MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
Taras Shevchenko National University of Kyiv
Institute of Philology
Department of English Philology and Intercultural Communication

AUTOMATIC ERROR ANNOTATION IN LEARNER CORPORA

(on the material of Ukrainian learner English corpus)

Master's Thesis
written by the 2nd year student
of master's programme
"Modern English Communication
and Translation and Two
Western European Languages"
Field of Science – 03 "Humanities"
Specialty – 035 "Philology"
Hupalyk Iryna Andriivna

Supervised by:
Pastushenko Liudmula Pavlivna, PhD

«Допущено до захисту»
Протокол засідання кафедри англійської філології
та міжкультурної комунікації
Протокол № 11 від 24.05.2021
Завідувач кафедри_____проф. Бєлова А.Д.

КИЇВ – 2021

CONTENTS

INTRODUCTION

The comparison of non-native and native linguistic production reveals that the utterances of learners form a distinct linguistic system. Patterns of the errors non-native speakers make depend on several factors including the speakers' native language, other languages they might know, the stage and ways of learning the language, etc. Investigating this system is beneficial both to the study of second language acquisition and as a stimulus for the development of teaching methods, instructional materials and software tools for non-native language analysis. This is becoming even more important due to the advent and continuous advancement in Natural Language Processing, which has an enormous potential when applied to the analysis of texts produced by non-native speakers as well as certain challenges connected with the peculiarities of non-native texts.

Thus, the topicality of the research is determined by the increasing interest in the rules and principles that govern non-native speech production and the need for comprehensive analysis of these principles, as well as the importance of understanding how the language processing tools which were trained on native language data perform on non-native texts.

The novelty of the work: even though there have been previous attempts to compile the corpus of Ukrainian learner English, the collected data was quite scarce, the corpus annotation was mostly limited to part-of-speech tagging and the analysis of the data - to the classification of learner errors and establishing their frequency. Our work makes an attempt to compile the corpus of Ukrainian learner English annotated for parts of speech and syntactic dependency relations as well as learner errors denoting also the type of the error and its possible correction and thus applicable for extensive linguistic analysis.

The aim of the research: to explore the possibilities of applying automatic corpus annotation tools to a compiled learner corpus and the possible challenges of the automatic annotation process.

Research objectives:

- review the main theoretical principles of corpus design and annotation;
- determine the distinctive features of learner corpora in general corpus taxonomy, establishing the peculiarities of their compilation;
- collect and properly annotate a corpus of Ukrainian learner English;
- describe the existing POS tagging and syntactic parsing algorithms and tools and assess their accuracy when employed to annotated the texts produced by language learners;
- develop a comprehensive error taxonomy applicable to the compiled corpus;
- describe the most common types of learner errors and their impact on the performance of automatic annotation tools;
- develop the algorithm of an automatic learner corpus annotation, focusing in particular on part-of-speech tagging, dependency relations distribution and error tagging.

The object of the research: the performance of automatic corpus annotation tools on Ukrainian learner English corpus.

The subject of the research: the peculiarities of Ukrainian learner English and their influence on the accuracy rate of the performance of automatic corpus annotation tools.

The corpus consists of 340 essays of approximately 150,000 words created by the students of Taras Shevchenko National University of Kyiv in September 2019. The collected texts were then transcribed, automatically annotated and

manually reviewed to ensure the annotation accuracy. Each text is provided with metadata containing the information about the author of the text.

The practical significance: the findings will significantly advance learner corpus research in terms of methodological innovation. They can become a theoretical basis for learner corpora compilation and significantly ease the process of automatic corpus annotation. The collected data might also comprise a useful material for establishing the peculiarities of Ukrainian learner English at all language levels as well as SLA research; or be used as training data for various NLP tasks, in particular for Native Language Identification purposes.

The structure of the research is determined by the scientific logic, the aim of the research and its objectives. The work consists of an introduction, 3 chapters, conclusions and references.

CHAPTER 1. Learner Corpora: Compilation and Use

1.1. Corpus Linguistics and Linguistically Annotated Corpora

Corpus linguistics is one of the fastest-growing methodologies in contemporary linguistics. Being a rather new approach to language study, it supplies samples and linguistics information for all the branches of linguistics. It seeks to provide comprehensive samples of real-world usage in a particular language and use this empirical data to test language hypotheses. Today, corpora are used to advance virtually every aspect of linguistics, from computer processing techniques such as machine translation to literary styles, social aspects of language use, to improved methods of teaching language.

Employing language corpora enables the scholars to observe a natural language in the light of its actual use in normal regular life; provides ample evidence to analyze language with a degree of authenticity that was lacking in earlier language studies and helps scholars to make scientific observations on any aspect of a language through inductive inference, i.e. using numerous individual examples.

A corpus is usually defined as a large body of linguistic evidence typically composed of attested language use [41]. It is methodically designed to contain many millions of words compiled from different texts across various linguistic domains to encompass the diversity a language usually exhibits through its multifaceted use. It should be stated, though, that the term should be applied only to a well-organized collection of data, collected within the boundaries of a sampling frame designed to allow the exploration of a certain linguistic feature (or set of features) via the data collected.

Gries and Berez view the characteristics of a prototypical corpus as following:

- the corpus consists of one or more machine-readable Unicode text files

- the corpus is meant to be representative for a particular kind of speaker, register, variety, or language as a whole;

- the corpus is meant to be balanced, which means that the sizes of subsamples (speakers, registers, varieties) are proportional to the parts of such speakers, registers, varieties, etc. in the population that the corpus is intended to represent;

- and the corpus contains data from natural communication environments. This means that the language data in the corpus at the time of creation was not created solely for the purpose of entering it into a corpus [24].

Stressing the benefits of using corpora, Dash and Arulmozi (2018) point out the following types of information that can be retrieved from a properly annotated corpus:

(a) detailed information about all properties, elements, and components used in a language, such as sounds, characters, punctuations, phonemes, morphemes, words, stems, bases, lemmas, compounds, reduplications, multiword units, idioms, proverbs, set phrases, phrases, sentences, and so on;

(b) grammatical and functional information (e.g., forms, compositions, patterns of using affixes and inflections, patterns of constituent structure, contexts of use, usage patterns, meanings and sense variations) of words, phrases, sentences, idiomatic expressions, and so on found in a language;

(c) usage-based information (regular, specific, stylistic, metaphorical, allegorical, idiomatic, figurative, proverbial, etc.) of segments, morphemes, words, compounds, phrases, and sentences used in a language;

(d) textual and contextual cues of a text by way of providing information relating to time, place, and agent of a language event;

(e) information of the extralinguistic world relating to linguistic discourse [15].

The information of the extralinguistic world obtained from a corpus is analyzed simultaneously with intralinguistic information collected from

linguistic elements of a language in order to understand how a piece of text is composed and developed; how text is used; in which context it is used; and how it serves the needs of text users.

Taking all these findings into account, one cannot underestimate the importance of using corpora in many domains and subdomains of linguistics and neighboring disciplines. For instance, one can think of using varied speech corpora in the works of speech technology and tools development; using annotated text corpora in translation, lexicography and language teaching; using dialect corpus in description and analysis of dialects and local language varieties; using comparable corpora, parallel corpora, and translation corpora for machine translation, machine learning, and cross-lingual studies and resource generation etc.

At the same time one should be aware of the limitations of corpora and their use. Hunston (2002) has summarised some of the limitations to be taken into account, stressing the following issues:

(1) corpora present the language out of its context. Despite several possibilities to include information about textual and contextual data into the corpus, this kind of annotation is time consuming and consequently relatively little used.

(2) any corpus is a limited sample of language. Therefore the linguist must be very careful at making generalisations from a single corpus, as "*conclusions about language drawn from a corpus have to be treated as deductions, not as facts*" [28]

(3) a corpus can only provide information about whether something is used or frequent, but not whether something is correct or impossible.

(4) a corpus can offer linguistic evidence but not linguistic information. The corpus only lists several examples of language in use, or

frequency counts, but making sense of them is left to the researcher; it does not automatically provide answers to linguistic questions [28]

Awareness of these limitations is one of the reasons that has led corpus linguists to work more and more on specialised corpora, and use general ones for comparison. It has also become a frequent practice for corpus linguists to carry out the same type of analysis on several different corpora, or to compare corpus results to other types of empirical data or to a specific theory, before drawing generalised conclusions.

1.1.1. Corpus Annotation

The first major step in corpus building is the determination of the criteria on which the texts that form the corpus will be selected.

Among the most frequently selected criteria Wynne mentions:

1. the mode of the text; whether the language originates in speech or writing, or perhaps nowadays in electronic mode;

2. the type of text; for example if written, whether a book, a journal, a notice or a letter;

3. the domain of the text; for example whether academic or popular;

4. the language or languages or language varieties of the corpus;

5. the location of the texts; for example (the English of) UK or Australia;

6. the date of the texts [73].

Kennedy states that there are three stages to corpus compilation: corpus design, text collection or capture and text encoding or markup [32].

Corpus annotation refers to the practice of adding linguistic information to a corpus of written or spoken language. The types of linguistic information that can be added to corpora are wide-ranging, including lexical, morphological, syntactic, semantic, pragmatic, and discoursal, among others. In the case of spoken corpora, one can also include phonetic and prosodic

information [41]. The principles, practices, schemes and formats for corpus annotation at different linguistic levels have been widely discussed in corpus linguistics (e.g., Biber et al. [4], Huston [28], Kennedy [32], Lüdeling and Kytö [37], McEnery et al. [37], O'Keeffe and McCarthy [34], Teubert and Čermáková [63]).

Each annotation layer is important for a variety of uses, including serving as input for processing of other annotation layers. For example, part-of-speech (POS) annotated text is used as input for syntactic processing, for practical applications such as information extraction, and for linguistic research making use of POS-based corpus queries.

Corpus annotation may be achieved entirely automatically, by a semi-automated process, or entirely manually. Some NLP tools, such as part-of-speech taggers and lemmatizers are currently at a level where a fully automated approach could be considered. On the contrary, pure manual annotation occurs where the accuracy of available systems is not high enough to compensate for the time invested in manual correction as, for example, in encoding anaphoric and cataphoric references (Botley and McEnery [6], Mitkov [44]).

Linguistic annotation is available in many different forms, and on many different levels of linguistic information.

By lexical annotations, we mean annotations that are restricted to individual words, i.e. they do not cover more than one word. Lexical annotation tags terms or phrases with synonyms, similar words, translation and abbreviations, so a term rather equals the entity described by it instead of the word. Furthermore, words are tagged with their corresponding lexicalization, word stems and derivations.

Part-of-Speech-Tagging is one of the most common and most frequently used types of annotation, as it is relevant for many corpus linguistic studies and it is involved in many other annotation processes such as lemmatization, syntactic analysis, semantic annotation, etc.. Here each tokenized word is

assigned a label that minimally identifies the part of speech of the word, but typically also includes some grammatical category data. The accuracy of the automatic part-of-speech tagging depends to a large extent on many factors, including the language represented by the corpus and its morphological properties, the complexity of the texts in the corpus, the type of tagger used (symbolic or, more often today, statistical), the size and precision of the corpora the tagger was trained on, the size of the tag set, etc [24].

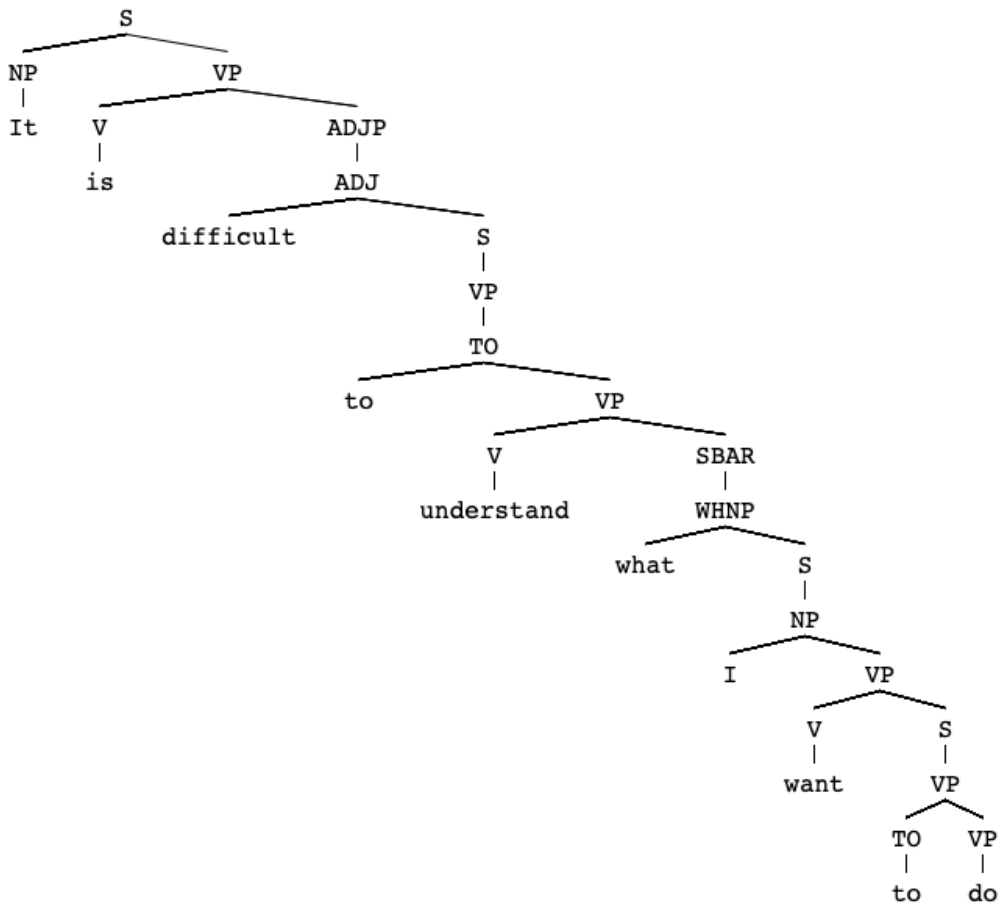
Another type of lexical annotation is lemmatization, the process of identifying and marking each word in a corpus with its base form . Lemmatization can be done on the basis of an existing form-lemma database, a (semi-) automatic approach called stemming, in which word forms are truncated to arrive at a more general representation of a lemma, or some hybrid approaches from these two strategies that can also include morphological and / or syntactic analysis to distinguish ambiguous forms.

