UNIWERSYTET IM. ADAMA MICKIEWICZA

WYDZIAŁ NEOFILOLOGII

Lexical convergence in Polish dialogues

Konwergencja leksykalna w dialogach w języku polskim

mgr Karolina Jankowska

Rozprawa doktorska napisana pod kierunkiem
Prof. zw. dr hab. inż. Grażyny Demenko

Poznań, 2023

Acknowledgements

The research presented in this thesis was carried out within the scope of the National Science Centre project *Automatic Analysis of Phonetic Convergence in Speech Technology Systems* (no. 2014/14/M/HS2/00631) directed by prof. Grażyna Demenko.

Abstract (English)

Convergence in communication concerns the assimilation of verbal and non-verbal behaviours during interpersonal interactions. This phenomenon has been studied at the linguistic, paralinguistic and non-linguistic levels in various languages. Materials in Polish were used for research focused mainly on phonetic and acoustic parameters, which confirmed the convergence in dialogues at this level. Far too little knowledge is available on lexical convergence in Polish dialogues.

The literature contains numerous examples of the methodologies, parameters and measures used that enabled the analysis of linguistic materials in terms of linguistic convergence in dialogues. The analyses were based on acoustic-prosodic features of speech, e.g. speech rate, pauses, fundamental frequency (F0), etc., lexical features such as utterance length, lexical variety, lexical adjustment (reuse of words from the previous utterance of the interlocutor). There are also complex methodologies for examining the level of convergence between interlocutors, such as *Language Style Matching*[1], which are based on the analysis of non-content words.

The theses of the dissertation assume the occurrence of the lexical convergence phenomenon in dialogues in the Polish language, observable in the choice of vocabulary at the level of parts of speech, the change in the intensity of lexical alignment depending on the communication situation, the adjustment of the degree of formality of the language utterance to the interlocutor and the circumstances and differences in the length of dialogues depending on the task and goal. In order to verify the theses, an analysis of the level of formality of the language, choice of vocabulary and collaborative effort (level of involvement measured in the number of words spoken) of the interlocutors in the dialogues was carried out. A Language Style Matching analysis was also performed, the methodology of which was adapted to the Polish language. The research material was part of the Harmonia corpus created within the frames of *Automatic Analysis of Phonetic Convergence in Speech Technology Systems* project (no. 2014/14/M/HS2/00631). The recordings selected for the lexical convergence analysis were obtained from dialogues conducted according to a fixed scenario.

The choice of vocabulary (individual parts of speech and the level of formality of the language) changes with the topic, purpose and partner of the conversation. Depending on the task according to which the participants conducted dialogues, the share of individual parts of speech varied. Based on the conducted research, it can be concluded that the interlocutors use polite forms in dialogues with a teacher and when playing the roles of strangers. In the diapix[2] task, partners used the most nouns and adjectives and matched the descriptions of the objects in the picture. Once used, the term was usually repeated by both interlocutors in a dialogue. In the map tasks, the interviewees tended to duplicate terms describing directions and ways of moving. The interlocutors created a distance suitable for strangers and used polite forms as well as official greetings and farewells in tasks where one interlocutor impersonated a tourist and the other impersonated an employee of a tourist information office. In tasks where participants were to express opinions on controversial art, an increase in the share of particles, conjunctions and pronouns in the statements was noticed. Dialogues in which participants agreed on the provocative work showed a lower level of vocabulary repetition than in dialogues in which participants disagreed. The results of the Language Style Matching index calculations were the highest on average (0,717) in student dialogues in which participants unanimously criticised provocative art. Slightly lower values occurred in student-teacher dialogues, in which the interlocutors were to praise provocative art (0,702) and express distinct opinions about it (0,707). On average, the lowest index values (0,611) occurred in the task of providing information to the tourist about events in the city. The results of the research confirmed the theses – lexical convergence occurs in dialogues in Polish in the presented aspects.

1 The technique of analysing stylistic similarities in the language and the linguistic match index.
2 Spot the difference task with a pair of pictures to induce spontaneous speech.

Abstract (Polish)

Konwergencja w komunikacji dotyczy upodabniania zachowań werbalnych i pozawerbalnych w trakcie interakcji interpersonalnej. Zjawisko to było badane na poziomie lingwistycznym, paralingwistycznym i pozalingwistycznym, w różnych językach. Na materiałach w języku polskim przeprowadzano badania skupione głównie na fonetycznych i akustycznych parametrach, które potwierdziły zjawisko konwergencji w dialogach na tym poziomie. Zdecydowanie zbyt mało wiedzy jest dostępnej w zakresie leksykalnej konwergencji w dialogach w języku polskim.

W literaturze można znaleźć liczne przykłady stosowanych metodologii, parametrów i miar, które umożliwiły analizę materiałów językowych pod względem konwergencji lingwistycznej w dialogach. Analizy opierały się na akustyczno-prozodycznych cechach mowy np. tempo mowy, pauzy, częstotliwość podstawowa (F0) itp., leksykalnych cechach takich jak długość wypowiedzi, różnorodność leksykalna, dopasowanie leksykalne (ponowne użycie słów z poprzedniej wypowiedzi rozmówcy). Istnieją również złożone metodologie badania poziomu konwergencji między rozmówcami takie jak *Language Style Matching*[3], które opierają się na analizie wyrazów niesamodzielnych znaczeniowo.

Tezy dysertacji zakładają występowanie zjawiska konwergencji leksykalnej w dialogach w języku polskim widocznej w doborze słownictwa na poziomie części mowy, zmianę intensywności dostosowania leksykalnego w zależności od sytuacji komunikacyjnej, dostosowanie stopnia formalności wypowiedzi języka do rozmówcy i okoliczności oraz różnice w długości dialogów w zależności od zadania i celu rozmowy. W celu weryfikacji postawionych tez przeprowadzono analizę poziomu formalności języka, doboru słownictwa oraz udziału (poziomu zaangażowania mierzonego w liczbie wypowiadanych słów) rozmówców w dialogach. Wykonano również analizę *Language Style Matching*, której metodologia została dostosowana do języka polskiego. Materiałem badawczym była część korpusu Harmonia stworzonego w ramach projektu *Automatic Analysis of Phonetic Convergence in Speech Technology Systems* (no. 2014/14/M/HS2/00631). Nagrania wybrane do przeprowadzenia analizy konwergencji leksykalnej uzyskano z dialogów przeprowadzonych z ustalonym scenariuszem.

Dobór słownictwa (poszczególne części mowy oraz poziom formalności języka) zmienia się wraz z tematem, celem i partnerem rozmowy. W zależności od zadania, zgodnie z którym uczestnicy prowadzili dialogi, zmieniał się udział poszczególnych części mowy. Na podstawie przeprowadzonych badań można stwierdzić, że rozmówcy stosują formy grzecznościowe w dialogach z osobą starszą w relacji (w tym przypadku student – nauczyciel) oraz w sytuacji odgrywanych ról osób nieznajomych. W zadaniu diapix[4] partnerzy używali najwięcej rzeczowników i przymiotników oraz dopasowywali opisy obiektów widocznych na obrazku. Raz użyty termin był zwykle powtarzany przez obu rozmówców w dialogu. W zadaniach z mapą, rozmówcy wykazywali tendencję do powielania określeń opisujących kierunki i sposób poruszania się. Rozmówcy tworzyli dystans odpowiedni dla nieznajomych i stosowali formy grzecznościowe oraz oficjalne powitania i pożegnania w zadaniach, w których jeden rozmówca wcielał się w turystę, a drugi w rolę pracownika biura informacji turystycznej. W zadaniach, w których uczestnicy mieli wyrażać opinie na temat kontrowersyjnej sztuki zauważono wzrost udziału partykuł, spójników i zaimków w wypowiedziach. W dialogach, w których rozmówcy zgadzali się ze sobą w ocenie prowokacyjnego dzieła, zauważono niższy poziom powielania słownictwa niż w dialogach, w których uczestnicy mieli odmienne zdania. Wyniki obliczeń wskaźnika *Language Style Matching* były najwyższe średnio najwyższe (0,717) w dialogach studentów, w których uczestnicy zgodnie krytykowali prowokacyjną sztukę. Nieco niższe wartości wystąpiły w dialogach studentów z nauczycielem, w których rozmówcy mieli zgodnie pochwalić prowokacyjną sztukę (0,702) oraz wyrażać odmienne zdanie na jej temat (0,707). Średnio najniższe wartości wskaźnika (0,611) wystąpiły w zadaniu polegającym na udzielaniu informacji turyście o wydarzeniach w mieście. Wyniki badań potwierdziły postawione tezy – konwergencja leksykalna występuje w dialogach w języku polskim.

3 Technika analizy podobieństw stylistycznych w języku oraz wskaźnik dopasowania lingwistycznego.
4 Zadanie „znajdź różnicę" z parą obrazków, służące do wywoływania spontanicznej mowy.

Table of content

Definitions

Annotation	Adding relevant information in the form of tags to text segments. In the case of this work, annotation consisted in determining the parts of speech of individual words.
Communication alignment	Adapting communication behaviours, both verbal and non-verbal, to the communication situation and the interlocutors.
Diapix	*Diapix is a problem-solving 'spot the difference' picture task used for eliciting spontaneous speech interactions between two participants* (Van Engen, et al., 2010).
Harmonia Corpus	A linguistic corpus in Polish consisting of recordings of individual speeches and dialogues, implemented in accordance with the scenario of sixteen tasks.
Language Style Matching (LSM)	A technique for analysing stylistic similarities in the language of different groups and individuals and an indicator of interpersonal and group alignment.
Language technology	A scientific area dealing with computer language processing and methods of analysing, creating, modifying or responding to human texts and speech.
Lemmatization	The process of reducing to the basic form (word's lemma or dictionary form) inflected forms of a word.
Lexical mimicry	The activity of copying the vocabulary used by the interaction partner in the conversation.
Paralanguage	*A term used in suprasegmental phonology to refer to variations in tone of voice (…) Examples of paralinguistic features would include the controlled use of breathy or creaky voice, spasmodic features (such as giggling while speaking), and the use of secondary articulation (such as lip-rounding or nasalization) to produce a tone of voice (…)* (Crystal, 2008).
Prosody	*A term used in suprasegmental phonetics and phonology to refer collectively to variations in pitch, loudness, tempo and rhythm* (Crystal, 2008).

Acronyms

CAT	Communication Accommodation Theory
F0	Fundamental frequency
IAM	Interactive Alignment Model
IPIPAN	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Eng. Institute of Computer Science Polish Academy of Sciences)
LFPC	Log-Frequency Power Coefficients
LILLA	Lexical Indiscriminate Local Linguistic Alignment
LLA	Local Linguistic Alignment
LPC	Linear Prediction Cepstral
LSM	Language Style Matching
LTAS	Long-Term Average Spectra
MFCC	Mel Frequency Cepstral Coefficient
NKJP	Narodowy Korpus Języka Polskiego (Eng. The National Corpus of Polish)
POS	Part of Speech
SAT	Speech Accommodation Theory
SILLA	Syntactic Indiscriminate Local Linguistic Alignment
TEO	Teager Energy Operator
VOT	Voice Onset Time

I Introduction

Language is the most natural form of communication for people, therefore it plays a fundamental role in the exchange of information, transfer of knowledge on sociological, historical, cultural topics, traditions, customs and values important for human heritage. Therefore, it is a carrier of information used, in various forms, by all people. Researchers analyse language for a number of reasons and for the needs of various scientific fields. The analysis covers both verbal and nonverbal communication. Along with the linguistic analysis comes the analysis of body movements, which can also be an additional means of communication supporting the message and common understanding. The analysis of the entire, complex communication process is a powerful tool for understanding various psychological, biological and physical phenomena in the human body as well as sociological, cultural and political processes.

Communication is a process that requires key elements - *sender*, *message*, *receiver* (Jakobson, 1960) and can be defined as a variety of behaviours, including the interactions of living organisms and objects, such as computers (Kimura, 1993). Referring to human beings, the basic means of communication is language in spoken and written form. In a general sense, people communicate through a process of *encoding* and *decoding* - a sender turns thoughts into communication means and a receiver processes this code into thoughts, which is a conscious, purposeful action. In the course of verbal communication, a person is not able to express all the details and nuances in relation to the conveyed meanings. This enables the addressee to introduce various changes and additions, consistent with the knowledge, intentions or context, which means that each time it leads to the reconceptualization of the original meanings emerging as the addressee's interpretation of the utterance (Lewandowska-Tomaszczyk & A. Wilson, 2009). However, not only the spoken or written words can deliver information. Considering the process that must take place in the human body for the speech sound to be produced, it can be concluded that from the speech signal physical and biological

characteristics of the sender can be examined and evaluated. With the prosodic and acoustic features of speech, it is possible to determine the age, gender, mood and even diagnose various diseases speaker could be suffering. Many other fields of science make use of the linguistic and paralinguistic features extracted from speech, such as medicine, psychology, social sciences and information technologies. For the medical application, speech analysis is a very useful tool, which supports diagnosis and monitoring of the progress and advance of certain diseases. These include mental, neurodegenerative diseases and endocrinological disorders. Psychologists use speech signal to extract information about the emotional state, moods, stress, etc. which is strictly correlated with the social sciences, well-being and even longevity (Marmot, 2005). Furthermore, it is possible to recognise alcohol or drugs intoxication and fatigue via speech, which finds its application for example in intelligent vehicle and machines assistants (Levit, Huber, Batliner, & Noeth, 2001). Text analysis is an equally important research area within which the current trends move towards the analysis of Internet content in order to obtain information on opinions (marketing and political use) (Lia & Dash Wu, 2010), detection of fake news (Aldwairi & Alwahedi, 2018) and illegal activities on the Internet (Hernandez-Castro & L. Roberts, 2015). Scientists strive to create and improve technological systems of speech generation and recognition – effective and natural human-machine interfaces. There are many dialogue systems based on language processing, for example various types of chatbots, voicebots that perform tasks as customer service, help desk[5] or even psychological support[6]. Smart user interfaces can be implemented in a variety of software and hardware, which greatly facilitates the use and interaction between humans and machines.

