDCE)** my imagination . or (LSF) **fantasy** $imagination$
				the movie **(sDCE)** I've . I (GVT) **saw** $have seen$ and I th= . I think that is really good . is . Pride and Prejudice
Repetitions				also (eh) {one one} more (er) thing which was pretty important for me
				{we we} went by bus which is . a little annoying because it was a long long way .. but . it was definitely worth it (erm) {we we} went to
				{they {they} were . they were} pink
				a really good experience to . {think of think of} this novel
				she's creating . {a a} different . work of art .
				{what {what} would he what would he do}
				{when I . (erm) . (erm) when I was} at the . entrance exams here
				something {which (er) which one does not . (er) . which one does not do}
				so {you you . you} definitely {learn (er) . learn} {how to: . (er) . how to}
Not tagged				(er) because (eh) it waas girls who a= who accompanied me . (eh) in the end .. (em) . there had to be plenty of water