The capability afforded by lemmatization to treat different inflectional forms of a lemma as the same word instead of different words is a very useful one in corpus analysis. Whereas in some types of corpus analysis inflectional forms are of utmost importance (e.g., the analysis of tense and aspect), in other types of analysis it is lemmas that we are concerned with. For example, in the analysis of frequency distribution of words in a large corpus, it is important to be able to perform the analysis using lemmas in addition to surface forms.

Syntactic annotation concerns the annotation of structural information, and is generally carried out on the sentence level. Currently, there are two major syntactic theories that are used for syntactic annotation: constituent-based and dependency-based annotations.

In constituent-based annotations, the goal is to identify groups of words that function as a unit. Smaller constituents are grouped into larger ones to build

a hierarchical structure, also called a syntax tree. The figure below show an example of a syntax tree taken from the Penn Treebank:



In computational linguistics corpora serve as a source of raw material; they are used to obtain an overview of the data occurring in natural language and to determine the scope of the phenomenon that one wants to examine. Corpus annotations encode diverse kinds of information, such as part of speech, lemma, word sense, syntax trees as well as the logical document structure or the content structure of texts.

Due to the interest in annotated corpora, a lot of work in computational linguistics has been devoted to the development of corpus tools, such as tools to assist the annotator in the annotation. In this area, computational and corpus linguistics completely overlap in their interest in tools and methods.

1.2. Learner Corpora and the Peculiarities of Their Compilation

Like any corpus, the learner corpus is a '*collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety*' [41]. What makes the learner corpus special is that it seeks to be representative of a certain language variety. This peculiarity is highlighted in the definition of the notion given by Nesselhauf (2004), where it is understood as '*systematic computerized collections of texts produced by language learners, where systematic means that the texts included in the corpus were selected on the basis of a number of – mostly external – criteria (e.g. learner level(s), the learners' L1(s)) and that the selection is representative and balanced'* [49]. The definition also stresses the importance of design criteria, which in many regards determine the potential of the application of a particular corpus.

Jarvis and Paquot state that an ideal learner corpus would be one where

(1) all of the texts in the corpus are written on the same topic or at least in the same genre with a symmetrical distribution of topics across L1 groups,

(2) the texts are all of similar lengths or have similar length means and standard deviations across L1 groups,

(3) all of the learners are at precisely the same level of L2 proficiency or at least the levels of proficiency are evenly balanced across L1 groups,

(4) the learners within and across groups have similar educational, socio-economic and psychological profiles, and

(5) the learners within and across groups have had comparable amounts and types of instruction in and exposure to the target language [31].

Gilquin classifies the types of learner corpora along to certain dimensions, some of which are common to all corpora while others are specific to learner corpora:

1) medium: written/spoken discourse. Written learner corpora are in general more numerous than spoken learner corpora but a number of spoken learner corpora have become available over the last few years as well. Another potential field of research is constituted by multimodal (or audio-visual) learner corpora including video recordings, which allow for some new spheres of investigation like the analysis of learners' gazes or gestures.

2) genre, where potentially any genre (or combination of genres) may be represented. However, here Gilquin stresses that in practice, the variety of genres tends to be limited as a result of the restricted number of genres for which a second language variety is actually used and learner corpus compilers' preference for certain genres, for example argumentative essays among written learner corpora [21].

3) the target language that corpora represent. At the moment, English is the most predominant target language. However, over the last few years, some new projects have been launched that collect data representing other target languages, e.g. German, French, Czech etc. While most learner corpora are monolingual, containing data from only one target language, a small number of learner corpora are multilingual, like the MiLC Corpus.

4) the learners' native language. Currently, a larger part of learner corpora is represented by mono-L1 learner corpora, among which Asian learners are the most widely represented. About a third of all the learner corpora are multi-L1; in this case, learners from several L1 populations have contributed to the corpus [64].

5) time period: synchronic, i.e. assessing the learners' knowledge of the target language at a particular moment, or longitudinal, i.e. gathering learner output produced at different stages in their development [21]. Belz and Vyatkina here use the term 'developmental learner corpus' to refer to the corpora '*in which learner performance is documented at close intervals or at all points of production*' [3]. This kind of corpora makes it possible to

investigate learners' progress over time and is therefore a precious resource. However, because such corpora are difficult to compile, there are very few currently available.

6) scale: global or local. Most learner corpora are global, being part of large-scale projects. However, Mukherjee and Rohrbach talk about the compilation and use of local learner corpora, which are typically collected by teachers among their students, who are both contributors to and users of the corpus. They stress that the objective of such an approach is to identify one's own learners' specific needs through a corpus analysis of their output and thus provide tailor-made solutions to their problems [46]. Rankin and Schiftner also recognise a further type of corpus in between global and local learner corpora, namely in-house learner corpora, i.e. '*local reference learner corpora which reflect the production of a given learner population*' [53]. In this case, the contributors and the users are not the same students, but they come from the same population (typically, the same school/university), which enhances the relevance of the analyses of these data for the users.

7) origin and purpose: commercial or academic. Commercial learner corpora are collected by publishing houses with a view to developing pedagogical materials (dictionaries, coursebooks, etc.) based on authentic learner output. Most of the time, these corpora are not publicly available. The two most notable examples of commercial learner corpora are the Longman Learners' Corpus and the Cambridge Learner Corpus. Unlike commercial learner corpora, academic learner corpora are initiated by researchers and/or teachers working in educational settings and interested in learning more about interlanguage [20].

The most important features of non-native English corpora is the availability of large metadata characterising the author of each text and the context of its production. It allows the investigation of complex feature

interactions, the possibility to filter data according to a particular characteristic or to focus only on one level of proficiency.

Tono presented the possibilities of using learner corpora, which can be summarized in five points:

(1) description of developmental levels of the learner interlanguage,

(2) studying the influence of mother tongue and language transfer,

(3) defining the overuse and underuse of linguistic expressions in the learner language,

(4) distinction between universal errors and errors due to the learner's mother tongue; and

(5) distinction between elements of communication in native and non-native speakers that are responsible for the foreign touch [65].

1.2.1. Overview of the existing corpora

The 'Learner corpora around the world' website maintained by the University of Louvain currently contains 137 learner corpora, 82 (60%) representing L2 English, the rest focusing on other languages (Arabic, French, German, Korean, Spanish, etc.). In terms of medium and text type, the dominant focus is on writing, in particular essay writing, but there is a general diversification of data types and, especially, a growing number of projects on learner speech.

Information about the most representative corpora of non-native English is summarized in the following table:

Name of corpus	**Size of corpus**	**L1**	**L2 Level**
The International Corpus of Learner English (ICLE)	6,085 essays	Various	Advanced

The NUS Corpus of Learner English (NUCLE)	1,400 essays	Chinese	Various
TOEFL11	12,100 essays	Various	Various
The Cambridge Learner Corpus for First Certificate in English exam (CLS FCE)	17,450 essays	Various	Upper-Intermediate
The Chinese Learner English Corpus (CLEC)	1,000,000 words	Chinese	Various
The Advanced Learner English Corpus (ALEC)	1,300,000 words	Swedish	Advanced
The Cambridge Learner Corpus (CLC)	50,000,000 words	Various	Various

The International Corpus of Learner English (ICLE) is based at the Universite Catholique de Louvain in Belgium and is one of the largest academic learner corpora. It is considered the first learner corpus created in an academic setting, its compilation having begun in 1990. The first version of the corpus contained 2.5 million words produced by learners from 11 mother tongue backgrounds. The second version was larger in terms of both words (3.7 million) and language backgrounds (16). The current version is even larger, as it includes data from 25 mother tongue backgrounds, amounting to 5.7 million words [23].

The main aim of ICLE is to study the inter-language of the foreign language learner. In particular, ICLE was set up to provide a resource for large-scale comparative studies of the interlanguage of advanced EFL learners with

significantly different native language backgrounds. The research goals of ICLE are the following:

(1) to collect reliable evidence of learners' errors and compare them across languages to determine whether they are universal or language-specific. In addition, the comparison is carried out to determine how far they are influenced by factors in the learner's cultural or educational background.

(2) to examine aspects of the foreign language use in non-native essays that are normally exposed through overuse or underuse of words or structures in relation to the norm of the target language. This investigation is carried out by comparing individual L2 subcorpora and native English corpora [23].

ICLE was innovative in that it drew on the rich methodological background of corpus linguistics while improving data collection beyond what had previously been achieved, both in terms of the volume of data collected and the number of respondents, providing a machine-readable, reusable resource that was easily accessible; and covered a number of L1 backgrounds for the L2 speakers in the corpus [40].

CLC (Cambridge Learner Corpus) is one of the largest commercial learner corpora. The commercial corpora stood out from the rest because their aim is to assist English Language Teaching / Training (ELT) publishers in compiling ELT dictionaries and other ELT resources such as ELT course books. CLC consists of exam scripts written by students who take English exams around the world and serve as the basis for a comprehensive analysis of the most common errors made by learners. In addition, the data is used by CLC to answer questions about how students learn at different skill levels and to ensure that the assessment of students' exams is consistent across countries and years.

TOEFL11 consists of essays written during the TOEFL®R test. The corpus contains 1,100 essays for each of the following 11 native languages: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish,

Telugu and Turkish. The essays were written in response to eight different prompts and are presented along with the grades for each essay. It was specially developed to support the NLI task and has become a standard frame of reference for NLI research. The TOEFL11 dataset was used in the NLI Shared Task 2013: a set of 900 essays per language was used as training data, while 100 essays per language were used as development data and a further 100 essays per language formed the sentence test [11].

1.3. The Use of Learner Corpora in Natural Language Processing

Natural Language Processing (NLP) deals with the representation and the automatic analysis and generation of human language [32]. It overlaps with learner corpora in the representation and automatic analysis of learner language [43]. We can distinguish three main uses of NLP involving learner corpora:

(1) NLP tools are used to annotate learner corpora with a variety of general characteristics and, based on this, to gain insights into the nature of language acquisition or the typical needs of learners. This includes general linguistic properties of part of speech and morphology through syntactic structure and dependency analyzes to aspects of meaning and discourse, function and style as well as properties specific to the language of the learner, such as different types of learner errors from lexical and syntactic to discourse, function and use. The use of NLP annotation tools can be combined with human post-processing to eliminate potential problems created by automated analysis.

(2) NLP tools are used to provide specific analyzes of learner language in the corpus. For the task of Native Language Identification (NLI), for example, classifiers are trained to automatically determine the mother tongue of the second language learner who has written a particular essay, or to determine the learner's level of competence.

(3) Learner corpora are used as data sets to train NLP tools, especially the statistical or machine learning components. The trained NLP tools can then be applied to learning languages that occur in other contexts. A tool trained on a learner's corpus to identify certain types of learning errors can be used to provide instant individual feedback to learners doing exercises in an intelligent tutoring system.

Let’s look into some of these uses in detail.

1.3.1. NLP and Learner Corpora in Second Language Acquisition

The use of NLP in language learning includes the development of NLP techniques for the analysis of learner language by tutoring systems in Intelligent Computer-Assisted Language Learning [23], automated scoring in language testing, as well as the analysis and annotation of learner corpora. The potential applications here can be summarized as followed:

(1) NLP technologies are used to analyze authentic spoken or written texts in the target language at a wide range of linguistic levels (e.g., lexical, morphological, and syntactic levels). The information generated by the analyses and annotations are then used to select, sequence, or enhance texts as language input for learners or to automatically generate exercises contextualized in those texts. In particular, the researchers have used NLP technologies to automatically generate exercises contextualized in input texts so that learners can perform different tasks on the noticed forms. Examples of such tasks include recognition of occurrences of specific linguistic structures in text and multiple choice or cloze exercises for specific linguistic forms.