Language production is therefore a very complex process involving many aspects of the speakers themselves as well as the interactions between them. In addition to the analysis of language production itself, an equally important, interdisciplinary aspect is the analysis of interactions and language

[5] i.e. https://talkie.ai/en/
[6] i.e. https://woebothealth.com/

changes that occur during the dialogue. One of the phenomena described in the literature is *convergence*, which concerns the process of alignment, becoming more similar to one another. Linguistic convergence refers to interlocutors' adjustment of communicative behaviour during interaction (Schweitzer & Lewandowski, 2013). This phenomenon is observable in many life situations, where one person adapts the language to the interlocutor in a more or less conscious way. When conversation partners know each other, they are able to use vocabulary and grammatical structures based on their assumptions that will facilitate understanding and all communication. An example would be a conversation between a doctor and a patient, during which a specific disease is discussed. The doctor is familiar with specialist terms to describe changes in the body, but usually assumes that the patient may not understand the message well if it contains medical terms, so they use general terms. This type of change, adapting the language to assumptions about the interlocutor, is a good example of lexical convergence. Another example, that can cover more communication-related aspects, is intercultural communication. When talking in the same foreign language, people from different parts of the world can use different concepts, gestures, facial expressions and adapt them to the interlocutor during the dialogue. These are examples of situations in which people largely consciously adapt the form of communication in order to increase its effectiveness. However, in everyday interactions there is also convergence or divergence, which is not always conscious and possible to observe perceptually. Based on previous research, it can be concluded that the phenomenon of convergence and divergence occurs between speakers during interaction on many levels. They are observed in interpersonal communication and in human-machine communication, covering many linguistic, paralinguistic and extralinguistic aspects (Pardo, Pellegrino, Dellwo, & Möbius, 2022). Convergence may concern the way of speaking, writing, choice of vocabulary, grammatical forms, gestures, facial expressions, etc., which will be described in detail later in this work.

The reasons or sources for this type of communication behaviour are not entirely clear, but there are theories that attempt to explain this phenomenon.

There are two main theoretical approaches: Interactive Alignment Model (IAM) (Pickering & Garrod, 2004) and Communication Accommodation Theory (CAT) (Giles H., 2008). The first theory assumes automatic, mechanical adjustment resulting from priming. The second, on the other hand, takes into account social closeness in the case of convergence and social distance in the case of divergence. These theories are still being researched, and the phenomenon of convergence and divergence is so complex and dependent on many psychological and social factors, individual characteristics of speakers and circumstances that there is a lot of room for further research in this area. The study of communication behaviour in terms of alignment will enable better understanding of the physiological, psychological and sociological mechanisms that occur during human-human interaction. Nowadays, when modern technologies enable conversations with computers in a natural language, an interesting topic is the analysis of human-computer conversations also in terms of adapting communication behaviour (Vinciarelli, et al., 2015). An in-depth knowledge in this area may enable modelling of dialogues and creating dialogue systems that will simulate communication in a way very similar to human-human communication.

1 Motivation

Research in the field of communication alignment is not a new topic in the literature, but there is still a lot of room for innovative research in this scientific field. Research methodologies are different, most often conducted and adapted to the English language. Relatively little research has been done on paralinguistic and linguistic convergence in the interaction of Polish speakers.

New technologies in the domain of intelligent systems based on natural language are dynamically developed for various languages, especially for English. However, systems based on human-machine communication in natural language are also being developed for the Polish language. Examples include voicebots, which are becoming more and more popular. For instance, the extraordinary epidemiological situation related to COVID-19 has created

a great need for prompt and reliable information about the symptoms and other related issues. Therefore, the Polish National Health Fund (NFZ) launched a telephone information hotline[7]. In March 2020, when there was little knowledge about the new virus and the number of cases was increasing across Europe, the hotline was heavily loaded. According to the website of the National Health Fund, the hotline was operated by over 300 people at the same time. Despite that, the lines to consultants were long, so the hotline was partially operated by voicebots. In the event of an accumulation of calls, the voicebots provided support in answering the most frequently asked questions, for which consultation with a specialist was not required.

Living in Poland, using various services, complete formalities at offices, it is easily noticeable that more and more companies and offices use voicebots as the first line of customer service and for marketing purposes. However, the quality of communication is not satisfactory and users usually quickly realise that they are talking to a machine and it is difficult for them to achieve their goals – to be correctly understood by the voicebot. Difficulties related to low quality of communication mean that people do not trust and are reluctant to use this type of technological solutions. Improving the quality of the systems may contribute to the increase in the achievement of communication goals and the satisfaction of using them. In order to improve the quality of systems based on natural language communication, basic research is needed that will bring closer the mechanisms and patterns in interpersonal interaction and enable modelling of natural human-computer dialogues. Due to the differences between languages related to grammar, syntax or prosody and its functions, similar studies should be carried out for individual languages. In this way, theoretical knowledge of natural communication can be acquired and used for modelling.

This work aims to take a step towards a better understanding of the mechanisms occurring during dialogues in terms of lexical selection in Polish. The research material is the Harmonia corpus created for the purposes of

[7] https://www.nfz.gov.pl/aktualnosci/aktualnosci-oddzialow/infolinia-nfz-10-tysiecy-polaczen-na-temat-koronawirusa,393.html

phonetic convergence research (see Chapter III.1.1.1) the results of which are described in the book *Phonetic Convergence in Spoken Dialogues in View of Speech Technology Application* (Demenko, 2020). The topic of lexical convergence is no less interesting and important, and at the same time rarely taken up in studies of the Polish language.

2 The thesis and objective

Taking into account the multidimensionality and importance of research on convergence in communication and development opportunities in interdisciplinary areas, an attempt was made to analyse lexical convergence in dialogues in Polish. The main theses of the work are:

- **lexical convergence in dialogues in Polish occurs in lexical choice at the part of speech level,**
- **the intensity of convergence and the level observable in the choice of vocabulary of parts of speech varies depending on the communicative situation,**
- **the interlocutors adjust the level of formality in the language to their own role, the interlocutor and the communication situation**
- **dialogues on provocative, emotional issues last longer on average than goal-oriented dialogues.**

It is suspected that in the dialogues from Harmonia corpus, apart from the phonetic convergence, there is also a lexical convergence observable. Lexical convergence, i.e. the similarity or repetition of words used by the interlocutor, can be observed in the selection of vocabulary from particular grammatical categories. For example, interlocutors can adjust their speech style in terms of the number of nouns and pronouns used. It is predicted that the study shall prove that in dialogues focused on completing a task (e.g. a task with a map, providing directions) more nouns and verbs will be used than in dialogues in which the interlocutors are to express their opinion on controversial topics that evoke emotions. It is suspected that Polish interlocutors use more pronouns in dialogues concerning their own opinions, feelings and emotions. Moreover, it is assumed that people tend to adapt their communication

behaviour to both the interlocutor and the communication situation. The study will focus on lexical analysis, which will also show whether the interlocutors adapt the level of formality in the language to the partner and the circumstances. It is suspected that interlocutors talk longer, that is, they say more words in conversations when they disagree. The desire to explain one's point of view to a person who has a different one and attempts to convince a conversation partner to other opinions involves the need to present arguments, which may be reflected in the length of dialogues and the number of words spoken. Similarly, in the case of conversations that are aimed at persuading someone to take a certain action or make a choice. Pardo et. al. (2017) concludes that low-frequency words evoke greater convergence than high-frequency words and that female talkers converge more than males. It is hypothesised that these tendencies will be noticeable in the dialogues of the Harmonia corpus.

The main purpose of this work and research is to check the level of lexical convergence in dialogues in Polish in various communication situations, where the variable is the subject and the roles of the participants in the study. The specific objectives are as follows:

- adaptation of the existing Harmonia language corpus for the purposes of lexical convergence research,
- analysis of lexical convergence in dialogues between people of the same sex, of similar age, of similar status (student) in various communication situations,
- analysis of lexical convergence in dialogues between people of the same sex, of similar age, of various status (student and teacher),
- use and evaluation of existing part-of-speech tagging programs for the Polish language,
- adaptation of the Language Style Matching methodology for the needs of the Polish language,
- the use of the adapted Language Style Matching methodology in the study of the convergence of dialogues in Polish.

Achieving the above-mentioned goals will enable the research and verification of the theses.

3 Overview

This work is divided into two main parts: theoretical and empirical. The first part includes an Introduction (current chapter) and a theoretical introduction to the subject of communication convergence. The second part of the work (empirical) describes the linguistic resource used in the study, methodology and results of own research.

Chapter II.1 describes theories explaining the psychological mechanisms that occur during verbal interaction between partners and lead to adjustment.

Chapter II.2 presents the levels of communication convergence in general and a detailed description of phonetic and lexical convergence. These chapters present the results of a literature review that sought to extract the methodologies and measures used in convergence research.

Chapter II.3 describes methodologies used in lexical convergence research with particular attention paid to Language Style Matching, which was adopted and used in own study.

Chapter II.4 outlines linguistic materials for computing in general. This chapter describes language corpora in various languages and Polish resources used in natural language processing.

Chapter II.5 contains a description of convergence research in Polish. In the first subchapter, examples of research on lexical convergence are cited, in the second – phonetic-acoustic convergence, and in the third – research on non-linguistic convergence with the participation of Polish speakers.

Chapter II.6 introduces the topic of language technology and contemporary challenges related to advanced systems based on dialogue systems and interfaces based on natural language, especially spoken language.

Chapter II.7 summarises the theoretical part of the dissertation, presents conclusion on the literature review.

Chapter III is the empirical part, which begins with Chapter III.1 describing the author's own research procedure. The research material and

methodology together with the features and measures used for each type of analysis, tools and annotation procedures are presented. For the purposes of this study, two POS tagging tools adapted to the Polish language were used. Subchapter III.1.1.3 describe the taggers and the quality of auto-tagging results with these tools. Subchapter III.1.1.4 contains information on manual annotation of the corpus used in the study.

Chapter III.2 presents the results of the lexical convergence analysis in terms of formality level, lexical choices, collaborative effort. In this chapter, the Language Style Matching methodology own adaptation to the Polish language and the results of LSM factor calculations are presented and discussed.

Chapter III.3. summarises the empirical part of the dissertation and discussed the results of the whole study.

Chapter IV summarises all the work described in this dissertation, addresses theses, the goals achieved, and the potential for exploiting the results and further developing research in the field.

The Appendix V.1 to the thesis presents a summary of review of linguistic corpora in various languages. Appendix V.2 provides scenarios for recordings used in this research created within the *Automatic Analysis of Phonetic Convergence in Speech Technology Systems* project. In Appendix V.3, the Python scripts written by the author of this work, which were created to automate annotation processes and create appropriate analyses, are presented. Appendix V.4 contains the manually annotated part of Harmonia corpus that was used in this work. In Appendix V.5, results of LSM factor measurements for each grammatical category are available.

II Communication convergence

People involved in a conversation communicate through language, voice, gestures and body language. Many studies have proven that people, either consciously or unconsciously, imitate their interlocutors, which has a number of positive effects, such as better understanding, group unity emphasising, easier persuading someone to their point of view or getting someone to do something (e.g. Manson, Bryant, Gervais, & Kline, 2013). Research on interpersonal communication has shown that the language and manner of speech of interlocutors during a dialogue becomes more similar as the conversation progresses. Interlocutors during the dialogue adapt their communication behaviour to each other. This phenomenon is observable in many aspects of language, in the paralinguistic, extralinguistic and lexical layers (Pardo, Pellegrino, Dellwo, & Möbius, 2022). The mechanism of adapting the way of speaking and language to the interlocutor is called convergence, entrainment, alignment, imitation, synchrony or mimicry (Pardo, Pellegrino, Dellwo, & Möbius, 2022; Tschacher, Rees, & Ramseyer, 2014; Louwerse, Dale, Bard, & Jeuniaux, 2012; de Looze, Oertel, Rauzy, & Campbell, 2011). Such terms appeared in the literature already at the end of the 20th century (e.g. Scollon & Scollon, 1980; Dingwall, 1979) but there are still many aspects that have not been addressed in scientific research on convergence in dialogues. Until recently, research on interpersonal communication has focused on the mechanisms of language comprehension and production, and on text processing in non-communicative situations (e.g. reading). This means that communication was studied partially, in some isolation, which does not give a full picture of the psychological and psycholinguistic mechanisms that occur during a dialogue (Pickering & Garrod, 2004). Currently, more and more research takes into account the entire context of the dialogue, not separating individual statements and analysing them as separate monologues, but as a single result of cooperation between the interlocutors (Pickering & Garrod, 2004).

Communication convergence is studied at different levels of communication and in different communication situations. In the literature, one can find numerous examples of research on the paralinguistic layer itself, prosodic and acoustic parameters of speech. No less interesting are the aspects related to the body language, i.e. facial expressions, gestures and posture during interaction. Many studies focus on the linguistic level of communication, analysing the lexis and syntax in conversations. Research on communication convergence is usually interdisciplinary, related to linguistics, psychology and sociology. Examples of research and conclusions will be presented in the following chapters of this work.