(2) NLP technologies are used to analyze the spoken or written production of learners. The analysis then constitutes the basis for assessing learner performance, identifying the learners’ first language, or providing individualized feedback to the learners.

There has been substantial research on building reliable models for automatic essay scoring. Such models typically incorporate a large set of linguistic features (e.g., lexical richness, syntactic complexity, cohesion, etc.) that have been shown to be indicative of writing quality and that significantly correlate with human ratings. Not only are such models a cost-effective alternative to human scoring, but also they can be useful for learners to obtain individualized feedback on their writing.

A related area of research that has received increasing attention is automated grammatical error detection and correction for language learners. Tools built for this purpose are useful for assessing the level of accuracy of learners' language production, an important dimension of second language proficiency. Such tools are also useful for assessing the distribution of different types of errors, and for providing language learners with immediate feedback on their errors. In addition, there is a growing body of research on automatic native language identification of learner text. Based on the theoretical assumption that learners' L1 background may have a significant effect on their L2 use, research in this area seeks to identify features of language use that are characteristic of learners of certain L1 backgrounds. Output of this line of research can also be used to provide tailored feedback to learners with different L1 backgrounds.

Different tutoring systems use NLP to provide individualized feedback to learners working on activities, and thus to individually adjust the sequencing of the material and to update the learner model [29]. For exercises that explicitly or implicitly require the learner to provide responses using forms from a small, predefined set, it often is possible to anticipate all potential well-formed and ill-formed learner responses, or at least the most common ones given a particular learner population. The intended system feedback for each case can then be explicitly specified for each potential response.

Finally, some research has attempted to build models for automatically determining the extent to which free text responses produced by learners answer open-ended reading comprehension questions, generally by comparing the meanings expressed in learner responses to models of expected responses. This line of research has the potential of dramatically increasing the scope of meaning-based activities that can be incorporated in ICALL systems.

(3) NLP technologies are used to build dialogue systems that allow learners to interact with a conversational agent in the target language in a meaningful and coherent way via text or voice. Dialogue systems generally require at least three components: language understanding, dialogue management, and language generation. For voice-based dialogue systems, the language understanding and language generation components also need to be able to perform speech recognition and speech synthesis, respectively. Dialogue systems designed for language learners may be oriented to specific tasks and may incorporate useful cultural and pragmatic knowledge. Such systems allow language learners to engage in free interaction with native-speaker like conversational agents and to receive feedback on their language production [74].

1.3.2. Learner Corpora and Native Language Identification

Native language identification (NLI) is the task of identifying the native language of a person based on their texts written in their second language. It is based on the assumption that the native language influences writing in the second language due to the cross-linguistic influence, i.e. "*the influence resulting from similarities and differences between the target language and any other languages that has been previously (and perhaps imperfectly) acquired*" [51]. In NLI research, machine-learning programs are used to create classifiers (or computational models or systems) that represent the relationship

between L1s (classes) and characteristics of learners' lan- guage use (features). Features are often operationalised as the relative fre- quencies of specific letters or combinations of letters, specific words or sequences of words, parts of speech (POS) or sequences of POS, specific errors or categories of errors, or abstract properties such as levels of cohe- sion, complexity and lexical diversity [59].

The features that can be employed for the aim of native language identification include among others lexical choices [7], grammatical patterns [60], [72] as well as error taxonomy.

Lexical choices are viewed not only as the carriers of topic-specific information but also as one of the strongest indicators of the speaker's native language, as proven by various studies on the ICLE dataset [7]. Some authors even speak about the lexical choices in second language writing influenced by sounds and sound patterns of the native language [66] or by so-called *cognates*, i.e. the words that have a common etymological origin and thus a similar form and meaning in two different languages [17]. The researchers argue that such phonology or cognate interference may cause the writer to misspell the intended L2 word under the influence of the L1 spelling. However, Markov stresses that caution should be taken when generalizing the finding obtained on a lexical feature set, e.g., word n-gram models, given that they may lead to unintended extraction of topic or domain information, rather than capturing general characteristics of the native language of the author.

Making use of the grammatical patterns as the indicators of the speaker's language background, researchers most often look at the errors non-native speakers make, trying to trace their native language characteristics as reflected in non-native writing. Wong and Drag [72] focus in this regard on the following types of syntactic errors: subject-verb disagreement, noun-number disagreement and misuse of determiners. From the structural point of view, a great deal of attention is devoted to dependency features via capturing the

overall structure of grammatical constructions and global syntactic patterns such as preferences for particular grammatical forms, e.g. active or passive voice etc. [62].

Function words can be seen as indicators of the grammatical relations between other words. Their role in assigning syntax to sentences is linguistically well-defined. They belong to a set of closed-class words and embody relations rather than propositional content. Due to their topic independence and structural usage, function words are often amongst the most frequent in the corpus. Examples of function words include articles, prepositions, determiners, conjunctions, and auxiliary verbs. As a result of language transfer, non-native speakers often misuse or ignore certain function words, as a consequence, function words are considered strong context- and topic-independent indicators of the native language of the writer.

POS tags and n-grams of such features have shown to be one of the most prominent reflections of morpho-syntactic aspects of the native language in non-native language writing. Basic categories include verbs, nouns, and adjectives, but can be expanded to include additional more fine-grained morpho-syntactic information such as tense, case, gender, number, person, verb transitivity, etc. This representation can encode word order and grammatical properties of the native language by capturing the use or misuse of well-established grammatical structures, e.g., verb-subject-object, subject-verb-object, and subject-object-verb, etc.

Beyond the standard types of features mentioned, the features that can be found useful in terms of NLI also include syntactic constituency structures, syntactic dependencies, Tree Substitution Grammar features, measures of textual cohesion, lexical sophistication, syntactic complexity and conceptual knowledge and psychological indices (e.g. sadness, negative emotion, overall affect).

One major problem that NLI shares with many other NLP tasks is the dependence on the text type of any learned models. For instance, the ICLE consists of semi-formal learner essays on personal experiences, which are presumably of limited use to induce an NLI model useful on other types of texts.

The correlation between a person's native language and aspects of their writing in a second language can be useful in various applications. It can become a part of second language acquisition research and thus a source of more targeted feedback to language learners as well as the basis for developing more interactive learning systems tailored to different language backgrounds. In business, it can be employed in the process of market segmentation, identifying the customer demographics or getting a broader understanding of customer behaviour. It can also become a part of author profiling and authorship in general, for example, as a tool for forensic investigations etc.

Increased interest in NLI brought unprecedented levels of research focus and momentum, resulting in the first NLI shared task being held in 2013. The shared task aimed to facilitate the comparison of results by providing a large NLI-specific dataset and evaluation procedure, to enable direct comparison of results achieved through different methods. The best teams achieved accuracies of around 80% on this 11-class classification task where the great majority of entries used standard features such as POS n-grams [62].

Current research on NLI tends to a wider variety of corpora, genres and conditions in order to determine how extensive and reliable L1 influence is, through which combinations of features it manifests itself and how well its manifestations can be detected through the proper combination of scholarship and technology.

CONCLUSIONS TO CHAPTER 1

In the first chapter of the paper, we have explored the existing approaches to corpus design and compilation, and outlined the main features of learner corpora and the potential of their application. The conducted analysis of the works of Ukrainian and foreign linguists allowed us to come to the following conclusions:

1. The use of corpus linguistics methodology opens immense possibilities for further linguistic analysis, including grammatical, functional as well as usage-based information about a particular text.

2. Corpus annotation is a multi-layered process which can be conducted on different linguistic levels and achieved automatically, manually or combining these two approaches.

3. Learner corpora constitute a particular type of corpora the peculiarities of which are determined by the medium, genre, scale and purpose of the corpus.

4. In the process of compiling a learner corpus great attention should be paid to the corpus annotation process, which might be more complicated in comparison with annotation of native language corpora, as well as providing large metadata containing the information about the text producer and the context of its production.

5. The potential of using learner corpora in second language acquisition lies in developing the tools for assessing learner performance, in particular automatizing essay scoring, or providing individualized feedback to the learners.

6. In terms of native language identification the NLI mechanisms and tools can make use of the features contained in learner corpora, such as lexical choices, grammatical patterns and error taxonomy.

CHAPTER 2. Creating the Corpus of Ukrainian Learner English

Learner corpora pose challenges in the process of corpus annotation, in particular of automatic annotation, given the fact that non-native language use is more likely to contain non-standard spellings, lexical items, and grammatical constructions that training data for the existing annotation tools are unlikely to contain. Thus, such annotation efforts will most probably require close attention in choosing the right tagging algorithm, and more manual checking than for native language use.

2.1. Corpus Design

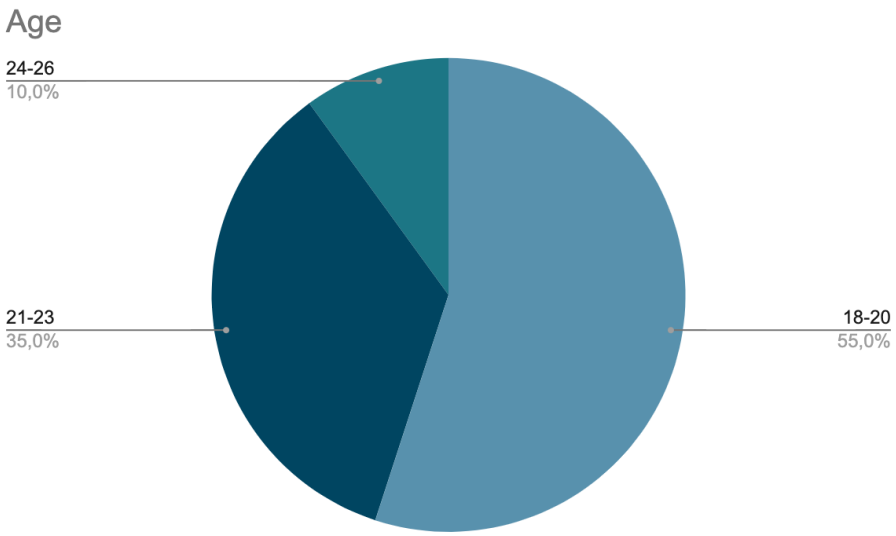
Our corpus consists of student essays written during the admission test for "Taras Shevchenko National University Language Club" student initiative. The corpus contains 340 essays.

The main data collection for the corpus took place in September and October 2019. The data then have been transcribed, processed and semi-automatically annotated. The collected texts are annotated using XML and linked to meta-data about the author of the text. In corpus design we have taken into account a number of variables that we found relevant for non-native language corpus building. Among the 29 variables which corpus builders should consider listed by Atkins, Clear and Ostler, we have focused on eight major learner variables - four general (learner's age, gender, region and mother tongue background) and four L2-specific (learning context, proficiency level, amount of L2 exposure and knowledge of other foreign languages) variables [1].

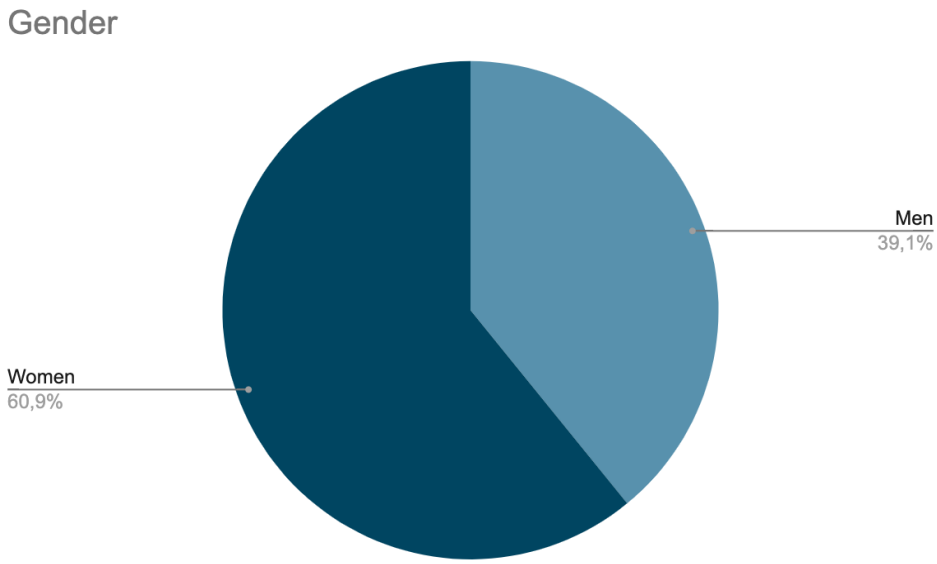
The data about the learners collected in our corpus include age, gender, first language, proficiency level in English according to CEFR, knowledge of

other (non-native) languages, English learning background, and time spent in an English-speaking country.