1 Reasons and sources of linguistic alignment in conversation

The topic of communication convergence appears in scientific publications from the second half of the 20th century. Initially, only speech-level convergence was considered, but over time it was extended to other aspects of communication. In the literature there are examples of publications dealing with the topic of origin and mechanisms related to convergence in communication (e.g. Elhami, 2020; Doyle & Frank, 2016). Based on the research, theories that attempt to explain this phenomenon taking into account various linguistic, psycholinguistic and sociolinguistic aspects have been developed. One of the main currents seeks to explain the phenomenon of convergence in a sociological and socio-historical context – Communication Accommodation Theory. The second takes into account priming – Interactive Alignment Model. CAT distinguishes between two types of adjustment: positive (convergence) and negative (divergence). Convergence is understood as a way to reduce social distance, it expresses the need for social approval, belonging to a group and can increase communication efficiency. Divergence is a way to increase social distance and maintain your own style. IAM disregards social aspects and issues related to the speaker's personality and views convergence as a simple and automatic mechanism based on precedence. In addition to these two main theories, researchers have

attempted additional explanations of the sources and effects of convergence in communication (e.g. Heath, 2017). However, the most popular and most cited theories behind alignment are CAT and IAM, which is why they will be presented in detail in this paper. The following chapters present the assumptions and theses of these two theories.

1.1 Communication Accommodation Theory (CAT)

Communication Accommodation Theory (CAT) was developed by Howard Giles and explains the behavioural changes that people make to adapt their communication to their partner, and the extent to which people perceive the partner as appropriately suited to them (Giles, Coupland, & Coupland, 1991). The theory evolved from Speech Accommodation Theory (SAT), which described changes in speech characteristics during human interaction (Giles H. , 1973). Since its definition, CAT has been modified and extended to other spheres beyond verbal communication. In addition to language features, the CAT also included other elements of behaviour, appearance, habits that may be related to the sense of belonging to a specific social group (Giles & Ogay, 2007).

CAT focuses on predicting and explaining the changes that interlocutors create during the interaction. This applies to both increasing and decreasing the level of differences in the form of information transfer in many aspects. CAT takes into account several basic assumptions. The first is the socio-historical context in which a given interaction takes place. The second concerns the extended concept of communication, which, apart from the exchange of information on specific topics, also includes social category membership regulated in the course of the conversation. The third assumes that interlocutors have specific expectations regarding the optimal level of communication accommodation. The fourth takes into account the communication strategies that interlocutors adapt in order to signal their attitude towards specific social groups (Giles & Ogay, 2007). Thus, CAT takes into account issues beyond the mere exchange of information, or rather sees sources of convergence in sociological and psychological aspects related to belonging to a group, being liked, creating or reducing distance.

One of the main goals of convergence is to win the sympathy and approval of the interlocutor. By making the communication behaviour similar to the conversation partner, it is more likely to be liked and respected by them. As a consequence, people receive social reward and a sense of belonging to a social group (Giles & Ogay, 2007). CAT assumes a conscious or semiconscious mechanism to adjust or vary communication behaviour in order to create, uphold, or reduce social distance. The goals of CAT have been defined by Pitts and Giles (2010):

> *Communication accommodation theory is primarily concerned with the motivation and social consequences underlying a person's change in communication styles (verbal and nonverbal features such as accent, volume, tone, language choice) to either accommodate or not accommodate their interactional partners* (Pitts & Giles, 2010).

The authors of the above quote indicate specific verbal and non-verbal features that may change as a result of convergence during interaction. More examples of speech features that have been taken into account in convergence studies can be found in the literature. They will be presented later in this work.

Convergence refers to the strategy of adapting communication behaviours in order to make them similar to the communication behaviours of the interlocutor. *Divergence* is about emphasising the differences between one's form of communication and that of the interlocutor. *Maintenance* is the lack of changes in communication behaviour regardless of the interlocutor's form of communication (Gallois, Ogay, & Giles, 2005). According to the assumptions of SAT and CAT, convergence occurs when interlocutors want to be liked by each other. Divergence or maintenance stands in opposition to convergence and is related to a lack of liking or a low need for social approval (Gallois, Ogay, & Giles, 2005). The perception of convergence or divergence is also an important element of communication leading to the creation of positive or negative associations with the interlocutor. According to the attribution theory (Kelley, 1973), which explains social perception and how

people make causal attributions, people tend to evaluate the behaviour of others by motive and intention. In short, this means that people appreciate, positively evaluate desirable behaviour if it comes from the willingness, internal motivation of the interlocutor, rather than from the need or pressure caused by circumstances. In the case of maladjusted behaviour, people are more indulgent, understanding when it can be explained by external circumstances than when it comes from internal causes.

1.2 Interactive Alignment Model (IAM)

The Interactive Alignment Model (IAM) was proposed by Pickering and Garrod (2004) as an explanation of increasing similarity in the language style and reuse of the same lexical items that occur during dialogue. Previously, researchers focused on the analysis of dialogue as separate utterances of interlocutors, which is basically a set of monologues. Pickering and Garrod (2004) proposed a new methodology for studying conversation as a single, undivided research material. As a result of these analyses, IAM was created, which assumes that the alignment process in conversation greatly facilitates linguistic processing and communication.

The Interactive Alignment Model assumes communicative alignment in dialogues to be a completely automatic and unconscious process. The theory completely ignores social issues and psychological mechanisms (the desire to belong, to be liked or to emphasise social status) described in the CAT. According to IAM, the goal of communication adaptation is to create a common ground that greatly simplifies the understanding and production of content in dialogue (Pickering & Garrod, 2004). Authors of IAM state:

> *The account assumes that in dialogue, production and comprehension become tightly coupled in a way that leads to the automatic alignment of linguistic representations at many levels. We argue that the interactive alignment process greatly simplifies language processing in dialogue. It does so (1) by supporting a straightforward interactive*

> *inference mechanism, (2) by enabling interlocutors to develop and use routine expressions, and (3) by supporting a system for monitoring language processing* (Pickering & Garrod, 2004).

Pickering and Garrod present the communication alignment as a situation where interlocutors achieve the same representation of a given level of speech. In their understanding, dialogue is coordinated behaviour, the success of which can be achieved by adapting the situation models - multi-dimensional representation of the situation under discussion. The situation model includes the following dimensions: space, time, causality, intentionality and references. Adjustment mechanisms are automatic and result mainly from priming. As the authors define, the appearance of a given utterance during a conversation activates a specific representation in the interlocutor, which increases the probability of using the same representations in subsequent utterances. Convergence at the global level comes from lower linguistic levels (Pickering & Garrod, 2004).

2 Levels of convergence

In the literature, there are many examples of research proving that interlocutors use various adaptive strategies in communication. Speech convergence was observed at the linguistic, paralinguistic and non-linguistic levels (Reitter & Moore, 2014; Street Jr., 1984; Misiek, Favre, & Fourtassi, 2020; de Jong, Theune, & Hofs, 2008; Xu & Reitter, 2016). As a result of numerous studies conducted on convergence in dialogues and group conversations, adjustments have been noticed in the acoustic features (Pardo, Pellegrino, Dellwo, & Möbius, 2022) and phonetic realisation for particular words (Pardo, 2006). Research shows that interlocutors also adjust response latency and utterance duration measured as the duration of speech signal (Joseph D., Arthur N., Ruth G., & George, 1968) and number of uttered lexical units (Bangerter, Mayor, & Knutsen, 2020). Interlocutors also tend to adapt the syntax – they are more likely to use the syntactic structure that their interlocutor has used than an alternative one (Branigan, Pickering, & Cleland,

2000). Analyses of recordings of task-oriented dialogues showed a tendency for speakers to repeat their own and partners' syntactic and structural patterns in conversation. Similar results were obtained by corpus research consisting in tracking the frequency of occurrence of the same language constructions in natural conversations (Howes, Healey, & Purver, 2010). Also, the convergence was observed at structure level, where different linguistic choices are associated with different conceptualizations. This was shown in the analysis of players' communication when they indicated their position in a maze (e.g. in terms of paths between two points or as column-row indices) in a study by Garrod and Anderson (1987). Depending on the interlocutors, circumstances, topic, etc., convergence or divergence may also be noticeable at the level of lexical complexity and lexical matching (Bangerter, Mayor, & Knutsen, 2020; Schneider, Ramirez-Aristizabal, Gavilan, & Kello, 2020). Convergence is also observable at the lexical level in the conversation of people who are multilingual or who use certain dialects. Research shows that bilingual speakers tend to adjust the level of language mixing in one conversation (code-switching) (Toribio, 2004). In the area of non-linguistic communication, adaptation at the level of mimicry, gestures and posture was noticed (Lakin, Jefferis, Cheng, & Chartrand, 2003). The researchers also payed attention to the adjustment at the level of formality of the language style (Mirzaiyan, Parvaresh, Hashemian, & Saeedi, 2010) as well as the adjustment related to the gender of the interlocutor (Tet-Mei Fung, Chuah, & Ting, 2020). Table 1 summarises the main areas where communication adaptation has been studied and observed. A more detailed analysis of the features in which convergence is observed in verbal communication (at the linguistic and paralinguistic level) is presented in the following chapters of the work.

Level	**Features converged**	**Sources**
Paralinguistic	Phonetic features	(Pardo, 2006)
Paralinguistic	Acoustic features	(Pardo, Pellegrino, Dellwo, & Möbius, 2022)
Paralinguistic	Response latency	(Street Jr., 1984)
Paralinguistic	Utterance duration	(Joseph D., Arthur N., Ruth G., & George, 1968)
Linguistic	Utterance length	(Bangerter, Mayor, & Knutsen, 2020)

Level	Features converged	Sources
Linguistic	Syntactic structures adaptation and syntactic complexity	(Branigan, Pickering, & Cleland, 2000), (Xu & Reitter, 2016)
Linguistic	Structure levels referring to conceptualisations	(Garrod & Anderson, 1987)
Linguistic	Lexical complexity and lexical matching	(Bangerter, Mayor, & Knutsen, 2020), (Schneider, Ramirez-Aristizabal, Gavilan, & Kello, 2020)
Linguistic	Information density	(Aronsson, Jönsson, & Linell, 1987)
Linguistic	Gender-related alignment	(Tet-Mei Fung, Chuah, & Ting, 2020)
Linguistic	Formality level	(Mirzaiyan, Parvaresh, Hashemian, & Saeedi, 2010)
Linguistic	Code-switching	(Kootstra, Dijkstra, & van Hell, 2020), (Toribio, 2004)
Linguistic	Topic and dialect	(Giles & Soliz, 2014)
Linguistic/ Paralinguistic	Turn-taking and vocal activity rhythms	(Campbell & Scherer, 2010), (McGarva & Warner, 2003)
Non-linguistic	Mimicry, Gestures, Posture	(Lakin, Jefferis, Cheng, & Chartrand, 2003)

Table 1. Summary of features converged in human interactions – own study.

2.1 Phonetic-acoustic convergence in dialogues

The list of acoustic and phonetic features of speech is long and, in the literature, examples of convergence studies using virtually all of them can be found. There are various approaches to speech analysis presented and examples of the types of speech features, which include (1) qualitative features, (2) teager energy operator (TEO) - based features, (3) spectral features, (4) continuous features (e.g. Surya Gunawan, Fahreza Alghifari, Arman Morshidi, & Kartiwi, 2018). Figure 1 presents four categories of speech features and examples of characteristics and measures used in each category.

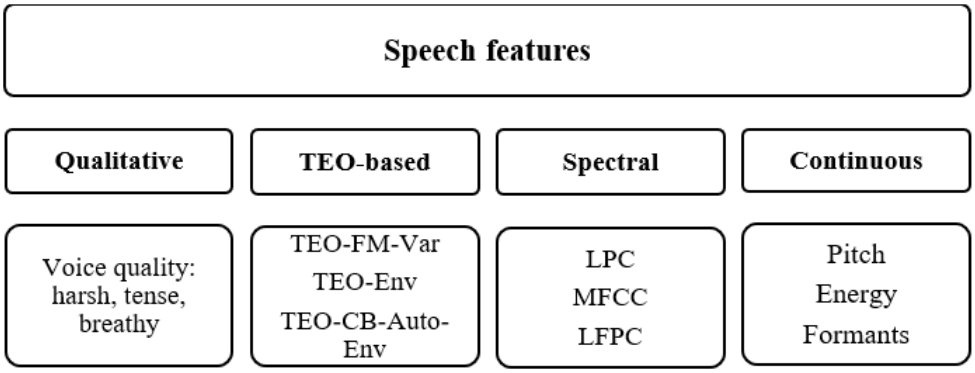


Figure 1. Speech features
(Surya Gunawan, Fahreza Alghifari, Arman Morshidi, & Kartiwi, 2018).

Research in the field of phonetic-acoustic convergence covers many features and parameters of speech. Acoustic convergence has been observed on many levels such as speech rate, pause rates, fundamental frequency, long-term average spectra, mel-frequency cepstral coefficients, voice quality, voice onset time, vowel formants, clicks, utterance duration, amplitude envelope signals (Pardo, Pellegrino, Dellwo, & Möbius, 2022). When analysing the methodologies applied in the speech convergence studies, some differences in the selection of parameters and measures can be noticed. For example, speech rate can be measured as mean syllable duration (Manson, Bryant, Gervais, & Kline, 2013), word duration (Wagner, Broersma, McQueen, Dhaene, & Lemhöfer, 2021), vowels uttered per minute and syllables uttered per minute (Demenko, 2020). In the analysis of dialogues, turn duration also turned out to be an important variable, which was examined as the duration of a single turn in a dialogue (Street Jr., 1984) or average time of an interlocutor speech in a conversation (Matarazzo & Wiens, 1972). The pauses between speaker turns (response latency) were also investigated by Street (1984) as well as the vocalisation duration, which is basically the duration of speech without pauses was examined in convergence studies.

Fundamental frequency (F_0) value, which is observable in voiced sounds and reflects movements of the glottis. The acoustic characteristics of voiced sounds are periodic, caused by the vibrations of vocal folds, and depend on the mass and elasticity of the glottis (Jiménez-Hernández, 2016). Numerous studies have shown that pitch range and pitch variation may differ

according to the age (Schötz, 2007), gender (Jassem, 2003), language and dialect of the speaker (Andreeva, et al., 2014). For example, the overall mean values of F0 measurements for Slavic languages are higher than for Germanic languages (Andreeva, et al., 2014). In convergence studies, F0 is usually measured and the statistics in a speech sample are calculated. For example, mean F0 value (in Hz) for each speaker in a dialogue (Demenko, 2020) or median F0 value (in Hz) of isolated voiced segments in each sentence (Bradshaw & McGettigan, 2021). Formants relate to a range of frequencies in which there is an absolute or relative maximum in the sound spectrum measured in hertz (Hz) (Acoustical Society of America, 1994). In other words, formant is a broad spectral maximum, a peak in the spectrum which is a result of acoustic resonance in vocal tract. The formant with the lowest frequency is F_1, the second is F_2, the third is F_3. Usually the formant value 1 and 2 are sufficient to identify a vowel. In speech convergence studies the formants were analysed and F1, F2 vowel space was drawn (Pardo, Urmanche, Wilman, & Wiener, 2017). Demenko (2020) investigated the mean duration of vowels and syllables, accented and not accented separately.

Another technique for speech feature extraction and analysis used in convergence studies is the Mel Frequency Cepstral Coefficients (MFFCs). Computation of MFCCs includes a conversion of the Fourier coefficients to Mel-scale. There are 39 features of MFCCs, namely 12 MFCC features, 12 Delta MFCC features, 12 Delta Delta MFCC features, 1 (log) frame energy, 1 Delta (log) frame energy and 1 Delta Delta (log) frame energy. In convergence studies, for instance, Bailly and Martin (2014) investigated alignment in a dialogue by the linear discriminant analysis of MFFC.

There are also studies focused on the analysis of vocal intensity, the results of which are presented in intensity courses during dialogue for each speaker (Demenko, 2020). Vocal quality was assessed based on jitter, shimmer and noise and harmonics ratio measurements (Levitan & Hirschberg, 2011).

Researchers points to convergence in the phonetic realisation for particular words which is usually studied perceptually. Phonetic realisation was tested in dialogues recorded for the study, based on the map task scenario

(Pardo, 2006) and in the so-called shadowing experiments (Gessinger, et al., 2018). Shadowing tasks are focused on the analysis of casual speech (baseline) produced at the beginning of the study and comparison to the realisation repeated after stimuli. In the shadowing experiment conducted by Gessinger et al. (2018) there was a natural and synthesised speech used as stimuli for participants. A similar study was conducted by Jankowska et al. (2021) for the Polish language, which will be described in detail in Chapter II.5.5.1.

The subject of convergence study was also the number of clicks, i.e. the sum of velaric ingressive stops (Gold, French, & Harrison, 2013) and Voice Onset Time (VOT) (Solanki, 2017). VOT refers to the period between the release of a plosive and the beginning of vocal cords vibration. Table 2 summarises the features, parameters and measures found in the literature in the phonetic and acoustic convergence studies.

Feature	**Parameters and measures**	**Source**
Speech rate	Mean syllable duration	(Manson, Bryant, Gervais, & Kline, 2013)
	Word duration	(Wagner, Broersma, McQueen, Dhaene, & Lemhöfer, 2021)
	Vowels per minute	(Kousidis, et al., 2008), (Demenko, 2020)
	Syllable per minute	(Street Jr., 1984), (Demenko, 2020)
Turn duration	Duration of a single turn in a dialogue	(Street Jr., 1984)
	Average amount of time that an interlocutor spoke	(Matarazzo & Wiens, 1972)
Vocalization duration	Duration of speech (without pauses)	(Street Jr., 1984)
Response latency	Pauses between speaker turns	(Street Jr., 1984)
Fundamental frequency (F0)	Median F0 value (in Hz) of isolated voiced segments in each sentence	(Bradshaw & McGettigan, 2021)
	Mean F0 value (in Hz) for each speaker in a dialogue	(Demenko, 2020)
Phonetic realisation	Perceptual assessments (additionally, measures of item duration and vowel spectra were compared to perceptual assessments)	(Pardo, Gibbons, Suppes, & Krauss, 2012)

Feature	Parameters and measures	Source
Pause rates	Number of pauses, mean and median pause duration	(Šturm & Volín, 2023)
Internal pauses duration	Duration of pauses within a speaker's turn	(Street Jr., 1984)
Long-term average spectra	Mean spectrum from LTAS sample	(Gregory & Webster, 1996)
Vocal intensity	Intensity courses during dialogue for each speaker (in dB)[8]	(Natale, 1975), (Demenko, 2020)
Voice quality	Jitter, shimmer, and noise to harmonics ratio	(Levitan & Hirschberg, 2011)
Voice onset time	Duration of negative and positive VOT	(Solanki, 2017)
Vowel formants	F0, F1, F2, and the F1 × F2 vowel space	(Pardo, Urmanche, Wilman, & Wiener, 2017)
Segmental and suprasegmental units duration	Mean duration of vowels and syllables (accented and non-accented)	(Demenko, 2020)
Click rates	Frequency of clicking (production of velaric ingressive stops)	(Gold, French, & Harrison, 2013)
Mel Frequency Cepstral Coefficients	Linear discriminant analysis	(Bailly & Martin, 2014)

Table 2. Summary of phonetic-acoustic features converged and parameters and measures used in previous research – own study.

2.2 Lexical and syntactic convergence in dialogues

Lexical entertainment refers to the tendency to choose and repeat certain words as referents during dialogue (Schneider, Ramirez-Aristizabal, Gavilan, & Kello, 2020). Similarly, in the case of used syntactic structures whose repetition may be caused by, among others, lexical priming or discourse register (Pickering & Branigan, 1998). Lexical convergence can be studied in many ways, taking into account various features of the language. This type of analysis is applicable not only to spoken but also to written language, which broadens its applications and increases the number of possible communication situations. In the case of lexical convergence analysis, both spoken (recording transcripts) and written materials (letters, e-mails, posts on Internet forums, chat messages, text messages, etc.) can be used. The methodologies of this type of analysis included, as in the case of the analysis

[8] In the study presented in the book the TAMA (Time Aligned Moving Average) approach was used.

of speech and acoustic-phonetic features, the length of utterances measured in the number of morphemes (Brown, 1973) or the number of words (Brownell, 1988). Bangerter, Mayor and Knutsen (2020) used the total number of words uttered by both participants in a dialogue to compare the so-called collaborative effort in different types of dialogues. The same authors analysed lexical diversity in dialogues by summing new verbs and nouns (content words) and dividing them by the sum of all occurrences of these word types. Similarly, in the case of indefinite references, which were summed up and divided by the total number of words uttered by a speaker (Bangerter, Mayor, & Knutsen, 2020). In the study of lexical convergence, parameters such as the length of utterances measured by the number of morphemes (Brown, 1973) or the number of words (Brownell, 1988) were also important. Similarly, the length of sentences in the utterances, which were important in the analysis of syntactic complexity, was examined. In this type of research, the syntactic tree depth and branching factor analysis, i.e. the analysis of parse tree elements in a sentence, was also taken into account (Xu & Reitter, 2016). In studies using other language materials (i.e. corpora, lexicons), syntactic and lexical matching was checked by the number of shared n-grams normalised by the number of all possible ngrams (Misiek, Favre, & Fourtassi, 2020). Table 3 summarises the lexical features converged as well as parameters and measures used in research found in the literature.

Feature	Parameters and measures	Source
Utterance length	Number of morphemes	(Brown, 1973)
	Number of words	(Brownell, 1988)
Collaborative effort	Total number of words uttered by both participants in a dialogue	(Bangerter, Mayor, & Knutsen, 2020)
Lexical diversity	Ratio of the total of new word types divided by the total of word types[9]	(Bangerter, Mayor, & Knutsen, 2020)
Indefinite reference	Number of pronouns (the number of pronouns divided by the total number of words)	(Bangerter, Mayor, & Knutsen, 2020)
Syntactic complexity	Sentence length (number of words in a sentence)	(Xu & Reitter, 2016)
	Tree depth	(Xu & Reitter, 2016)

[9] In the research presented in the paper content words (nouns and verbs) were counted.

Feature	Parameters and measures	Source
	Branching factor (average number of children of all non-leaf nodes in the parse tree of a sentence)	(Xu & Reitter, 2016)
Syntactic alignment (speaker's reuse of syntactic structure from the interlocutor's previous utterance)	Number of shared PoS ngrams (bigrams and trigrams) normalised by the number of all possible ngrams	(Misiek, Favre, & Fourtassi, 2020)
Lexical alignment (speaker's reuse of words from the interlocutor's previous utterance)	Number of shared ngrams (unigrams, bigrams and trigrams) across pairs of turns, normalised by the number of all possible ngrams	(Misiek, Favre, & Fourtassi, 2020)
Politeness and formality	Analysis based on a lexicon of more and less formal greetings and noun synonyms that have different levels of formality in Dutch	(de Jong, Theune, & Hofs, 2008)

Table 3. Summary of lexical features converged and parameters and measures used in previous research – own study.

3 Methodology of communication convergence research

In the literature, there are descriptions and formulas of complex methodologies for studying linguistic alignment. In this chapter, examples of methodologies created specifically for the study of convergence and others drawn from related fields will be provided.

Language Style Matching (LSM) is a technique for analysing stylistic similarities in the language of different groups and individuals and an indicator of interpersonal and group alignment. This is one of the methods of measuring verbal mimicry in conversations. LSM provides an algorithm for automatically assessing the level of linguistic adjustment, which takes into account the specific features and elements of the language. This method makes it possible to parse conversations in specific psychological dimensions. LSM methodology, examples of research, conclusions and correlations between high LSM factor and psychological-behavioural phenomena are described in detail in Chapters II.3.3.1 and II.3.3.2.

Other methods of measuring lexical adaptation can be found in the literature, e.g. probabilistic measures, which involve dividing texts into parts and counting the frequency of words and syntactic rules (Amit, Sturt, &

Keller, 2005). In order to perform this kind of calculation, a large amount of linguistic material is required in which trends can be observed and the probability of co-occurrence estimated (Xu & Reitter, 2015). Other methods used in the convergence study are taken from the document similarity measures methodology and includes measures such as Spearman's rank correlation coefficient, Zipping and Latent Semantic Analysis. In the case of Spearman's rank correlation coefficient document similarity is measured based on word frequency and co-occurrence. Zipping uses a data compression algorithm and Latent Semantic Analysis measures semantic similarity between documents (Xu & Reitter, 2015). Other approaches take repetition decay analysis, which focuses mainly on the measurement of syntactic alignment. For example, Reitter, Keller & Moore (2006) proposed a generalised linear model that used repetition of syntactic rules as the dependent variable and distance between prime and target as predictor. This method is suitable for large amounts of data, e.g. for transcription of long conversations between the same people.

Another example of a methodology for studying lexical and syntactic convergence is Local Linguistic Alignment (LLA). This methodology consists of two components: Lexical Indiscriminate Local Linguistic Alignment (LILLA) and Syntactic Indiscriminate Local Linguistic Alignment (SILLA) (Carrick, Rashid, & Taylor, 2016). LILLA deals with lexical units only, and SILLA requires sentence-level annotation with the phrase structure tree. The calculation results for this methodology are obtained by calculating:

$$LILLA\ (target, prime) = (p(target|prime))/\ (p(target))$$

The same formula is used to calculate SILLA. LILLA and SILLA are measured as item (word or syntactic structure) repetition in the target text after its occurrence in prime text. A high level of LILLA and or SILLA indicates increase in alignment. This methodology has been used i.e. in convergence studies in online forum discussions (Wang, Reitter, & Yen, 2014).

Linguistic convergence analysis techniques include dedicated methodologies and methodologies drawn from other language processing fields, such as document similarity analysis. The measures and parameters used are usually related to lexis and syntax. Language materials subjected to this type of analysis usually require appropriate normalisation (e.g. lemmatization of lexical units) and annotation (e.g. at the word level – part of speech, at the sentence level – syntactic tree). Most require the definition of a *prime* and *target* source, i.e. primary and stimulus text. However, there is a methodology that does not require dividing or grouping linguistic resources – Language Style Matching.

3.1 LSM methodology

LSM takes into account function words that are useful in analysing social psychological states through language (Gonzales, Hancock, & Pennebaker, 2010). It focuses not on the content words (nouns and verbs), but on the way the language is used, the analysis of function words, i.e. pronouns, articles, conjunctions, prepositions, auxiliary verbs, etc., that play a syntactical role (Ireland, et al., 2011). Function words have features that allow relatively easy and useful analysis. They usually occur with high frequency, covering a significant part of everyday speech. Function words are context-independent and are produced unconsciously, avoiding manipulation through specific patterns. The methodology proposed by the LSM authors focuses on counting the words used from nine main categories:

- auxiliary verbs (e.g., to be, to have),
- articles (e.g., an, the),
- common adverbs (e.g., hardly, often),
- personal pronouns (e.g., I, they, we),
- indefinite pronouns (e.g., it, those),
- prepositions (e.g., for, after, with),
- negations (e.g., not, never),
- conjunctions (e.g., and, but),
- quantifiers (e.g., many, few).