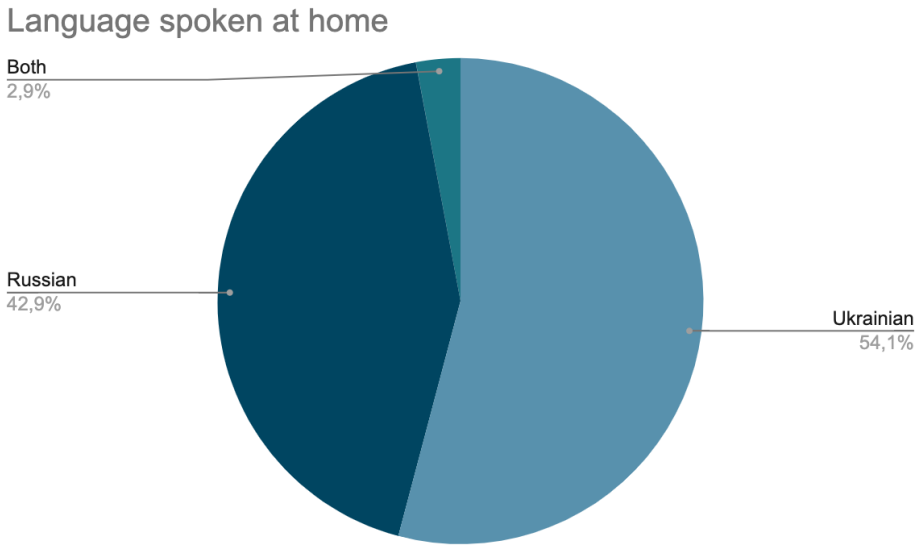
Age. All the participants are university undergraduates and therefore usually in their twenties. Thus the age distribution can be summarized as following:



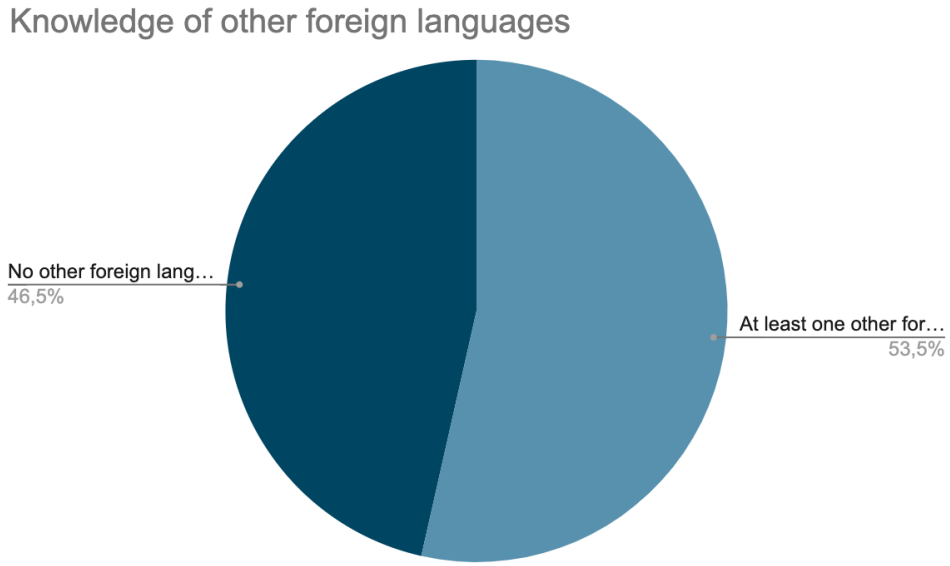
Gender. As the humanities tend to attract more female than male students, the corpus appeared to be slightly female-dominated with 60.9% of the data being produced by female learners.



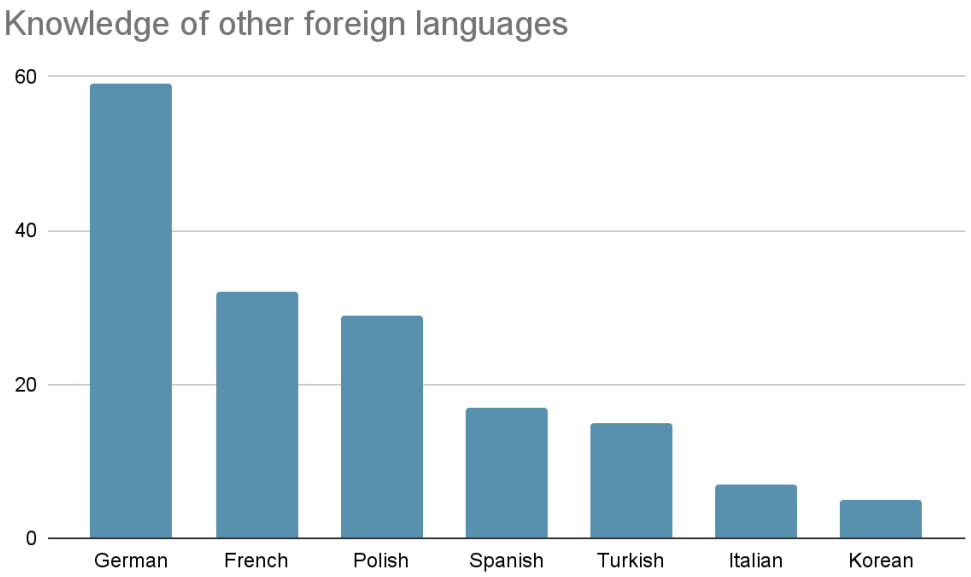
Mother tongue. The corpus is intended to contain the texts produced by the learners whose mother tongue is Ukrainian. However, in order to have a more precise picture of the learners' language backgrounds, we also recorded the information about the languages they speak at home.



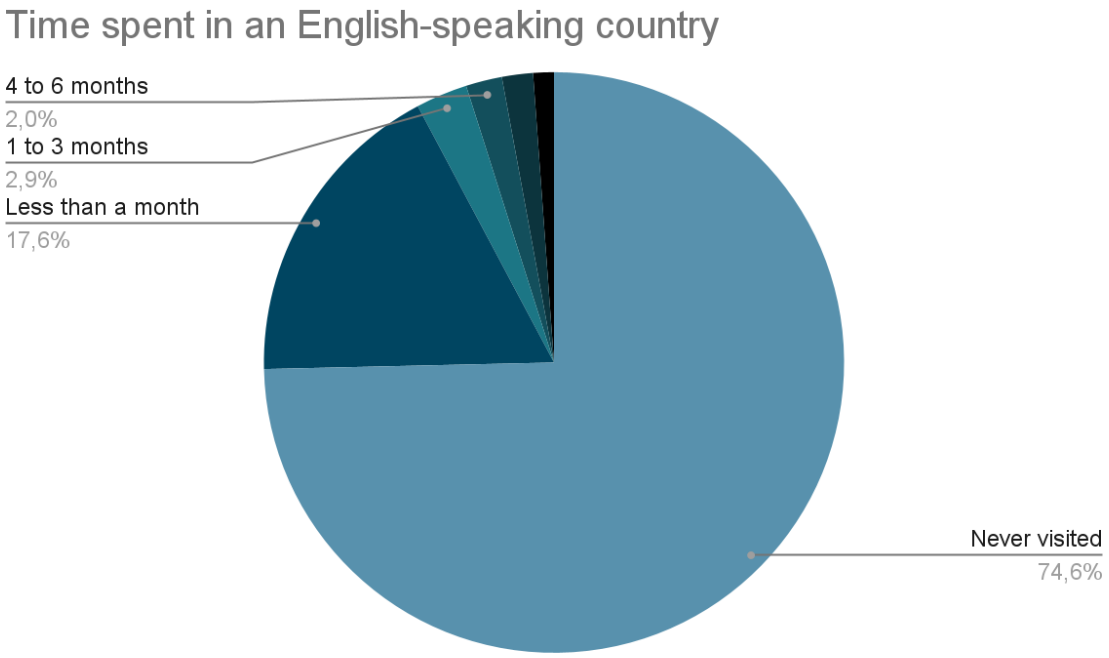
Knowledge of other foreign languages. This factor is useful to record as the learners' L2 may be influenced not only by their mother tongue, but also by their knowledge of other foreign languages. 53.5% of the participants reported to have an experience of learning at least one other foreign language apart from English while 46.5% have never had such an experience.



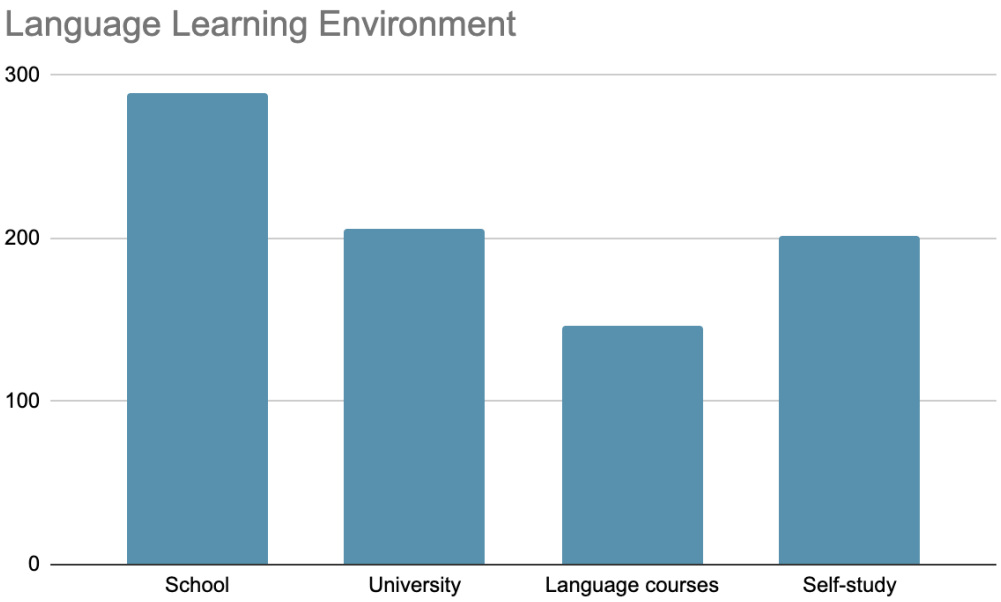
Top 7 learners' foreign languages are listed in the following figure:



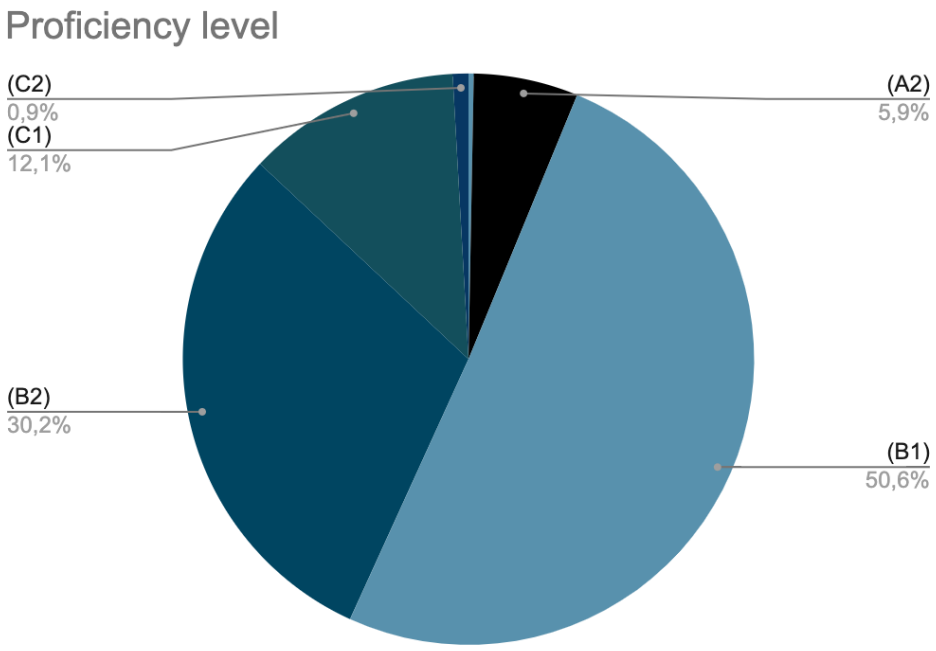
Time spent in an English-speaking country. While a large proportion of the learners (74.6%) reported no stay in an English-speaking country, 5.4% reported a stay of 3 months or more and 20% a stay of less than 3 months.



Learning context. Almost all the learners represented in the corpus learned English primarily in a classroom setting despite some individual variability in encountering English outside the classroom).



Proficiency. The proficiency level was established according to the Common European Framework of Reference for Languages. In order to ensure the consistency of the corpus, the participants were asked to take a standardized placement test, the results of which along with the overview of the essay were a basis for establishing the learners' proficiency level. Intermediate level (B1) appeared to be the most widely represented (50.6%), followed by Upper-Intermediate level (30.2%) and Advanced level (12.1%).



Due to the format of the admission test, the materials we collected were mostly hand-written. Therefore, each text was transcribed preserving the original spelling and grammar.

All learner essays in ICLEv3 were lemmatized and part-of-speech tagged with the Stanford POS tagger. The conducted analysis has shown that despite having been trained on native speaker corpora, the Stanford POS tagger has been found to have a high success rate with learner data. Its success rate has, however, proved to be very sensitive to morpho-syntactic and orthographic errors. It tends to decline as the number of errors increases. Therefore, at first we have conducted a spelling error annotation and provided each error occurrence with a tag containing the corrected spelling option, which has helped to significantly increase the POS tagger accuracy rate. Then the data were annotated for constituency and syntactic dependency relations using the Stanford Universal Dependencies Parser.

Corpus Overview

Documents	340
Sentences	7,198
Word tokens	146,706
Error annotations	5,602
Number of sentences per document	21.1
Number of word tokens per document	431.48
Number of word tokens per sentence	20.38
Number of error annotations per document	16.47

We intend it to be a useful resource for exploring the features of non-native English writing as well as the native language identification task. A corpus of such texts can be used to compare different varieties of non-native

language, or non-native and native language on the background of traditional native language corpora.

2.2. Corpus Annotation

2.2.1. Part-of-Speech Tagging

Part-of-speech (POS) tagging refers to the task of annotating every token in a text with a tag or label that indicates its part-of-speech category [36]. This level of linguistic annotation facilitates a number of different types of linguistic analysis that are otherwise difficult or impossible to perform. First, for a word that belongs to multiple POS categories, we can differentiate the occurrences of the word used as different parts of speech. POS tagging mostly relies on statistically observed sequences of certain parts of speech as well as lexical item information.

Generalised across different types of computational architectures, the task of POS tagging can be divided into three subtasks:

- Segmentation of text into tokens (tokenization)
- Assignment of all potential tags to tokens, which often means that more than one tag is assigned to a word (potential tag assignment)
- Determining the contextually appropriate tag from the potential tags (disambiguation)

The Stanford POS tagger was initially released by the Stanford Natural Language Processing Group in 2004 and has since been updated on a regular basis. For English, this tagger adopts the Penn Treebank POS Tagset, and the best reported accuracy is 97.24%, achieved on the Penn Treebank Wall Street Journal (WSJ) Corpus. Current downloads contain three trained tagger models for English, two each for Chinese and Arabic, and one each for French,

German, and Spanish. The tagger can be retrained on any language, given POS-annotated training text for the language [68].

The typical annotation looks as following:

```
1
2   Every/DT day/NN I/PRP start/VBP my/PRP$ morning/NN by/IN exercising/VBG.
3
4   This/DT habit/NN help/VBP me/PRP to/TO keep/VB a/DT good/JJ shape/NN.
5
```

The tagger makes use of the Penn Treebank POS Tagset, which has the following 36 POS-tags:

CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
		TO	*to*
JJ	Adjective	UH	Interjection
JJR	Adjective, comparative	VB	Verb, base form
JJS	Adjective, superlative	VBD	Verb, past tense
LS	List item marker	VBG	Verb, gerund or present participle
MD	Modal	VBN	Verb, past participle
NN	Noun, singular or mass	VBP	Verb, non-3rd person singular present
NNS	Noun, plural		
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

The Natural Language Toolkit (NLTK) is a set of computational linguistics and NLP program modules, annotated corpora and tutorials supporting research 25 and teaching for the Python language. NLTK allows various NLP tasks by providing implementation of various algorithms such as the Brill tagger, HMM based POS tagger, n-gram based taggers etc.The Universal tagset of NLTK comprises 12 tag classes: Verb, Noun, Pronouns, Adjectives, Adverbs, Adpositions, Conjunctions, Determiners, Cardinal Numbers, Particles, Other/ Foreign words, Punctuations.

```
import nltk
text = nltk.word_tokenize("This habit help me to keep a good shape.")
text
['This', 'habit', 'help', 'me', 'to', 'keep', 'a', 'good', 'shape']
nltk.pos_tag(text)
[('This', 'DT'), ('habit', 'NN'), ('help', 'VBP'), ('me', 'PPO'),
('to', 'TO'), ('keep', 'VBP'), ('a', 'AT'), ('good', 'JJ'), ('shape', 'NN')]


DT/This NN/habit VBP/help PPO/me TO/to VBP/keep AT/a JJ/good NN/shape.
```

The TreeTagger is a tool for annotating text with part-of-speech and lemma information which has been developed within the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger has been successfully used to tag German, English, French, Italian, Spanish, Bulgarian, Greek and old French texts and is easily adaptable to other languages if a lexicon and a manually tagged training corpus are available.

Sample output looks as following:

```
% echo 'Every day I start my morning by exercising.'
    'This habit help me to keep a good shape.' | cmd / tree - tagger - english
          reading parameters ...
          tagging ...
  Every       DET     every
  day         NN      day
  I           PN      i
  start       VBP     start
  my          PP$     my
  morning     NN      morning
  by          IN      by
  exercising  VBG     exercising
  .           SENT    .
  This        DT      this
  habit       NN      habit
  help        VBP     help
  me          PPO     me
  to          TO      to
  keep        VBP     keep
  a           AT      a
  good        JJ      good
  shape       NN      shape
  .           SENT    .
           finished.
```

Keeping in mind that the POS-taggers have been trained on the basis of native speaker corpora, we decided to primarily establish the accuracy rate of the POS-tagger analyzing non-native English data. For our analysis we have randomly selected five essays of roughly 400 words from the compiled corpus. The essays were tagged with each tagger, after which we have examined the tagger output and marked all the incorrect tags.

Error and ambiguity rate for POS-tagging

	Ambiguity rate	Error rate
NLTK	3.00%	6,7%
Stanford POS tagger	3.83%	5.4%
TreeTagger	5.18%	7.7%

One of the main sources of tagger errors has shown to be spelling mistakes. Following the findings of van Rooy and Schäfer, we have established the following categories of spelling errors:

Influence of spelling errors on tag correctness

Category	Total errors	NLTK correct	Stanford POS tagger correct	TreeTagger correct
Non-word errors	38	30	20	14
Real-word errors	14	5	5	1
Capitalisation	3	2	1	1
Space missing	10	0	0	0
Extra space	13	0	0	0
Total	78	37	26	16

Under non-word errors we understand the cases when the result of the misspelling is a word that doesn't exist; real-word errors include the cases when the result of the misspelling is a different real word of English; and capitalisation errors comprise the use of capital letter in middle of sentence, or sentence-initial word or proper noun used without capital letter. The category space errors comprises cases where two words are written as one word or where a single word is written as two words. Van Rooy and Schäfer here point out that errors with spacing always cause a tag error, so in all the cases a tag has to be added or deleted during manual tag correction to provide a correctly tagged token [70]. The data we received in the tagging process also support this suggestion.

Having analysed the tags wrongly assigned to the misspelled words, we have arrived at the conclusion that while word errors can be handled in some

cases, because the taggers employ their guessing modules to assign tags to non-words, real-word errors were far more problematic, since very often the actual form used is from a different part of speech as the intended form.

To determine the effect of spelling mistakes on tagger performance, we manually edited the corpus sample before retagging it, which has shown some impressive improvements in terms of tagging accuracy:

Tagger performance after spelling correction

Category	Total errors	NLTK correct	Stanford POS tagger correct	TreeTagger correct
Non-word errors	38	38	36	28
Real-word errors	14	13	12	13
Capitalisation	3	3	3	3
Space missing	10	10	4	2
Extra space	13	13	11	11
Total	78	77	66	57

The correction of spelling errors therefore proves to contribute significantly to the minimisation of tagger errors, without being such a time-consuming process as manually tagging the entire corpus, or manually editing the output of the automatic tagging.

2.2.2. Syntactic Parsing

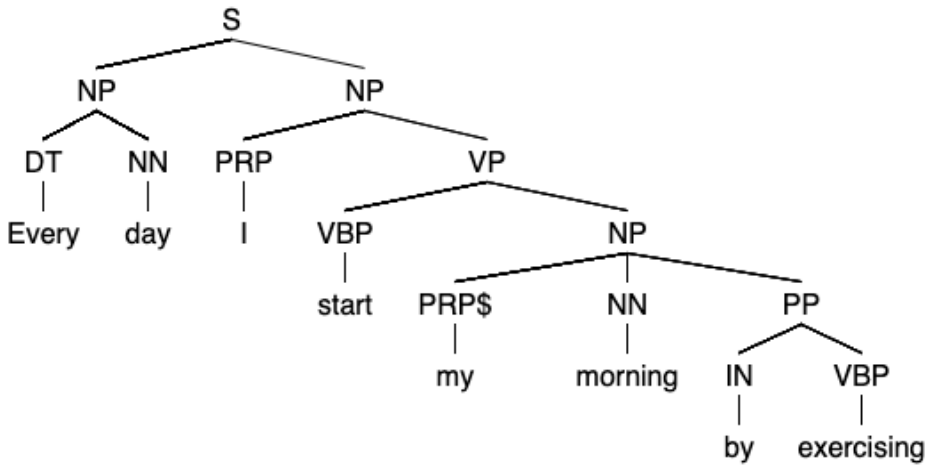
A closely related issue to POS tagging is that of parser performance. Since parsers usually rely on POS tag sequences, and are trained on, and expect, well-formed input, there is a high probability of incorrect parses being produced if the L2 input diverges greatly from typical L1 data and/or incorrect POS tags have been assigned.

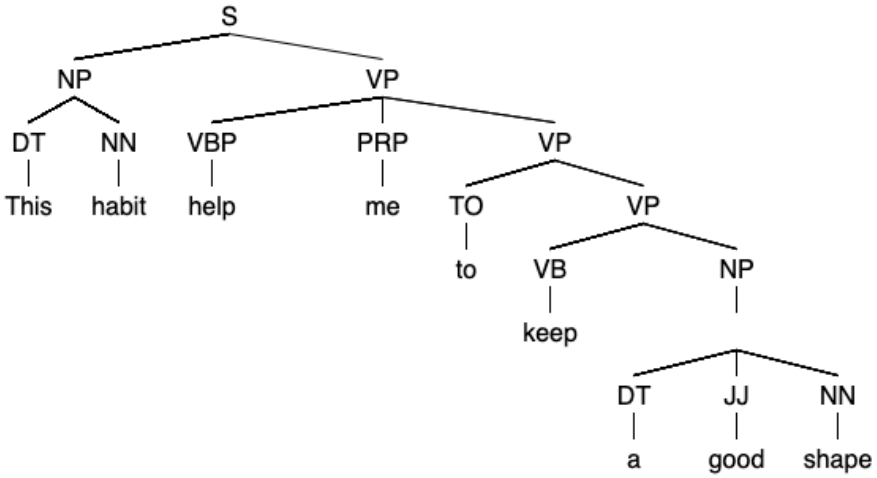
Syntactic parsing refers to the process of determining the syntactic structures of sentences, and the computational tools that are used to automate this process are referred to as syntactic parsers. This level of corpus annotation makes it possible for us to perform linguistic analyses that require information on the partial or full structure of the sentences in the text.

Sentence structure is viewed here in terms of constituency relation. A constituent of a sentence is a single word or a group of words that functions as a single unit in the sentence, such as a noun phrase that functions as the subject or object of the sentence, or a prepositional phrase that functions as the post-modifier of a noun. The constituents of a sentence are hierarchically related. For example, a prepositional phrase may be part of a noun phrase, which in turn may be part of a verb phrase [39]. The hierarchical relationship among the different constituents is indicated by means of bracketing and indention:

```
1
2   (ROOT
3     (S
4       (NP(DT Every)(NN day))
5       (NP(PRP I))
6       (VP(VBP start)
7         (NP(PRP$ my)(NN morning))
8         (PP(IN by)
9           (S
10            (VP(VBG exercising)))))
11      (. .)))
12
13  (ROOT
14    (S
15      (NP(DT This)(NN habit))
16      (VP(VBP help)
17        (S
18          (NP(PRP me))
19          (VP(TO to)
20            (VP(VB keep)
21              (NP(DT a)(JJ good)(NN shape))))))
22      (. .)))
23
```

This bracketed output can be visually represented using the phrase structure tree, with punctuation marks removed. This tree was generated using the Syntax Tree Generator:





Dependency grammars are a family of grammar formalisms that view sentence structure in terms of dependency relation instead of constituency relation.In dependency grammars, a dependency relation is defined as a relation that holds between a pair of words in a sentence, where one word (the dependent) is said to depend on or to be governed by the other word (the governor). Robinson proposed three axioms that govern the well-formedness of dependency structures. These are:

(1) one and only one word in a sentence is independent,

(2) all other words depend directly on some other word, and

(3) no word depends on more than one other word [54].

The independent word in a sentence is generally the matrix verb, which is also said to be the root of the sentence.

Annotating our corpus for dependency relations we tested the following state-of-the-art parsers on the annotated data: Stanford Statistical Natural Language Parser, Charniak-Johnson parser and MaltParser.

BLLIP (Charniak-Johnson) Parser is a statistical natural language parser including a generative constituent parser and discriminative maximum entropy reranker.