A high LSM score is associated with positive interactions, belonging, and being liked, while a low LSM score is associated with the opposite (Gonzales, Hancock, & Pennebaker, 2010).

The LSM analysis consists of calculating the number of occurrences of each type of function words and calculating the percentage of their occurrence in the utterances and calculating the LSM index - dividing the absolute value of the difference between the speakers by the sum for each category (Gonzales, Hancock, & Pennebaker, 2010). Based on the examples presented by the authors of the algorithm in the article, a general formula for calculating the LSM index can be presented:

$$fwLSM = 1 - ((|fw1 - fw2|) / (fw1 + fw2 + 000{,}1))$$

where *fw* stands for function word and number *1* determines the result of the person and *2* of his interlocutor. The results are calculated for each function word category separately. The final result is the average of the results of the individual calculations. The result of this operation is between 0-1 and the closer the value is to 1, the higher the language style matching, verbal mimicry level.

3.2 LSM and psychological factors and behavioural outcomes

Research on Linguistic Style Matching arouses the interest of many linguists, psychologists and sociologists. Based on the results, behaviours, attitudes, emotions, perception of messages and other psychological and behavioural variables can be inferred and predicted. In the literature, there are many examples of research in face-to-face, telephone and electronic interaction (internet forums, discussion groups on portals, e-mails, etc.). The LSM technique has been used in many studies, on the basis of which conclusions have been drawn about LSM in the field of cooperation, interdependence, liking, attraction, positive perception of other people, attachment, mutual understanding, child development and therapeutic benefits.

In terms of cooperation, the level of LSM in conversations between police officers and suspects was tested and an increase was observed in cases of confession (Richardson, Taylor, Snook, Conchie, & Bennell, 2014).

Another example is the analysis of mediation during a divorce trial, which showed that couples who agreed terms had a higher level of LSM than those who did not reach an agreement (Donohue & Liang, 2011). Another study of conversations between couples in relationships found that in conflict discussions, higher LSM correlated with negative emotions and a lower sense of being supported. In supportive discussions, couples with higher LSM experienced more positive emotions and a sense of support (Bowen, Winczewski, & Collins, 2017). Studies conducted in the field of LSM in conversations of potential partners during speed dates have shown that a higher LSM increases the likelihood of romantic interest in a partner. The same tendency is formed in the case of maintaining relationships (Ireland, et al., 2011). In addition, people who communicate in a similar linguistic style are more likely to form and maintain friendships and they increase linguistic convergence over the duration of the relationship (Kovacs & Kleinbaum, 2019). Linguistic accommodation was also explored in the virtual space, e.g. in Reddit communities (Sharma & De Choudhury, 2018), internet forum for the Moroccan minority in the Netherlands (Welbers & de Nooy, 2014), health bloggers posts and related comments (Rains, 2016) etc. Research has shown that in online communication people adapt their language style to the relevant posts and LSM contributes to a sense of social support on forums and discussion groups. LSM studies in task-oriented groups showed a higher LSM level correlation for success than failure (Purpura, Schwanda, Williams, Stubler, & Sengers, 2011). There are many more examples of LSM research. An extensive list with references, summaries and conclusions has been published by Shaw et. al (2019).

The LSM methodology can be used on both written and spoken language resources. However, these must be materials collected during a communicative interaction of at least two people, available and annotated as required – detailed part of speech tagging. In the following chapters, language resources available in various languages and Polish will be presented.

4 Resources for language processing

In language research and technology, quality databases are a key element that enables high quality studies and creation of speech analysis, recognition and synthesis systems (Li & Yin, 2007). For such purposes, linguistic corpora are created and used. According to the PWN Dictionary of the Polish Language[10], the corpus is *texts, data, etc. collected due to their representativeness, constituting the basis for scientific analysis*. Another source defines corpus as *a set of linguistic texts collected for the purpose of studying its system or subsystem* (Polański, 1999). In the literature other definitions can be found, for example, *language corpus is a body of documented evidence of the authentic use of natural language, a vast body of electronic texts deliberately collected as a reference source* (Tkaczewski, 2008). All the above-mentioned definitions present the linguistic corpus as a collection of various types of written and spoken texts gathered in a single database that enables automatic searching and analysis of linguistic data.

The most important issue for any linguistic database is the quality, which can be described within the following characteristics:

- representativeness of data,
- complexity, including information from many sources,
- heterogeneousness - not depended on any particular computer operating system or platform,
- annotation - containing linguistic and meta information,
- availability to researchers operating in an open environment,
- distribution - open file formats, supporting import and export various data formats,
- maintenance - corpora need to be routinely augmented with new information or else soon they will have only historical value.

A language corpus that meets the above conditions can be used for linguistic research, which guarantees the quality and credibility of research results.

[10] https://sjp.pwn.pl/szukaj/korpus.html

In addition to corpora, linguistic resources may also include lexicons. The field that deals with the creation of lexicons for the purpose of computer processing of natural languages is computational lexicography (Gibbon & Borchardt, 2007). It deals with the study and modelling of the automatic acquisition of lexical units from collections of texts, the construction of lexicons on the basis of corpus, automatic extraction of syntactic and semantic information, creating, extending and maintaining machine-readable dictionaries (Van Eynde & Gibbon, 2000). Lexicons are created for various research purposes and they are used in the automatic search of large amounts of language data or Internet content, in machine translation, recognition of specific features of the text, etc. Depending on the needs, lexicons are created in different languages containing specific terms related to the appropriate language domain or lexis allowing to recognize features in both spoken and written language (Gibbon & Lüngen, 2000). In terms of technological application, speech and language processing require access to large lexical data so as to achieve a high degree of accuracy (Van Eynde & Gibbon, 2000).

Linguistic resources can be constructed from existing materials such as literature, magazines, online texts, radio broadcast, TV program or they can be created (recorded or written) specifically for a given study. There are language sources designed for research in a specific field of science, e.g. emotional speech corpora, corpora of speech recordings of people suffering from various diseases, corpora for convergence studies and others. Examples of language corpora both in various languages of the world and in Polish will be described in the following chapters.

4.1 Linguistic corpora

There are many examples of corpora in many languages of the world created from existing materials (e.g. books, films, radio) and recorded or written according to scenarios created in a way that allows the study of selected phenomena. The use of language corpora is wide, research and scientific in the field of e.g. sociolinguistics, psycholinguistics, development in the field of modern applications based on language processing or learning and teaching

foreign languages. Due to the application and specificity of the texts contained in the corpus, corpus types are distinguished: (1) general – dealing with a wide range of language, containing texts from a variety of sources (often used to create dictionaries), (2) specialised – focused on one feature, containing texts from one field of science, literature from one author or genre, etc., (3) synchronous – presenting the current state of natural language, (4) diachronic – presenting changes that occur in natural language over time, (5) spoken language – containing recordings and/or transcripts of spoken language, (6) written language – containing only materials from written sources, (7) parallel – containing equivalent texts in two or more languages (Zasina, 2018).

The first computer-readable general corpus of texts for linguistic research was the Brown Corpus for the English language. The corps was created in the years 1963-1964 as part of a project led by W. Nelson Francis and Henry Kučer. The corpus contains approximately 1 million words, 500 samples and over 2000 words each. Romanian Academy, Faculty of Computer Science and Technology launched the Computational Reference Corpus for the modern Romanian language (CoRoLa[11]) in 2014. CoRoLa is a collection of contemporary texts (written and oral), with huge dimensions (approximately 1 billion words and over 300 hours of voice recording). The corpus is completed with a set of metadata (which relate to the author, publication, publication date, literary type of the text, etc.) and annotations presenting information of a linguistic and grammatical (morphological, lexicographic, syntactic, etc.) nature. There are many more language corpora in various languages created for research and commercial use, for example, data provided by the European Language Resources Association[12] (ELRA), the Linguistic Data Consortium[13] (LDC) or the ELSNET group[14]. The list of language corpora is presented in the Appendix V.1.

[11] http://corola.racai.ro
[12] http://www.elra.info/en/
[13] https://www.ldc.upenn.edu/
[14] http://www.elsnet.org/

4.2 Polish linguistic resources

Language materials for the Polish language are created according to rules and requirements and are often adapted to specific research and development needs. Collections of linguistic materials were created, which were appropriately developed and annotated to form a corpus. National Corpus of the Polish Language (pl. Narodowy Korpus Języka Polskiego – NKJP) is a corpus of the Polish language, launched in 2012. The corpus includes words from Polish literature, daily and specialist magazines, as well as recordings of dialogues and texts from the Internet. NKJP contains about 1,500 million text words, or about 1,800 million segments (Przepiórkowski, 2011). A dedicated PELCRA corpus search engine[15] has been created for NKJP, which allows extraction of relevant information from a balanced version of the corpus. Advanced search settings allow you to define such parameters as source type (e.g. literature, poetry, media sources, spoken language, etc.), channel (e.g. internet, book, press, etc.), source publication date, source and text title, words contextual required and not allowed. A useful feature of the search engine is the ability to display graphs of the frequency of use of specific words or phrases in publications over time, type of sources, channels and exporting search results to a spreadsheet. The search engine was created by the research team of Prof. Barbara Lewandowska-Tomaszczyk at the Institute of English Studies, University of Łódź. The same research team has created other useful corpus tools, e.g. HASK – collocation dictionaries for English and Polish corpora, Spokes - a multimedia search engine for a unique corpus of casual conversational Polish and other[16].

Another example is the JURISDIC Polish Word Corpus which is intended for training and testing the dictation system dedicated to legal texts. It can be used to model word-spotting systems and speaker- and text-independent systems that use modelling of words or other language units. The specification includes speech recording in read, semi-spontaneous and spontaneous styles. Assumptions regarding the structure of the database are

[15] http://nkjp.uni.lodz.pl/index_adv.jsp
[16] http://pelcra.pl/new/tools_and_resources

based on general linguistic conditions and specific phonetic and acoustic properties of the Polish language. The average duration of a recording session was about 60 minutes. The JURISDIC database consists of 2,219 recording sessions and contains a total of 704,520 statements with a total duration of more than 1,200 hours of speech (Demenko, et al., 2008).

The Polish Corpus of Wrocław University of Technology is a collection of text documents that have been tagged using the wcrft2[17] tool and described with various types of information such as chunks, relationships between syntax phrases, identification units (including their relationships and lemmatization), disambiguated meanings words, spatial expressions, verbs with implied subject, textual keywords, temporal expressions (locally and globally normalised), situations, semantic roles and coreferences. The corpus consists of 449,985 tokens, but it is constantly expanded and developed towards a balanced corpus, containing scientific, official, artistic, rhetorical, press, journalistic and colloquial texts in equal measure (Broda, Marcińczuk, Maziarz, Radziszewski, & Wardyński, 2012).

The Parliamentary Discourse Corpus was developed by the Linguistic Engineering Team of the Institute of Computer Science of the Polish Academy of Sciences. The corpus is a regularly supplemented collection of annotated texts from plenary sessions of the Sejm and Senate of the Republic of Poland, parliamentary interpellations and questions and committee meetings from 1919 to the present. Texts are described with metadata and automatically processed with linguistic tools at the level of segmentation, morphosyntax analysis, recognition of syntax groups and proper names (Ogrodniczuk, 2018).

A research team from the AGH University of Science and Technology has created a corpus of audiovisual recordings of Polish speech. The corpus consists of good quality facial recordings of 20 different speakers, male and female, and transcripts of speeches. The semantic content of each speaker's recordings is the same. The total duration of the recordings is 200 minutes (Igras, Ziółko, & Jadczyk, 2012). The same research team created a

[17] A morphosyntactic tagger for Polish (https://clarin-pl.eu/dspace/handle/11321/36).

corpus of emotional speech in Polish. 6 women and 6 men aged 20-30 took part in the recordings. Some of the participants were former actors, and some were student volunteers. The corpus contains recordings expressing five of the basic emotions: joy, sadness, anger, fear, surprise, irony, and a neutral state as a reference signal. In total, 282 words were recorded for each speaker for each of 6 emotional states (Ignas & Ziółko, 2013).

In addition to language corpora, there are also many dictionary materials and lexicons in Polish. The vast majority of them, however, are not adapted to computer processing, although there are collections of this type, e.g. *Computer dictionary of difficult words inflection* (pl. *Komputerowy słownik odmiany wyrazów trudnych*) (Lubaszewski, Moskal, Pisarek, & Rokicka, 1996). A list of language sources is available, for example, on the website of the Department of Phraseology and Culture of the Polish Language Adam Mickiewicz University[18]. The site contains a list of 289 items in the category of orthoepic dictionaries and language guides, basic general dictionaries of the Polish language, other dictionaries of the Polish language, linguistic encyclopaedias and terminology compendia.

The presented examples of linguistic materials in Polish constitute a good research material in related fields. NKJP is a well-developed, accessible, annotated corpus of written and spoken language. The tools created to use this corpus enable numerous linguistic analyses. JURISDICT is a comprehensive source of high-quality legal jargon material. Similarly in the case of the Parliamentary Discourse Corpus, which contains texts of official speeches of Polish politicians. The AGH audiovisual corpus and the AGH emotional speech corpus are equally interesting and useful, though less extensive, research material. In order to perform a linguistic convergence analysis in the dialogues of Polish speakers in various communication situations, recordings of spontaneous or semi-spontaneous speech in a two-person interaction in various communication situations, conducted by the same speakers, are required. None of the above-mentioned sources meets these requirements,

[18] https://kjp.amu.edu.pl/sip.html

therefore it was decided to choose another linguistic material described in detail in Chapter III.1.1.1.