MaltParser is a system for data-driven dependency parsing, which can be used to induce a parsing model from treebank data and to parse new data using

an induced model. MaltParser has been developed by Johan Hall, Jens Nilsson and Joakim Nivre at Växjö University and Uppsala University, Sweden. MaltParser is based on the transition-based approach to dependency parsing, which means that a parser consists of a transition system for deriving dependency trees, coupled with a classifier for deterministically predicting the next transition given a feature representation of the current parser configuration. In order to derive training data for the classifier, an oracle is used to reconstruct a valid transition sequence for every dependency structure in the training set. A transition system together with an oracle defines a parsing algorithm.

All three parsers make use of the tagset developed and employed

Grammatical relations in the Stanford dependencies

abbrev	Abbreviation modifier	nn	Noun compound modifier
acomp	Adjectival complement	npadvmod	NP as adverbial modifier
advcl	Adverbial clause modifier	nsubj	Nominal subject
advmod	Adverbial modifier	nsubjpass	Passive nominal subject
agent	Agent	num	Numeric modifier
amod	Adjectival modifier	number	Element of compound number
appos	Appositional modifier	parataxis	Parataxis
attr	Attributive	partmod	Participial modifier
aux	Auxiliary	pcomp	Prepositional complement
auxpass	Passive auxiliary	pobj	Object of a preposition
cc	Coordination	poss	Possession modifier
ccomp	Clausal complement	possessive	Possessive modifier
complm	Complementizer	preconj	Preconjunct
conj	Conjunct	predet	Predeterminer
cop	Copula	prep	Prepositional modifier
csubj	Clausal subject	prepc	Prepositional clausal modifier
csubjpass	Clausal passive subject	prt	Phrasal verb particle
dep	Dependent (default relation when specific relation cannot be determined)	punct	Punctuation (if retained)
		purpcl	Purpose clause modifier
		quantmod	Quantifier phrase modifier
dobj	Direct object	rcmod	Relative clause modifier
expl	Expletive	ref	Referent
infmod	Infinitival modifier	rel	Relative
iobj	Indirect object	root	Root
mark	Marker	tmod	Temporal modifier
mwe	Multi-work expression	xcomp	Open clausal complement
neg	Negation modifier	xsubj	Controlling subject

Here are given the examples of the syntactic dependency annotation process:

```
LexicalizedParser lp = LexicalizedParser.loadModel(
                           "edu/stanford/nlp/models/lexparser/englishPCFG.ser.gz",
                           "-maxLength", "80", "-retainTmpSubcategories");
TreebankLanguagePack tlp = new PennTreebankLanguagePack();
tlp.setGenerateOriginalDependencies(true);

GrammaticalStructureFactory gsf = tlp.grammaticalStructureFactory();

String[] sent = "This", "is", "an", "easy", "sentence", "." ;
Tree parse = lp.apply(Sentence.toWordList(sent));

GrammaticalStructure gs = gsf.newGrammaticalStructure(parse);
Collection<TypedDependency> tdl = gs.typedDependenciesEnhancedPlusPlus();
System.out.println(tdl);

```

and of a sample output:

```

 det(day - 2, Every - 1)
 obl: tmod(start - 4, day - 2)
 nsubj(start - 4, I - 3)
 root(ROOT - 0, start - 4)
 nmod: poss(morning - 6, my - 5)
 obl: tmod(start - 4, morning - 6)
 mark(exercising - 8, by - 7)
 advcl(start - 4, exercising - 8)

 det(habit - 2, This - 1)
 nsubj(help - 3, habit - 2)
 root(ROOT - 0, help - 3)
 obj(help - 3, me - 4)
 mark(keep - 6, to - 5)
 xcomp(help - 3, keep - 6)
 det(shape - 9, a - 7)
 amod(shape - 9, good - 8)
 obj(keep - 6, shape - 9)

```

Precision rate for each parser

Parser	Precision rate
Stanford	67.3%
Charniak	70.5%
MaltParser	73.0%

We have analyzed the overall effect of learner errors on parsing, having established that learner errors are among the most frequent reasons for parser errors. Among the words that contained parsing errors, 54.2% had at least one learner error tag. Most of the errors made by the parsers are due to incorrect phrase structures leading to higher or lower attachment as well as to the use of the imprecise generic dep relation. The latter is produced when the dependency extraction code has difficulty labeling a relationship within a parse tree. Among the learner errors that became the most frequent reasons for parsing errors we should mention punctuation errors, spelling errors, capitalization errors, wrong argument structure, missing determiners and missing prepositions. Therefore, correcting these learner errors will be an effective pre-processing technique to reduce parsing errors for NLP application on learner English.

However, despite the significant impact learner errors have on the parser performance, the precision rate is still relatively high. We assumed that the seemingly high accuracy rates might be caused by the prevalence in learner English of short and simple sentences. To investigate the performance of the parsers on shorter sentences as compared to the longer ones, we grouped the sentences by their length, calculating the average accuracy rate for each group. The conducted analysis has shown that the dependency parsing performance was better on shorter sentences. The accuracy of POS-tagging performance appeared to be another crucial factor: when fewer POS errors occurred, it also positively influenced the performance of the dependency parser.

CONCLUSIONS TO CHAPTER 2

In this chapter we have taken a look at the performance of automatic corpus annotation tools when applied to a learner corpus, taking into account the nature of learner mistakes and their impact on the accuracy rate. As the

material for our research we have taken a corpus of written texts produced by Ukrainian speakers, which was compiled and annotated by the author of the research. The performance of the tools was assessed on the randomly selected essays with a total amount of around 2 000 words.

For assessment of the accuracy of POS tagging we have chosen the Stanford POS tagger, the set of the Natural Language Toolkit for POS tagging and the TreeTagger, all of which adopt the Penn Treebank POS Tagset for English. All the taggers have shown a relatively high accuracy rate, with most mistakes being connected with errors in spelling or word boundaries. After the manual correction of spelling errors the performance of the taggers has significantly increased, indicating the importance of preprocessing of the text in the POS tagging process.

Assessing the accuracy of syntactic dependency parsing we used the Stanford Statistical Natural Language Parser, BLLIP (Charniak-Johnson) Parser and MaltParser. Grammatical relations were presented in accordance with the Stanford dependencies tagset. The accuracy rate appeared to be much lower than for POS tagging, which was mostly connected with a range of learner errors, in particular, spelling and punctuation errors, wrong argument structure and missing determiners and prepositions as well as errors in POS tagging process. However, among the factors that increased the tagger performance we should mention a general tendency of non-native English speakers to use shorter and more simple sentences.

CHAPTER 3. Error Detection in Compiling the Corpus of Ukrainian Learner English

Error-tagging is a specific type of annotation and the corpora on which error-tagging is done is normally associated with second language learners. Maybe the most frequent exploitation of an error-annotated corpus is to observe the frequency of errors made by second language learners and their type. Not only do we need to understand the types of errors that exist in learner corpora, but we also need to evaluate the effect such errors have on the reliability of existing corpus analytical tools. Not all learner errors affect automatic corpus analysis in the same way, and this evaluation helps us better understand the magnitude as well as the sources of the problem.

3.1. Error Tagsets and Error Taxonomies

The distinguishing characteristic of learner corpora is the fact that they can be used to learn what kinds of errors L2 writers are most prone to, taken either as a general population or only considering subsets by L1 or proficiency level, for example.

The process of error tagging is far from straightforward. Firstly, the nature of the error itself has to be taken into account.

It cannot be refuted that error taxonomies should be based on the description of observable data and include established language categories in order to minimize subjectivity in the process of error detection and classification. However, having looked into the existing research on error classification (James 1998; Tono 2003; Diaz-Negrillo 2006), we can state that learner corpus researchers have yet to agree on a general algorithm of error

annotation. A good overview of the general criteria involved in such taxonomies is found in James [30], where the various dimensions involved in characterising an error are described.

Rosen et al. summarize the existing error taxonomies as following:

- Linguistically-based taxonomies, with a varying degree of detail, ranging from general categories (morphology, lexicon, syntax) to specific labels (auxiliary, passive, negation).

- Taxonomies based on a formal classification of surface alternations of the source text, such as missing, redundant, faulty or incorrectly ordered elements.

- A combination of several taxonomies, e. g., a multi-dimensional scheme consisting of an error domain (formal, grammar, lexicon, style), an error category (agglutination, diacritics, inflection, derivation, gender, mode), and word class (POS) [55].

While choosing an error taxonomy, one should focus on detail of error description and specificity of its use on a specific kind of material. We agree with the assumption of Meunier [64] that "*the more refined the tagset the more refined the analysis*" and therefore make use of detailed error tagging tools in order to provide fine-grained analysis of error-tagged data.

Descriptive detail can be reached in the taxonomy by the incorporation of several sets of information in tags, consisting of:

- identification of the units under description. Alongside error information, the unit where the error is found is identified in the tag with, for example, a punctuation mark for punctuation errors, POS information for grammatical and lexical errors, syntactic functions for syntactic errors, etc.

- distinction between internal and external errors under the major category of word grammar, and

- narrow linguistic subcategorization of errors.

In our research we follow the approach of Tono [65] who distinguishes the following dimensions in error taxonomy:

(1) Grammar-based error types – a taxonomy classifying errors from a grammatical perspective: errors in spelling, morphology, word boundary, agreement, government, lexical issue, style, punctuation; also can be qualified as "linguistic category classification" [30] or "linguistically based errors" [37].

(2) Formal error types – error types capturing the formal nature of an error without referring to possible underlying grammatical reasons: diacritics, capitalization, metathesis, missing element; also called "target modification taxonomy" [30] or "edit-distance based errors" [37]

In annotation process we have decided to make use of the tagset employed in the creation of the ICLE, which comprises 52 tags based around 7 major categories: form, grammar (general rules of grammar related to POS), lexicogrammar (POS complementation, use of dependent prepositions, count/uncountable nouns), lexis (semantic and collocational properties, subcategorisation), word (repetitions or omissions of words, wrong word order), register, style (longer stretches of text, incomplete or unclear passages), punctuation (whether missing, redundant or misused) [23] .

A typical tag structure consisted of three parts, indicating the linguistic unit, error subcategory and language level of the error (e.g. *I dreamed* {WM - to W} *go to England for all my life* where 'WM' signifies 'word missing' - the error subcategory, '-to' signifies a missing particle, and W signifies a language level of the error).

Louvain Error Tagset

Error categories		Error subcategories	
(F)	Formal	(FM) (FS)	morphology spelling
(G)	Grammatical	(GA) (GADJCS)	article adjective, comparative/superlative

		(GADJN) (GADJO) (GADVO) (GNC) (GNN) (GP) (GVAUX) (GVM) (GVN) (GVNF) (GVT) (GVV) (GWC)	adjective, number adjective, order adverb, order noun, case noun, number pronoun verb, auxiliaries verb, morphology verb, number verb, non-finite/finite verb, tense verb, voice word class
(X)	Lexico-grammatical	(XADJCO) (XADJPR) (XCONJCO) (XNCO) (XNPR) (XNUC) (XPRCO) (XVCO) (XVPR)	adjective complementation adjective-dependent preposition conjunction complementation noun complementation noun-dependent preposition noun countable/uncountable preposition complementation verb complementation verb-dependent preposition
(L)	Lexical	(LCC) (LCLC) (LCLS) (LCS) (LP) (LS) (LSF)	conjunction, coordinating connector, logical, complex connector, logical, single conjunction, subordinating lexical phrase lexical single lexical single, false friends
(W)	Word redundant, word missing, word order	(WR) (WM) (WO)	word redundant word missing word order
(P)	Punctuation	(PM) (PR) (PW)	punctuation missing punctuation redundant wrong punctuation mark
(R)	Register		
(S)	Style	(SI) (SU)	incomplete unclear

In the process of annotation we have discovered the most frequent types of errors according to the categories:

Distributions of errors according to the categories

Tag	Number of occurrences	Percent
F	1422	25.4
G	1378	24.6
X	851	15.2
L	868	15.5
W	168	3
P	683	12.2
R	96	1.7
S	79	1.4

3.2. Error Detection

The language error detection problem is mostly considered as a sequence labeling task where a supervised learning approach is adopted to predict whether the input sequence is grammatically correct or not. Most of the existing studies use one out of two approaches to convert an English sentence into a sequence for the classification task. In the first approach, the sentence is processed as a sequence of words as they appear in the text, i.e. as a lexical sequence. For example, the sentence *"I am reading a book"* will be transformed into the sequence *<I> <am> <reading> <a> <book>*. In the second approach, a sentence is converted into the sequence of tokens which indicate its structural or syntactic information, i.e. a syntactic sequence. For example, the syntactic sequence of the same sentence will be ***<subject> <helping – verb> <verb> <article> <object>***. The syntactic sequences are mostly built on a basis of the output of a dependency parser and POS tagger and carry the structural information of a sentence.

Granger states that in order to be fully effective, an error annotation system should be:

1. informative but manageable: it should be detailed enough to provide useful information on learner errors, but not so detailed that it becomes unmanageable for the annotator;

2. reusable: the categories should be general enough to be used for a variety of languages;

3. flexible: it should allow for addition or deletion of tags at the annotation stage and for quick and versatile retrieval at the postannotation stage; and

4. consistent: to ensure maximum consistency between the annotators, detailed descriptions of the error categories and error tagging principles should be included in an error tagging manual [23].

Correct forms were also inserted in the text files next to the erroneous forms (a) to facilitate subsequent interpretation of the error annotations; and (b) to allow for automatic sorting on the correct forms.

3.2.1. Grammar-Based Errors

The errors found on the grammatical level are summarized in the following table:

Error type	**Description/Example**
Verbs	
Verb tense	You have to buy food by yourself, pay for flat and another things that your parents **[do - did]** before.
Verb modal	I believe, that it **[need to - will/should]** be quite easy

Missing verb	I also remember how **[worried - worried I was]** before the exam.
Verb form	Can you imagine, big town, so much interesting people, different meals and another wonderful things, that I have never **[try - tried]** before.
Subject-verb-agreement	On my opinion this a book **[have - has]** wonderful plot and author **[have - has]** strange imagination
Articles/determiners	
Missing article	I think it's very important to **[have good - have a good]** travelling companion.
Incorrect use of article	I would advice you to visit Paris and acquainted with **[a culture - the culture]**, good places and kind people
Incorrect use of determiner	But on **[the another - the other]** hand life give you many responsibilities
Nouns	
Noun number	On my opinion studying abroad has so many **[advantage and disadvantage - advantages and disadvantages]**.
Noun possessive	You live without **[parents - parents']** control, come home when you want and do everything you can`t do before.

Pronouns	
Pronoun form	Of course, they can earn money from fans on Patreon, but the main source of money for **[they - them]** remain legal internet stores and shops.
Pronoun reference	To avoid quarrel with such people, you need to talk to **[him - them]** before the trip.
Word choice	
Wrong collocation	Do sport, eat vegetables and fruits, sleep enough to **[recover - improve]** your health.
Wrong preposition	Night lights, the beauty of main streets, great parks and high buildings will probably stun you **[in - at]** first
Word form	You need to go to the doctor and **[consultant - consult]** with him.
Sentence Structure	
Comma splice	At our school, we did distance learning for a month, public transport was also canceled.
Parallelism	Despite on town's attractions, I could to add some possibilities for employment and career growth, in order to get more money for **[developing and build - developing and building]** our town.

Fragment	In which you can climb mountains and swim in the sea and go on a lot of excursions.
Word order	
Incorrect sentence form	From time to time procrastination take me over and I know **[how is it hard - how hard it is]** to work, when you must complete it, but you can`t.
Adverb/adjective position	I saw some photos of this place on the internet before my trip and that all really interest me, **[these lamps, style of buildings and restaurants especially - especially these lamps...]**
Transitions	
Linking words/phrases	But I advise you not to worry **[because - that]** you can write the exam badly.
Mechanics	
Punctuation, capitalization, spelling	**[Unfurtanetly - unfortunarely]**, I had to stop my "studies" for some time because for getting a good result you need to know at least one native speaker of this language
Others	
Other errors	Any error that does not fit into any other category, but can still be corrected

Unclear meaning	The quality of the passage is so poor that it cannot be corrected.

3.2.2. Formal Errors

Our system consisted of four categories:

1) omission {-} - a missing item (a word or a group of words) which would have appeared in a well-formed sentence;
2) addition {+} - a redundant item which would not have appeared in a well-formed sentence;
3) misinformation {#} - a mechanical error that involved the use of the incorrect form of morpheme (e.g. an incorrect past tense form of a verb), or the cases when the selection of the incorrect item entailed a more complex conceptual judgement (e.g. the incorrect choice of tense/aspect);
4) misordering {[]}- the incorrect placement of an item in a sentence.

Frequencies of errors according to the type

Type of error	det	noun	prn	adv	adj	be	verb	prp	modal	to	conj	Total
Addition	32	39	19	36	7	29	79	30	17	12	6	306
Omission	194	154	61	52	44	56	103	131	14	32	16	857
Misinformation	47	224	105	62	64	134	595	38	11	7	16	1302
Misordering	4	9	3	5	4	2	5	3	0	1	2	38
												2503

In terms of the number of error tags, misformation errors were found to be most frequent (52%), followed by omission errors (34%), addition errors

(12%) and misordering errors (2%). Overall, noun and verb errors are very frequent, followed by determiner errors. This has to be interpreted with caution because the total number of occurrences of nouns and verbs is usually greater than the other parts of speech.

Determiner errors are especially frequent in the case of omissions. The frequencies of omission errors are five to six times higher than addition errors, which shows that Ukrainian learners of English tend to omit determiners rather than oversupply them. Prepositions are also problematical and they are frequently omitted.

The top ten most frequent error features were:

Error feature	**Number of occurrences**	**% out of all errors**
missing definite article	176	7%
singular noun for plural	152	6%
redundant definite article	148	5.9%
misselection of preposition	107	5.2%
lexical misconception	87	4.2%
incorrect tense and aspect	66	3.2%
S-V non-agreement	41	2%
incorrect collocation	36	1.8%
missing indefinite article	35	1.7%
definite article for indefinite article	34	1.7%

The conducted analysis allowed us to develop the taxonomy which we had taken as a basis for our research, highlighting the way the structure of a sentence is altered:

1) Omission errors:

a) <u>omission of grammatical/function words</u>: the article (*a, an, the*), verb auxiliaries (*is, will, can, may,* etc) and prepositions (*in, on, or*, etc);

b) <u>omission of content words</u>.

Language learners omit grammatical words much more frequently than content words. If content words are omitted in L2, it is usually caused by lack

of vocabulary, and learners usually indicate their awareness of the missing constituent.

2) Addition errors:

a) double markings: the error which is caused by the failure to delete certain items which are required in some linguistic construction (e.g. *She didn't went/goed back*);

b) regularization: the type of errors in which a marker that is typically added to a linguistic item is erroneously added to exceptional items of the given class that do not take a marker (e.g *sheeps* instead of *sheep*);

c) simple addition: the type of errors which characterize all addition errors. It is the use of an item which should not appear in well-formed utterances (e.g *the fishes doesn't live in the water*).

3) Misformation errors:

a) regularization errors: the errors that fall under the misformation category are those in which a regular marker is used in place of an irregular one (e.g *runned* instead of *run*);

b) archi-forms: the selection of one number of a class of forms to represent others in the class (e.g *I see her yesterday. Her dance with my brother*);

c) alternating forms: the errors caused by the learners' vocabulary and grammar development (e.g *I seen her yesterday*).

4) Misordering errors.

Error category	**Error subcategory**	**Frequency**	**Percentage**
Omission	Omission of grammatical/ function words	223	8.9%
	Omission of content words	634	25.3%
Addition	Regularization	59	2.4%

	Double marking	31	1.2%
	Simple addition	216	8.6%
Misinformation	Regularization	303	12.1%
	Archi-forms	420	16.8%
	Alternating forms	579	25.1%
Misordering		38	1.5%

In general, there is no established view on whether automatic error detection is at all viable. When learner writing samples have not previously been annotated for grammatical errors, each error flag produced by the system must be evaluated, or verified, manually.

Manual error detection and classification, however, poses a question of how reliably human annotators can agree on whether a word or sentence is grammatically correct. This is especially important where lexical errors are concerned, as it is up to the annotator to decide what is right and wrong.

Another issue to be taken into account is the notion of consistency. As we have already mentioned, in some aspects the plurality of analyses is possible for any given annotation. Therefore, we cannot be absolutely sure that for the same set of decision-making conditions the interpretation the human annotator imposes upon a text will be the same for different parts of the corpus.

One way to improve the reliability of manual evaluation is to have multiple raters evaluate the same error flags and then discuss discrepancies in judgments and produce a gold standard adjudicated set.

CONCLUSIONS TO CHAPTER 3

The third chapter of our chapter was devoted to the attempt of error tagging and analysis of error distribution in the compiled corpus. We have reviewed the existing error taxonomies, having chosen for the aims of our research the taxonomy that distinguishes between grammar-based and formal error types. The tagset we used for error annotation consisted of 52 tags belonging to 7 categories: form, grammar, lexicogrammar, lexis, word repetition, omission or word order, register and style. For each of the categories we estimated the frequency rate and the most common error types.

In the process of error annotation we used automatic error detection tools in line with manual error detection. Each instance of learner errors was indicated in the text and assigned a tag containing the information about the error category and type as well as the corrected form for the cases where it was possible to establish a target hypothesis.

The errors then were processed and analyzed in accordance with the chosen error taxonomy. On the grammatical level the errors were classified according to the linguistic category; on the formal level we differentiated between omission, addition, misinformation and misordering errors, having established the subcategories for each error type.

The analysis we have conveyed allows us to state that the existing NLP tools are not yet sufficiently accurate so as to allow fully automated annotation. If the researcher also wants to retrieve clean data and for this purpose, the automated tagging needs to be manually checked or at least, the researcher should know what is the percentage of error and how much additional manual checking they need to do.

CONCLUSIONS

In accordance with the aim of the research we have analysed the main methods and tools of corpus linguistics focusing on the potential of automatic corpus annotation tools when applied to a compiled learner corpus. For the aims of our research we have compiled a corpus of written English texts produced by Ukrainian speakers of different proficiency level and language learning background. The texts were then annotated for parts of speech, syntactic dependencies and types of errors using the combination of automatic and manual annotation.

In the process of analysis we have developed an algorithm of non-native text annotation, which allows the researcher to make use of the existing automatic annotation tools most efficiently and reduce the need for manual annotation and correction.

First, before the part-of-speech tagging and syntactic dependency annotation itself the text is checked and annotated for spelling errors, as our analysis has shown that the latter have the greatest impact on POS tagger performance. Current spell checkers have shown to have a high accuracy rate on non-native texts; they don't require any manual performance review for higher proficiency levels and for lower levels reviewing the output of the spell checker doesn't take much time and effort. Therefore, they can be widely employed for such tasks. However, in order to preserve the original features of the text, which might be of scientific interest in the first place, we suggest that the correct spelling option be provided as an error tag and in the process of POS tagging the input should be retrieved from the tags themselves without making any changes to the original structure of the text.

Second, the text is annotated for parts of speech. Having established that in most cases tagging errors were caused by spelling errors, we might suggest that after the text has been preprocessed with the help of spell checkers, the

output of an automatic POS tagger can be considered accurate enough. As for the cases when the learner erroneously uses one part of speech instead of the other, we believe that they should be classified according to the learner's word choice and in case of discrepancies between the output of POS tagger and dependency parser or in cases where the latter establishes certain dependency models which are not typical of the English language, such discrepancies can be regarded as a basis for error detection.

Third, the text is annotated for syntactic dependencies. Here, a more thorough review might be needed, as the performance of the parser is highly sensitive to the errors in word order and verbal forms. However, manual review of the parser output spares the annotator's time and effort and allows for relative consistency in the annotation process.

Fourth, the text is annotated for errors based on the output of the POS tagger and syntactic parser. Each error is assigned a tag containing the information about the type and subcategory of the error as well as its possible correction. The challenges we faced here include the subjectivity of annotator decisions, especially in terms of establishing a target hypothesis, and overlapping of certain error categories.

Therefore, having analysed the most common learner mistakes we made an attempt to establish more rigid boundaries between the error categories and subcategories. For each error detection level - grammar-based vs formal errors - we have developed a broader taxonomy, having established the frequency of each error type and its influence on the performance of automatic annotation tools. Being compared with the corresponding results for other learner corpora as well as native language corpora, these data can become the basis for extracting the distinctive features of English texts produced by Ukrainian speakers. Therefore, we view the potential for further research in the contrastive analysis of the texts produced by language learners from different linguistic and ethnic backgrounds and developing NLP tools for native

language identification based on error distribution using machine learning algorithms.

REFERENCES

1. Atkins S. Corpus design criteria. / S. Atkins, J. Clear, N. Ostler // Literary and linguistic computing 7.1 / S. Atkins, J. Clear, N. Ostler., 1992. – C. 1–16.
2. Automated Grammatical Error Detection for Language Learners. / C.Leacock, M. Chodorow, M. Gamon, J. Tetreault. – Morgan & Claypool Publishers, 2010.
3. Belz J. The pedagogical mediation of a developmental learner corpus for classroom-based language instruction. / J. Belz, N. Vyatkina., 2008.
4. Biber D. Corpus linguistics: Investigating language structure and use. / D. Biber, S. Concrad, R. Reppen.. – Cambridge: Cambridge University Press., 1998.
5. Bird S. Natural language processing with Python / S. Bird, E. Klein, E. Loper. – Sebastopol: O'Reilly, 2009.
6. Botley S. Multilingual corpora in teaching and research / S. Botley, A. McEnery, A. Wilson., 2000. – (Rodopi).
7. Brooke J. Robust, lexicalized native language identification. / J. Brooke, G. Hirst // In Proceedings of the 24th International Conference on Computational Linguistics / J. Brooke, G. Hirst. – Mumbai, India: The COLING 2012 Organizing Committee, 2012. – C. 391–408.
8. Buchholz S. CoNLL-X shared task on multilingual dependency parsing / S. Buchholz, E. Marsi // Proceedings of the tenth conference on computational natural language learning (CoNLL-X) / S. Buchholz, E. Marsi., 2006.
9. Building a Large Annotated Corpus of English: The Penn Treebank", Computational Linguistics Journal / [P. Mitchell, B. Santorini, M. Mary Ann та ін.]., 1994.