5 Linguistic convergence in Polish

Overall, little research on convergence in Polish has been conducted and published in the literature. However, there are several examples of publications that describe research in this area. Examples of such studies will be presented in this chapter.

Placiński (2019) investigated interactive alignment models in Polish in computer-mediated communication. The study was designed to test Pickering and Garrod's interactive alignment and showed convergence in repetitive verb usage and word order. In addition, the study found that the more lexical items shared, the shorter the conversation, which is related to reaching understanding and conversational success. Study conducted by Łyskawa et. al (2016) showed that for second generation heritage Polish speakers, individuals' code-switching (English-Polish) rates are positively correlated with their rates of word-final obstruent devoicing. Furthermore, the authors suggest that frequent code-switching provides the context in which knowledge of Polish and English patterns converge as the speakers present a convergence of both languages' grammars (Łyskawa, Maddeaux, Melara, & Nagy, 2016).

5.1 Phonetic-acoustic convergence in Polish

There are few publications in the literature devoted to phonic-acoustic convergence in Polish. However, there are several groups of scientists who work in this niche field. One of the leading teams in Poland in this field of science is the team led by prof. Grażyna Demenko, who received funding for research on phonetic-acoustic convergence in Polish. As a result of the project work, a language corpus (Harmonia corpus - see Chapter III.1.1.1) and publications describing the results of research in this area were created.

The following analyses were performed in the study:

- duration of syllables at the beginning, middle and end of sentences also with division into stressed and unstressed syllables
- the duration of stressed and unstressed vowels at the beginning, middle, and end of sentences
- analysis and comparison of F0 values
- intensity analysis
- speech rate (syllables per minute).

The results are described in detail in the publication by Demenko (2020).

Another example of the studies on phonetic-acoustic convergence in Polish was the analysis of two types of Polish task-oriented dialogues, mutual visibility and lack of mutual visibility. The authors focused on analysis of the speech rate (number of syllables divided by their total duration) and speech rhythm irregularity (syllable timing pairwise variability). The results of the study showed no significant differences between the groups (Karpiński, Klessa, & Czoska, 2014).

PhD Jolanta Bachan conducts research on modelling phonetic convergence in dialogue systems (Bachan, 2022). PhD Magdalena Zając took up the topic of phonetic convergence in the speech of Polish students of English. In her doctoral thesis, she focuses on the analysis of speech convergence in L2 pronunciation by Polish speakers (Zając, 2015).

Another important example of the studies on phonetic-acoustic convergence was the shadowing experiment. The experiment was conducted for the Polish language under the supervision of Prof. Grazyna Demenko by the author of this work and PhD Tomasz Kuczmarski, who was responsible for creating the synthesised speech. The study and results were published in the journal *Lingua Posnaniensis* (Jankowska, Kuczmarski, & Demenko, 2021). In order to discover the linguistic behaviour of people while interacting with the artificially generated speech in Polish, a shadowing experiment was performed. This experiment required the creation of a spoken language corpus that included several types of recordings of spoken sentences by a man, a woman, and artificially synthesised speech. Table 4 presents the sentences used as the linguistic material for recordings.

Orthographic word-medial letters *ę* before vowels other than fricative				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Ta służba to mordęga.	ę	*ɛw/ ɛ̃*	en	*ɛn*
Wszędzie jest spory bałagan.	ę	*ɛw/ ɛ̃*	en	*ɛn*
Wczoraj było jakieś święto.	ę	*ɛw/ ɛ̃*	en	*ɛn*
Do dziś cierpią męczarnie.	ę	*ɛw/ ɛ̃*	en	*ɛn*
Gęba sama się wykrzywia.	ę	*ɛw/ ɛ̃*	em	*ɛm*
Orthographic word-medial letters *ą* before vowels other than fricative				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Dął straszliwy wiatr.	ą	*ɔw/ ɔ̃*	o	*ɔ*
Z tej mąki nie upieczesz chleba.	ą	*ɔw/ ɔ̃*	on	*ɔn*
Obcy nie może tu rządzić.	ą	*ɔw/ ɔ̃*	on	*ɔn*
To nie jest zbyt rozsądne.	ą	*ɔw/ ɔ̃*	on	*ɔn*
Zagraj to teraz na trąbce.	ą	*ɔw/ ɔ̃*	om	*ɔm*
The letter *ń* in various positions				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Nad wejściem wisi końska podkowa.	ń	*ɲ̊*	ń	*J̃*
Przyznawał się do duńskiego pochodzenia.	ń	*ɲ̊*	ń	*J̃*
Był zupełnym jej przeciwieństwem.	ń	*ɲ̊*	ń	*J̃*
Niańczą dwójkę swoich dzieci.	ń	*ɲ̊*	ń	*J̃*
Ukończyła kurs dworskiego tańca.	ń	*ɲ̊*	ń	*j̃*
Realisation of *em(n), om(n)* word-initially in loanwords				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Komfort onieśmielał ich coraz częściej.	om	*ɔm*	ą	*ɔw (/ɔ̃)*
Przybył właśnie pan konsul.	on	*ɔm*	ą	*ɔw (/ɔ̃)*
Nimfa przewróciła mu w głowie.	im	*im*	im	*iŋ*
Widać w tym dziele niezwykły kunszt.	un	*un*	un	*uŋ*
Powiedz, jaki to ma sens.	en	*ɛn*	ę	*ɛw (/ɛ̃)*
Combinations of letters *trz, strz*				
Sentence	Pronunciation variant 1		Pronunciation variant 2	
Ona ma już trzeciego męża.	trz	*t̪ʂ*	cz	*t̪͡ʂ*
Potrzeba matką wynalazków.	trz	*t̪ʂ*	cz	*t̪͡ʂ*
Jeszcze tam przytrzymaj!	trz	*t̪ʂ*	cz	*t̪͡ʂ*
Nie może powstrzymać kaszlu.	strz	*st̪ʂ*	szcz	*ʂt̪͡ʂ*
Basia jutro cię ostrzyże.	strz	*st̪ʂ*	szcz	*ʂt̪͡ʂ*

Table 4. The sentences and the occurring pronunciation variants.

All the recordings were carried in a professional studio with the following equipment:

- overhead microphones - DPA 4066 omnidirectional headset microphone,
- stationary microphones - Neumann TLM 103 condenser,
- large diaphragm microphone,
- Cakewalk Sonar X1 LE Software,
- Roland Studio Capture hardware.

The speech synthesis was prepared using the revised version of the Polish HMM-based speech synthesiser which was built using the HMM-based Speech Synthesis System (HTS). The speech samples were trained on a Polish BOSS unit selection synthesiser corpus which was also used as the text analysis tool for HTS. The default speech analysis configuration included data sampled at 48 kHz, 25ms frame and 5 ms shift using Hamming window (Jankowska, Kuczmarski, & Demenko, 2021).

The speech corpus consists of two main parts. The first is model recordings: male, female and a synthetic voice in two variants of pronunciation. The second part includes 60 recordings of baseline production (casual reading the sentences by participants) and shadowed sentences. The whole corpus contains 66 files with a total of 1,506 spoken sentences recorded (Jankowska, Kuczmarski, & Demenko, 2021).

The convergence analysis in the recordings included perceptual methods aimed at checking whether the speaker changed the pronunciation variant under the influence of a natural or synthesised stimulus and whether there was a change in the execution of the whole sentence. Convergence was observed in the case of shadowing both natural and synthesised stimuli, however in the case of natural stimuli the convergence level was higher. The results show that accent and intonation was reproduced exactly the same way as the natural stimuli. For the task of shadowing synthesised stimuli, the speech was perceived and reported as flat. In addition, an F0 analysis was performed in all statements. The F0 analysis included statistics (mean, standard deviation, median, minimum and maximum value, 25th and 75th quantile) for baseline production, synthetic and natural speech shadowing. No

statistically significant changes were observed under the influence of any stimulus (Jankowska, Kuczmarski, & Demenko, 2021).

5.2 Non-linguistic convergence in Polish

Non-linguistic communication convergence includes analysis of facial expression, gaze, posture, body movements, including hand gestures and head movements. Very little research has been done on convergence analysis of non-linguistic aspects of communication for Polish speakers. This type of study was conducted by Karpiński et. al. (2018) featuring Polish and German teenagers. The analysis of gestural behaviours included such aspects as gesture frequency, function (referential vs. pragmatic), duration of gesture strokes and features of strokes and their co-occurrence and recurrence (Karpiński, Czoska, Jarmołowicz-Nowikow, & Juszczyk, 2018). The authors of the study on the adjustment of gestures in task-oriented dialogues with the participation of Polish and German teenagers proved that more gestures were made by Polish than German children. The analysis of the duration of the original and repeated strokes of gestures for gestures with the same function showed similarities in the dialogues of the Polish speakers. Polish-speaking participants show a higher average duration of original and repeated strokes in reference gestures in competitive dialogues (Karpiński, Czoska, Jarmołowicz-Nowikow, & Juszczyk, 2018).

6 Convergence studies in view of language technology

6.1 Language Technology

The growing interest in computer natural language processing and the rapid development of the field of computer science dealing with artificial intelligence have contributed to the emergence of new fields of study known as Computational Linguistics, Natural Language Processing (NLP) or Language Technology (Agerri, et al., 2021). Research on Natural Language Processing began with machine translation efforts in the mid-twentieth century. The first translation systems used a simple mechanism of selecting word equivalents from individual vocabularies and changing the order of

words. These systems did not take into account numerous important elements of languages, e.g. lexical ambiguity, inflection, grammatical structures (Hutchins, 1995). Scientists realised that natural language processing is a much more complex process that requires research in the field of language theory. The publication of Chomsky's Syntactic Structures in 1957 was a milestone. In his work, Chomsky presented a formalised general theory of linguistic structure and syntactic investigation which was crucial for language analysis (Chomsky, 1957). Thanks to the development of syntactic theory of language, parsing algorithms were created, and the research community found that there is very little to do with creating a high-quality, fully automatic machine translation system (Bender, Sag, & Wasow, 1999). However, the then technological advancement and the state of knowledge did not allow for the creation of such a system in the 1950s. Another work of Chomsky, which had a significant impact on the field, was the transformation model of linguistic competence, proposed in 1965 (Chomsky, 1965). However, this work did not allow for major changes enabling not only syntactic but semantic analysis, which remained a significant problem. In response to Chomsky's theory, many other concepts and works on grammar were created, for example Fillmore's case grammar (Mazarweh, 2010), Quillian's semantic networks (Quillian, 1967), Schank's conceptual dependency theory (Schank, 1969), Wilks' preference semantics (Wilks & Fass, 1992), Kay's functional grammar (Kay, 1979). All activities were aimed at explaining and systematising syntactic anomalies and proposing semantic representations.

At the same time as linguists were working on the development of language theory, prototype systems such as the ELIZA dialogue system were being developed. The system developed by Weizenbaum at the MIT Artificial Intelligence Laboratory was designed to simulate a psychologist's conversation with a patient (Weizenbaum, 1966). The operation of the program is simple, based on the analysis of patterns in the user's sentences and building the question by rearranging words and replacing keywords. Another example of a system developed at the same university is SHRDLU by Terry Winograd (Winograd, 1971). Initially, it was a language parser that

enabled the user - computer to dialogue in English. SHRDLU was executing user commands - moving objects in a simulated world made of blocks. The program was very simple and it took about 50 words to execute user commands such as „move blue block". It was the first program found to be convincing in terms of computer's understanding of natural language. Another example of one of the first dialog systems is PARRY, created by Kennth Colby in 1972. Implemented by a psychiatrist, PARRY attempted to simulate a person with paranoid schizophrenia (Colby, Weber, & Hilf, 1971). An example of a non-therapeutic system is LUNAR - the database interface to lunar rocks samples information which made use of the augmented transition network and procedural semantics (Woods, 1978). Subsequent research in the field of natural language processing has focused on semantic issues, discourse analysis, and the planning and purpose of communication. Research has been carried out on the structure of discourse and the analysis of task-oriented dialogues. At the same time, significant progress was observed in natural language generation technology. McKeown created the text discourse planner that generated coherent responses online (McKeown, 1985). Created by McDonald's, MUMMBLE generated short texts using theoretical predicates (McDonald & Pustejovsky, 1985).

In the 1990s, language technology, NLP began to develop very dynamically thanks to the access to large resources of research materials in electronic form, the development of computers, high computing power and the development of the Internet. Much progress has been made in the field of NLP thanks to new resources, tools and applications. At that time, work began on specialised linguistic resources that enabled the development of modern technologies. Such resources include thesauri or annotated language corpora, one of the main results of which is WordNet (Miller, 1995). The effect of many-years' work of scientists in many fields of science is visible nowadays in everyday life. In 2010, a radical technological change in NLP was observed. In 2011, a group of researchers form NEC Laboratories America presented a multilayer neural network corrected by back propagation that was able to solve various sequential labelling problems (Collobert, et al., 2011). Gradually, data-driven systems started replacing rule-based systems, and

today it is hard to imagine an NLP system that does not include any machine learning based component. The availability of large volumes of texts, along with advances in unsupervised machine learning and the development of high-performance hardware (in the form of graphical processing units - GPUs) has enabled the development of a very effective deep learning system in a variety of application areas (Agerri, et al., 2021).