10. Carlson A. Scaling up context-sensitive text correction. / A. Carlson, J. Rosen, D. Roth // In Proceedings of the 13th Conference on Innovative Applications of Artificial Intelligence / A. Carlson, J. Rosen, D. Roth., 2001. – C. pages 45–50.
11. Centre for English Corpus Linguistics. Learner Corpora around the World. [Електронний ресурс] // Louvain-la-Neuve: Université catholique de Louvain. – Режим доступу до ресурсу: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.
12. Cognate and misspelling features for natural language identification / [N. Garrett, H. Bradley, M. Salameh et al.] // In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications / [N. Garrett, H. Bradley, M. Salameh et al.]. – Atlanta, GA, USA: ACL, 2013. – C. 140–145.
13. Common European Framework of Reference for Languages: Learning, teaching, assessment. [Електронний ресурс] // Cambridge: Cambridge University Press. – 2001. – Режим доступу до ресурсу: www.coe.int/lang.
14. Dahlmeier D. Grammatical error correction with alternating structure optimization / D. Dahlmeier, H. T. Ng // In Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies / D. Dahlmeier, H. T. Ng. – Stroudsburg, PA: Association for Computational Linguistics, 2011. – C. 915–923.
15. Dash N. History, Features, and Typology of Language Corpora. / N. Dash, S. Arulmozi., 2018.
16. Díaz-Negrillo A. A tagging tool for error analysis on learner corpora / A. Díaz-Negrillo, M. García-Cumbreras. // Internation Computer Archive of Modern and Medieval English (ICAME) Journal. – 2007. – №31. – C. 197–203.

17. Garrett N. Cognate and misspelling features for natural language identification / Nicolai Garrett // Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications / Nicolai Garrett., 2013.
18. Geertzen J. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT) / J. Geertzen, T. Alexopoulou, A. Korhonen // Proceedings of the 31st Second Language Research Forum. / J. Geertzen, T. Alexopoulou, A. Korhonen. – Somerville: Cascadilla Proceedings Project, 2013.
19. Gilquin G. Corpora and experimental methods: A state-of-the-art review. / G. Gilquin, S. Gries // Corpus Linguistics and Linguistic Theory / G. Gilquin, S. Gries., 2009. – C. p. 1–26.
20. Gilquin G. Errors and disfluencies in spoken corpora: Setting the scene. / G. Gilquin, S. De Cock // International Journal of Corpus Linguistics / 2011. – C. p. 141–172.
21. Gilquin G. Learner corpora. / Gaëtanelle Gilquin // Magali Paquot & Stefan Gries, A Practical Handbook of Corpus Linguistics / Gaëtanelle Gilquin., 2020. – C. pp. 283–303.
22. Golding A. A Bayesian hybrid method for context sensitive spelling correction. / Andrew Golding // In Proceedings of the Third Workshop on Very Large Corpora – 1995. – C. pages 39–53.
23. Granger S. Learner Corpora / S. Granger // The Encyclopedia of Applied Linguistics / S. Granger. – John Wiley & Sons, Ltd, 2019.
24. Gries S. Linguistic annotation in/for corpus linguistics / S. Gries, A. Berez // Handbook of Linguistic Annotation / S. Gries, A. Berez. – New York: Springer, 2017. – C. pp.379–409.
25. Heift T. Computer-assisted corrective feedback and language learning / T. Heift, V. Hegelheimer // Corrective feedback in second language

teaching and learning: Research, theory, applications, implications 66 / T. Heift, V. Hegelheimer., 2017. – С. 51–65.

26. Hirschmann H. Syntactic annotation of non-canonical linguistic structures / H. Hirschmann, S. Doolittle, A. Lüdeling., 2007.
27. Hirst G. Correcting real-world spelling errors by restoring lexical cohesion / G. Hirst, A. Budanitsky // Natural Language Engineering – 2005. – С. 87–111.
28. Hunston S. Corpora in Applied Linguistics / S. Hunston. – Cambridge: Cambridge University Press, 2002.
29. Intelligent CALL / [M. Schulze, T. Heift, M. Thomas та ін.]., 2013.
30. James C. Errors in Language Learning and Use. Exploring Error Analysis. / C. James. – London: Longman, 1998.
31. Jarvis S. Native language identification / S. Jarvis, M. Paquot // The Cambridge Handbook of Learner Corpus Research. / S. Jarvis, M. Paquot. – Cambridge: Cambridge University Press, 2015.
32. Kennedy G. An introduction to corpus linguistics. / G. Kennedy. – London: Longman, 1998.
33. Krivanek J. Comparing rule-based and data-driven dependency parsing of learner language. / J. Krivanek, D. Meurers.. – 2011 с.
34. Leech G. Introducing corpus annotation / G. Leech, R. Garside, T. McEnery // Corpus Annotation. Linguistic Information from Computer Text Corpora. / G. Leech, R. Garside, T. McEnery. – London: Longman, 1997. – С. 1–18.
35. Lingzhen C. Improving native language identification by using spelling errors. / C. Lingzhen, C. Strapparava, V. Nastase // In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics / C. Lingzhen, C. Strapparava, V. Nastase. – Vancouver, Canada: ACL, 2017. – С. 542–546.

36. Lu X. Computational Methods for Corpus Annotation and Analysis / X. Lu., 2014.
37. Lüdeling A. Corpus Linguistics. An International Handbook. / A. Lüdeling, M. Kytö. – Walter de Gruyter, 2008. – 776 c.
38. Malmasi S. Oracle and human baselines for native language identification / S. Malmasi, J. Tetreault, M. Dras // In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications / S. Malmasi, J. Tetreault, M. Dras. – Denver, CO, USA: ACL, 2015. – C. 172–178.
39. Manning C. Foundations of statistical natural language processing / C. Manning, H. Schutze., 1999. – (MIT press).
40. McEnery T. Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. / T. McEnery, V. Brezina, D. Gablasova // Annual Review of Applied Linguistics / T. McEnery, V. Brezina, D. Gablasova., 2019. – C. 74–92.
41. McEnery T. Corpus linguistics / T. McEnery, A. Wilson. – Edinburgh: Edinburgh University Press Ltd, 2001.
42. Meurers D. Compiling a task-based corpus for the analysis of learner language in context." / D. Meurers, N. Ott, R. Ziai // Proceedings of Linguistic Evidence / D. Meurers, N. Ott, R. Ziai. – Tübingen, 2010.
43. Meurers D. Natural language processing and language learning. / Detmar Meurers // The encyclopedia of applied linguistics / Detmar Meurers, 2012.
44. Mitkov R. The Oxford Handbook of Computational Linguistics. / Mitkov. – Oxford: Oxford University Press, 2003.
45. Moshe K. Determining an author’s native language by mining a text for errors. / K. Moshe, S. Jonathan, Z. Kfir // In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in

Data Mining / K. Moshe, S. Jonathan, Z. Kfir. – New York, NY, USA, 2005. – (ACM). – С. 624–628.

46. Mukherjee J. Rethinking applied corpus linguistics from a language pedagogical perspective / J. Mukherjee, J. Rohrbach // New departures in learner corpus research / J. Mukherjee, J. Rohrbach., 2006.
47. Multi-level error annotation in learner corpora / A.Lüdeling, W. Maik, E. Kroymann, P. Adolphs // The Corpus Linguistics Conference – The Corpus Linguistics, 2005.
48. NLTK. The Natural Language Toolkit [Електронний ресурс] / NLTK – Режим доступу до ресурсу: http://nltk.sourceforge.net/index.html.
49. Nesselhauf N. Learner corpora and their potential for language teaching / Nadja Nesselhauf // How to use corpora in language teaching 12 / Nadja Nesselhauf., 2004. – С. 125–156.
50. O'Keeffe A. The Routledge handbook of corpus linguistics. / A. O'Keeffe, M. McCarthy., 2010
51. Odlin T. Language transfer: Crosslinguistic Influence in language learning. / T. Odlin. – Cambridge: CUP, 1989.
52. O’Keeffe A. The Routledge handbook of corpus linguistics / A. O’Keeffe, A. McCarthy. – London: Routledge, 2010.
53. Rankin T. Marginal prepositions in learner English / T. Rankin, B. Schiftner // Applying local corpus data / T. Rankin, B. Schiftner., 2011. – С. 412–434.
54. Robinson J. Dependency structures and transformational rules. / Jane J Robinson // Language / Jane J Robinson., 1970. – С. 259–285.
55. Rosen A. Compiling and Annotating a Learner Corpus for a Morphologically Rich Language. / A. Rosen..
56. Rozovskaya O. Annotating ESL errors: Challenges and rewards. / O. Rozovskaya, D. Roth. // In Proceedings of the Fifth Workshop on

Innovative Use of NLP for Building Educational Applications. – 2010. – С. 28–36.

57. Semantic role parsing: Adding semantic structure to unstructured text / [S. Pradhan, K. Hacioglu, W. Ward та ін.] // In Third IEEE International Conference on Data Mining / [S. Pradhan, K. Hacioglu, W. Ward та ін.]., 2003. – С. pp. 629–632.
58. Shibamouli L. Using n-gram and word network features for native language identification. / L. Shibamouli, R. Mihalcea // Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications / L. Shibamouli, R. Mihalcea., 2013.
59. Supervised machine learning: A review of classification techniques / Kotsiantis, B. Sotiris, I. Zaharakis, P. Pintelas // Emerging artificial intelligence applications in computer engineering 160.1 / Kotsiantis, B. Sotiris, I. Zaharakis, P. Pintelas., 2007. – С. 3–24.
60. Swanson B. Exploring syntactic representations for native language identification / Ben Swanson // Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications / Ben Swanson., 2013.
61. Syntax Tree Generator [Електронний ресурс] – Режим доступу до ресурсу: http://mshang.ca/syntree/.
62. Tetreault J. Native tongues, lost and found: Resources and empirical evaluations in native language identification / J. Tetreault, D. Blanchard, A. Cahill // In Proceedings of the 24th International Conference on Computational Linguistics / J. Tetreault, D. Blanchard, A. Cahill. – Mumbai, India: The COLING 2012 Organizing Committee., 2012. – С. pages 2585–2602.
63. Teubert W. Corpus linguistics: A short introduction. / W. Teubert, A. Čermáková. – London: Continuum, 2007.

64. The International Corpus of Learner English. / S.Granger, F. Dagneaux, E. Meunier, M. Paquot. – Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain, 2002.
65. Tono Y. Learner corpora: design, development and applications / Y. Tono. // Proceedings of the Corpus Linguistics 2003 Conference. – 2003. – C. 800–809.
66. Tsur O. Using classifier features for studying the effect of native language on the choice of written second language words. / O. Tsur, A. Rappoport // In Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition / O. Tsur, A. Rappoport. – Stroudsburg, PA, USA: ACL, 2007. – C. 9–16.
67. Using the Web for language independent spellchecking and autocorrection. / C.Whitelaw, B. Hutchinson, G. Chung, G. Ellis // In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) / C.Whitelaw, B. Hutchinson, G. Chung, G. Ellis. – Singapore, 2009. – C. pages 890–899.
68. Van Halteren H. Automatic taggers: an introduction. / H. Van Halteren, A. Voutilainen // Syntactic Wordclass Tagging / H. Van Halteren, A. Voutilainen. – Dordrecht: Kluwer, 1999. – C. pp. 109–115
69. Van Rooy B. Annotating learner corpora. / B. Van Rooy // The Cambridge Handbook of Learner Corpus Research. – Cambridge: Cambridge University Press, 2015.
70. Van Rooy B. The effect of learner errors on POS tag errors during automatic POS tagging / B. Van Rooy, L. Schäfer // In Southern African Linguistics and Applied Language Studies / B. Van Rooy, L. Schäfer., 2002. – (20). – (4). – C. 325–335.
71. Voutilainen A. Does tagging help parsing? A Case Study On Finite State Parsing / Atro Voutilainen. – Finland: University of Helsinki.

72. Wong S. Exploiting parse structures for native language identification / S. Wong, M. Dras // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing / S. Wong, M. Dras., 2011.
73. Wynne M. Developing Linguistic Corpora: a Guide to Good Practice. / M. Wynne. – Oxford: Oxbow Books, 2005.
74. Xiaofei L. Automatic analysis of syntactic complexity in second language writing / L. Xiaofei // nternational journal of corpus linguistics 15.4 / L. Xiaofei., 2010. – C. 474–496.
75. Čermáková A. Directions in corpus linguistics. / A. Čermáková, T. Wolfgang / Lexicology and corpus linguistics / A. Čermáková, T. Wolfgang., 2004. – C. 113–166.