Thanks to advanced technology, we can use various types of generally available dialogue systems on computers, smartphones and other devices. On a daily basis, we use machine translation, automatic e-mail classification, language auto-correction and many other assistants that work more and more effectively. In mobile devices, we can use voice systems such as Siri[19] in Apple devices or Bixby[20] for Samsung. Amazon offers voice-controlled smart speakers called Alexa[21] that work quite similar to Google Nest with Google Assistant voice user interface.

6.2 Current challenges in language technology

Such significant progress in the development of tools using language technologies has been achieved mainly thanks to modern methods such as machine learning, artificial intelligence and corpus research. Thanks to extensive and high-quality language materials (annotated corpora), language systems are created with which you can conduct conversations in natural language at a level very close to human. Due to the popularisation and easy availability of voicebots, chatbots, voice assistants and other tools based on human-computer interaction in natural language, the requirements for the quality of these systems are growing. Current requirements include aspects such as processing spontaneous, emotional, whisper, disfluent, noisy, distorted speech. It is also important to be able to process the language regardless of the speaker's individual characteristics and adapt it to a specific communication situation (e.g. adapting vocabulary to a specific topic). Synthesised speech should have appropriate paralinguistic parameters so that

[19] https://www.apple.com/siri/
[20] https://www.samsung.com/us/apps/bixby/
[21] https://developer.amazon.com/en-US/alexa

it is as complex as possible to the natural and pleasant to the human interlocutor (Demenko, 2020). The question of the specificity of different languages also has to be taken into account. Many studies on convergence have been conducted on English, German and other materials, much less in Polish. The methods and degree of communication adjustment should also be appropriate to the communication situation and the interlocutor at the linguistic and paralinguistic level. Speech and language technology models cannot achieve the expected sophistication and approximation to the natural, human-human like level of interaction if they omit the interactive process of taking place during a conversation. In addition to the transmission of information, systems must take into account interpersonal aspects and the specific communication situation (Demenko, 2020).

Research on convergence in dialogues may contribute to the creation of algorithms that will be able to simulate natural conversations at a level very close to human-to-human communication. However, in order to gain the knowledge to describe and model the processes of communicative adjustment, this process must be thoroughly investigated both in linguistic and paralinguistic aspects.

7 Summary

In the literature, numerous examples of research on convergence and attempts to explain the mechanisms that occur during verbal interaction can be found. The effect of these psychological and social mechanisms is the language, whose study can explain the source of communication adaptation and enable its modelling. There are theories that try to explain the sources of communication alignment based on sociological and psychological assumptions. CAT assumes that communication adaptation is a partially conscious strategy used by interlocutors to achieve common and individual goals. IAM adopts the less conscious priming mechanism as the source of convergence. The existence of communication alignment is scientifically proven, and there is no doubt that during dialogues, interlocutors adjust their

behaviour at various levels. Many internal and external factors influence the type and intensity of these behaviours.

Communication alignment was studied in various aspects and levels of verbal and non-verbal communication. Research covers the linguistic, paralinguistic and non-linguistic levels. Researchers used a variety of methodologies, features and measures to study speech assimilation in acoustic and phonetic aspects, lexical aspects, facial expressions and gestures. Studies on communication adjustment have been conducted in various languages, but the vast majority of them concern English. Much less research has been conducted in this regard for the Polish language and they concern mainly phonetic and acoustic convergence as well as gestures and facial expressions. Little is known about verbal behaviour and lexical adjustment in various communicative situations in Polish. The study of phonetic-acoustic convergence in Polish was based on methodologies also used for other languages. In the literature, there are various approaches and extensive methodologies for analysing and calculating lexical convergence for a variety of languages. An example is LSM, which is a popular language matching methodology currently in use for the English language. LSM has been recognized as a reliable and effective tool for calculating the level of lexical mimicry. There is even an LSM calculator available online[22]. Due to the differences and the specificity of individual languages, methodologies for this type of analysis need to be adapted. However, it is assumed that they may find applications and be used in research for the Polish language.

Communication alignment analysis is a research area whose scope goes beyond the field of linguistics. The sources and effects of this phenomenon are directly related to psychological, cognitive and sociological mechanisms. On the other hand, a good understanding of how, at what levels, under what circumstances and with whom people change their communication behaviour can enable the modelling of more natural dialogues. Nowadays, when modern technologies are developing very dynamically, the requirements for the quality of systems and interfaces based

[22] http://www.utpsyc.org/synch/input.php

on natural language are also growing. While this type of technological solutions are at a high level for the English language, the Polish version needs improvement. In order to achieve high-quality, intelligent systems based on communication in natural language, both spoken and written, theoretical knowledge of human-to-human communication is needed. On this basis, it will be possible to model high-quality, reliable and useful human-computer communication.

III Lexical convergence in Polish dialogues

1 Own research on lexical convergence

All the activities and analyses described below were developed by the author of this work. The linguistic material (Harmonia corpus) created by the research team of Professor Grażyna Demenko and PhD Jolanta Bachan was partially used in this research. The transcription and annotation of the entire corpus of recordings, which was originally used for phonetic-acoustic analysis, was done by the author of this work in the Praat program. In order to perform the lexical convergence analysis, appropriate processing of data and files was required. This process as well as the methodology and tools used are described in detail in the following chapters.

1.1 Research material

The research material created as part of the *Automatic analysis of phonetic convergence in speech technology systems*[23] project (no. 2014/14/M/HS2/0063) founded by National Science Centre Poland was used for this study. The project aimed to objectively assess phonetic convergence in human-human and human-computer interactions. The specific aim of the project was to present the quantitative description of accommodation phenomena appearing in different properties of speech, such as acoustic, prosodic, temporal and spectral. The research was aimed at making progress, presenting theoretical foundations and proposing solutions for use in speech technology. The linguistic material developed within the project consists of scenario-based recordings that include both individual speakers and dialogues and has been published under the name Harmonia Corpus. For the purposes of the current study, only tasks involving two people in a conversation were used. As part of the project, phonetic convergence analyses were carried out, the results of which were published in the book *Phonetic Convergence in Spoken Dialogues in View of Speech Technology Application* (Demenko,

[23] http://wczt.pl/technologia_mowy/speech_convergence.html

2020). Scenarios, recordings and annotation instructions were developed by PhD Jolanta Bachan, Prof. Grażyna Demenko and MSc Mariusz Owsianny (Bachan, Owsianny, & Demenko, 2017). Transcription, segmentation and multi-level annotation of the recordings was made by the author of this work, in accordance with the guidelines.

1.1.1 Tools and set-up

The recordings took place in a professional studio at the Faculty of Modern Languages at the Adam Mickiewicz University in Poznań. The first speaker's voice was recorded in a sound insulation cabin (anechoic chamber) and the second was acoustically separated by sound-absorbing panels in the corner of the studio. Four professional microphones were used for recordings:

- two overhead microphones (DPA 4066 omnidirectional headset microphone)
- two stationary microphones (condenser, large diaphragm studio microphone with cardioid characteristic – Neumann TLM 103).

The software used for the recordings was Cakewalk Sonar X1 LE software. This setup provided 4 mono channels of recordings, 2 for each speaker, at 44.1 kHz sampling frequency and 16 bit depth (Bachan, Owsianny, & Demenko, 2020).

1.1.2 Participants

32 persons took part in the recordings, of which 16 were women and 16 were men. In addition, the recording of the last three tasks was attended by the recording leader - PhD Jolanta Bachan (called Teacher for the sake of simplicity). People of the same sex participated in the dialogues. The youngest participant was 19 and the oldest 58, but most participants were under the age of 29. The average age was 27 years (Bachan, Owsianny, & Demenko, 2017).

1.1.3 Recordings scenario

The recording scenario consisted of 16 tasks including statements of individual participants, dialogues of participants and dialogues of participants with the Teacher. The first tasks consisted of reading and repeating sentences.

Others required cooperation in order to perform a specific task. Some of the tasks were supposed to evoke specific emotions and attitudes of the interlocutors towards each other's opinions and the objects they were supposed to talk about (Bachan, Owsianny, & Demenko, 2020). The exact content of individual tasks is presented in Appendix V.2 of this work.

1.1.4 Annotations

Harmonia corpus was annotated at several linguistic levels, for which the Praat program was used. For Tasks 1-3, an orthographic (transcription) and prosodic annotation was performed. For Tasks 4-16, segmentation was performed into statements with separate speakers and seven layers:

- ort_A - transcription with prosodic information, speaker A
- DA_A - dialogue act, speaker A
- info_A - information about the speaker (e.g. neutral, excited, irritated, etc.)
- ort_B - transcription with prosodic information, speaker B
- DA_B - dialogue act, speaker B
- info_B - information about the speaker (e.g. neutral, excited, irritated, etc.)
- agree / disagree - fragments of dialogue in which the speakers agree or disagree.

The annotation instructions were prepared by the team implementing the Harmonia project, mainly Prof. Grażyna Demenko and PhD Jolanta Bachan. The annotation was made by the author of this work.

The main difficulties related to the manual annotation of speech recordings in Polish were mainly: 1) pronunciation errors, 2) pronunciation corrections, 3) unclear, interrupted utterances, uttering fragments of words, 4) incomprehensible utterances, 5) insertion of words from foreign languages, 6) non-linguistic sounds from the speaker (clicks, coughs, fillers „yyy", laughter etc.), 7) distractions from the speaker, 8) language manners of the speakers. Pronunciation errors are common and are often corrected immediately by the speakers. Slippages, unconscious mistakes or repetitions are very common, especially in spontaneous speech. In annotation, this problem can be solved in two ways. The first way is to write down the misspelt or wrong words exactly as they were spoken. The second solution is

to write down the statements in the correct forms by adding an appropriate mark, e.g. „*". The same is true of slurred and broken statements. In the case of incomprehensible statements, it is necessary to put a symbol that replaces the given phrase. Often there are also inclusions from foreign languages in the statements. In such a case, several solutions can be used, for example, write the word phonetically in International Phonetic Alphabet (IPA), write it correctly in the source language with the appropriate marking and information about the language, or write down the wording in Polish (e.g. *playback - plejbek*). Any non-linguistic sounds from the speaker and disturbance from the environment should also be marked with appropriate symbols. The greatest difficulty is the problem of intonation-related linguistic manners. Oftentimes, individuals tend to utter sentences with increasing intonation when context and other indications suggest that the sentence is affirmative. The ambiguity of increasing intonation is problematic due to its function – posing a question. When the intonation increases and the sentence is certainly not a question the annotator has to make a decision which tag to use. Therefore, it seems necessary to create an additional label to mark sentences in which the intonation is clearly increasing, but other features clearly indicate that the sentence is a statement.

All the issues were foreseen and planned by the authors and the appropriate tags were specified in the annotation instructions. Spelling mistakes, i.e. words distorted by the speaker in a way that does not hinder their understanding, were written with an asterisk, e.g. * word. This designation was used in the case of obvious errors, omissions of a letter or syllable, and not in the case of using different variants of the pronunciation of a given word, i.e. saying [em] or [e ~] at the end of a verb in a singular present tense were not treated as errors. Words and phrases in languages other than Polish were written as follows: [len = EN] welcome, where *len* means language, *EN* means English. Appropriate codes were used for other languages, e.g. for German – DE, French – FR, Latin – LA, etc. In case of difficulties with understanding the speech sequence, two asterisks were introduced **. Due to the high quality of the recordings and excellent soundproofing conditions in the studio, there was no need to use markings

about permanent and temporary disturbances. In the case of fillers, laughter and other extra-linguistic sounds made by speakers, e.g. clear breathing, the signs [fil = m], [laugh], [spk = b] were used. All onomatopoeias (e.g. *aaa*, *aha*, *ouch*, etc.) were written as words with additional FPW annotation for hesitant onomatopoeia and FPE for expansion. In spite of such a detailed instruction, there were questionable statements, the annotation of which required reflection as well as meticulousness and consistency of the annotator.

- **+** syllable/word emphasis (not the stress)
- **/** phrase boundary
- **//** strong phrase boundary of an affirmative sentence
- **//?** strong phrase boundary for interrogative sentences
- **//!** strong boundary of an exclamation sentence phrase
- **$** grammatical sentence (complete utterance); ambiguous prosodic features
- **{** non-grammatical boundary
- **!!** word or phrase clearly highlighted in emphasis /expression
- **/..** addition
- **/@** backchannel
- **/~** utterance interrupted at the end
- **~/** utterance interrupted at the beginning

Figure 2 shows a screenshot of a portion of the recording annotation in Praat.

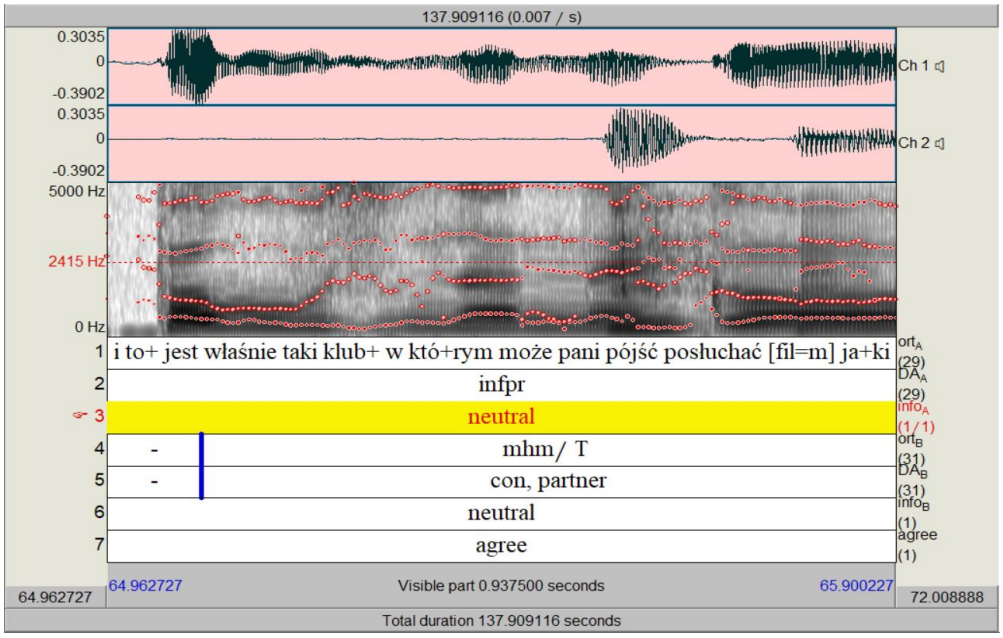


Figure 2. Fragment of annotated Harmonia Corpus in Praat.

1.1.5 Subcorpus of scenario-based dialogues

In order to examine and prove the theses of the dissertation, recordings of dialogues (not one-person utterances, i.e. reading, repeating after the model recording) were needed which is why some of the recordings were rejected (Tasks 1-4). It was also important that the participants were of a similar age. Due to the fact that the majority of the dialogues were conducted with students, it was decided that the research material would consist of recordings of people aged 20-30. Therefore, it was necessary to reject some of the recordings (N01-03, N05). Eventually, a corpus was created consisting of 180 recordings based on 12 scenario tasks with the participation of 25 Polish native speakers who were asked to analyse the lexical convergence in dialogues. The 24 people who took part in the recordings were matched in pairs and conducted conversations in 12 tasks. In three tasks (14-16), each person spoke separately with the person supervising and conducting the recordings (named Teacher for simplicity). Recordings were based on tasks:

1. Z05 (Speaker 1 and Speaker 2) Work with your partner to find 3 differences between the pictures.
2. Z06 (Speaker 1 and Speaker 2) One speaker assumes the role of a hotel receptionist, the other – a guest. A guest calls the hotel and asks

for information on how to get to the hotel from the station. Both interlocutors have access to the map of the city.

3. Z07 (Speaker 1 and Speaker 2) Same as in Z06 – speakers switch roles.
4. Z08 (Speaker 1 and Speaker 2) One interlocutor plays the role of a tourist information employee of a large city and is supposed to encourage the interlocutor (tourist) to take advantage of one of three proposals for spending an evening in the city. If the employee convinces the tourist, one will get a high reward from the boss.
5. Z09 (Speaker 1 and Speaker 2) Same as in Z08. – speakers switch roles.
6. Z10 (Speaker 1 and Speaker 2) One interlocutor plays the role of a tourist information employee of a large city and is supposed to encourage the interlocutor (tourist) to take advantage of one of three proposals for spending an evening in the city. The situation is difficult because the city has raised the alarm about terrorist attacks. The employee's task is to convince the tourist to choose the safest option.
7. Z11 (Speaker 1 and Speaker 2) Same as in Z10 – speakers switch roles.
8. Z12 (Speaker 1 and Speaker 2) The interlocutors are to exchange opinions about the photos of the artwork they see, to be in agreement and to praise the art form.
9. Z13 (Speaker 1 and Speaker 2) The interlocutors are to exchange their opinion about the photos of the artwork they see, to be in agreement and to criticise the art form.
10. Z14 (Speaker 1 and Teacher, Speaker 2 and Teacher) Interlocutors are to exchange their opinion about the photos of the artwork they see, to agree and to praise the art form.
11. Z15 (Speaker 1 and Teacher, Speaker 2 and Teacher) The interlocutors are to exchange opinions about the photos of the artwork they see, to agree and to criticise the art form.

12. Z16 (Speaker 1 and Teacher, Speaker 2 and Teacher) A conversation between a supporter and an opponent of provocation in art. Everyone stands by their own opinion.

Tasks 5-7 were based on cooperation between speakers, information providing and achieving a common goal. Tasks 8-11 were supposed to evoke expression, persuasiveness, and 12-16 were provocative and arousing emotions. In Table 5, the task columns are grouped and color-coded as described.

Table 5 presents the summary of recordings, tasks and participants in each dialogue. The codes used for one speaker's speech in one dialogue are formed by the recording number, the task number, and the speaker information, e.g. N04_Z05_SPK1 (N stands for recording (pl. “nagranie”), Z stands for tasks (pl. “zadanie”), P stands for pair (pl. “para”), SPK stands for “speaker”

	Cooperation, common goal			**Expression, persuasiveness**				**Provocative, arousing emotions (students)**		**Provocative, arousing emotions (students with Teacher)**					
	Task 5	**Task 6**	**Task 7**	**Task 8**	**Task 9**	**Task 10**	**Task 11**	**Task 12**	**Task 13**	**Task 14**	**Task 15**	**Task 16**	**Task 14**	**Task 15**	**Task 16**
N04	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N06	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N07	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N08	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N09	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N10	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N11	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N12	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N13	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N14	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N15	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher
N16	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 SPK2	SPK1 Teacher	SPK1 Teacher	SPK1 Teacher	SPK2 Teacher	SPK2 Teacher	SPK2 Teacher

Table 5. The summary matrix of recordings, tasks and participants in each dialogue.

This page is intentionally left blank

1.2 Methodology

The literature review revealed many approaches, methodologies, measures and parameters used in communication alignment research in different languages. Few similar studies have been conducted for the Polish language, especially in the field of lexical convergence. Considering the scope and a certain niche in the study of lexical convergence for the Polish language, the focus was on this aspect. The following analyses were carried out to shed some light on lexical alignment in Polish dialogues:

- Formality level,
- Collaborative effort – the length of the sum of each speaker's speech in each dialogue,
- Lexical choices – analysis of word selection and lexical mimicry level,
- Language Style Matching – analysis of non-content words in Polish.

The level of formality and politeness in the language is expressed in several ways in Polish and at the same time is an important element of communication. The analysis of the forms used is aimed at showing how peers address each other in various communication situations and whether they change the forms used when talking to the Teacher. Collaborative effort analysis can show tendencies for longer utterances in women's or men's dialogues, individual tendencies for longer or shorter utterances, and adjusting the amount of spoken words to the interlocutor (conversation with a partner and with a Teacher). Lexical choices analysis seems to be the most obvious measure of lexical convergence, which is why it was carried out and described. The manual analysis of lexical mimicry was supported by the results of the corpus analysis - the number of individual parts of speech in the statements of individual speakers in each task. Language Style Matching is quite a popular methodology used to study lexical mimicry in the English language. There are many examples of its application and interpretation in the

literature[24]. The possibility of using this methodology for the Polish language seems to be possible, promising and applicable in various types of research.

1.2.1 Formality measures

Two parameters were used to determine the formality level - the form of greeting and the form of addressing the interlocutor used in dialogues. Greetings are honorifics that begin or signal a meeting, one of the most important forms of linguistic politeness. There are two main functions of the greeting: (1) an indication that we see a familiar person, (2) informing the partner that we will be talking to him (Bańko, 2022). People with close relationships, friends, use greetings such as "Cześć" (Eng. "Hello") which is the most frequent form, used by interlocutors of all ages, and other such as "Witaj", "Siemasz", "Siema", "Siemka", "Siemanko", "Elo", "Hej", "Piątka", "Strzała" etc. Their choice often depends on the age of the interlocutors. A more formal form of greeting used by strangers is "Dzień dobry" (Eng. "Good morning" – literally "Good day"). The form of greeting "Witam" (Eng. "Greetings") is gaining more and more popularity and is used as an intermediate form between the unofficial "Cześć" and the official "Dzień dobry". In addition, religious forms such as "Niech będzie pochwalony Jezus Chrystus" (Eng. "Praise be to Jesus Christ") or "Szczęść Boże" (Eng. "God bless") are used.

In Polish, there are two main forms of addressing interlocutors, depending on the level of acquaintance and relationship. In the case of friends of the same level (e.g. the same age), direct phrases are used – "Ty" (Eng. "You"). If the conversation is between adults who do not know each other, the form "Pan" (Eng. "Mister") or "Pani" (Eng. "Mrs") is used. In cases of unequal relations, e.g. child-adult, the younger interlocutor uses the polite form and the older – the direct one.

The study included an analysis of greetings in individual tasks and dialogues, as well as the form of addressing each other by the interlocutors. More formal forms of greetings and personal expressions are considered to

24 i.e. https://psyarxiv.com/yz4br/

indicate a higher level of dialogue formality. It was important in which tasks what forms were used and whether the interlocutors used the same polite forms or differentiated them within one dialogue.

1.2.2 Collaborative effort measures

Collaborative effort was the total number of words uttered by both participants. In other words, it is the length of the dialogues and the total length of both interlocutors. The word is defined as *a unit of expression which has universal intuitive recognition by native-speakers, in both spoken and written language* (Crystal, 2008). This measure can indicate whether the interlocutors' participation in the dialogue is well-balanced or one of the interlocutors is dominant. In addition to predispositions, natural tendencies of speakers, differences may occur depending on the communication situation, the topic of conversation and the relationship between the interlocutors.

1.2.3 Lexical choices

Lexical choices analysis was carried out by analysing the transcription of whole texts and by analysing content words (nouns and verbs) in a manually tagged corpus. In addition, attention was paid to adjectives and adverbs, especially in Tasks 12-16, where the interlocutors are tasked with agreeing or disagreeing, positively or negatively assessing controversial art. This analysis aims to show whether the interlocutors chose the same terms for objects, phenomena, activities or diversified, introduced new concepts during one conversation.

1.2.4 Language Style Matching

After reviewing and analysing the methodology of computing at the lexical mimicry level, it was decided that LSM would be the most suitable for this study. This is mainly due to the focus of this methodology on lexis, taking into account parts of speech and the content and non-content words. The recordings, and thus the transcriptions of the dialogues in the selected linguistic material for study, are not suitable for dividing them into prime and target. The LSM methodology was used in many studies and their results gave grounds for drawing interesting conclusions.

The LSM methodology is adapted to the specificity of the English language. The analysis includes vocabulary and grammatical categories that do not have identical equivalents in Polish. Polish grammar differs significantly from English grammar, so the LSM methodology needs to be adapted.

1.2.4.1 LSM methodology adaptation for Polish

The adaptation of the LSM methodology includes a holistic approach analogous to the English version. The aim was to create the most faithful reflection of language categories while maintaining the correctness in accordance with the grammar of the Polish language.

Auxiliary verbs have the syntactic and morphological features of verbs, but they perform only grammatical functions and have no meaning of their own. In Polish, this function is performed by the verb "być" (Eng. "to be"). English articles have no equivalents in Polish. Prepositions and conjunctions are also considered parts of speech in Polish.

The issue of pronouns in Polish is not obvious and linguists and Polish philologists have presented different approaches to categorizing pronouns. For example, Zygmunt Saloni (1974) associated pronouns with related word root classes, which resulted in the elimination of pronouns as a grammatical category. Other approaches consider substantival pronouns e.g. "ja","ktoś", "nikt" (Eng. "me", "someone", "nobody") and other pronouns are treated as adjectives or numerals (Grzegorczykowa, 1984). The Encyclopedia of General Linguistics (1999) identifies the following types of pronouns: (1) possessive, (2) indefinite, (3) definite, (4) personal, (5) interrogative, (6) indicative, (7) relative, (8) reflexive (Polański, 1999). Current approaches take into account the pragmatic function of pronouns – *they link the elements of the utterance with the situation in which the utterance occurs, referring to the contemporary world of the sender and receiver* (Nagórko, 2007). This function is also called deictic. Nagórko (2007) categorizes pronouns into: (1) personal, (2) reflexive, (3) indefinite, (4) negative, (5) demonstrative, (6) possessive, (7) interrogative-relative. Szczepankowska (2012) proposed a detailed typology of Polish pronouns,

taking into account (1) personal pronouns referring only to persons, (2) demonstrative pronouns, (3) possessive personal pronouns, (4) possessive pronouns referring to persons or non-persons, (5) interrogative pronouns relative, (6) indefinite pronouns (6.1) formed from interrogative pronouns by means of particles "-ś", "-kolwiek" or words "bądź", "byle", (6.2) lexemes indicating an indefinite object only for the recipient, (6.3) quantifying: ditributively, collectively, (6.4) negatives. The author additionally distinguishes each category into pronouns replacing the names of persons, pronouns replacing the names of persons and non-persons, and replacing the names of the way, place, path, direction, time. Wierzbicka-Piotrowska (2011) also distinguishes a category of generalising pronouns (positive and negative), which are also included in a separate category in this work.

For the purposes of this work, it was decided to use the division of pronouns according to their functions, disregarding their inflectional and syntactic properties:

- possessive pronouns
- personal pronouns
- reflexive pronouns
- interrogative pronouns
- indefinite pronouns
- demonstrative pronouns
- negative pronouns
- relative pronouns
- generalising pronouns.

The full specification with examples is presented in Table 6.

The LSM methodology takes into account personal pronouns and indefinite pronouns. It also does not include pronouns indicating possession, for example "mine", "yours", "hers", "theirs". Accordingly, in the LSM version for the Polish language, possessive pronouns will not be taken into account. English negations can correspond to Polish negative pronouns and some particles. Particles (pl. partykuły) are invariable parts of speech that have no syntactic function in a sentence. The Polish particles include: "no",

“niech”, “by”, “nawet”, “właśnie”, “bodaj”, “to”, “tak”, “nie”, “tu”, “lada”, “niech”, “chyba”, “ci”, “co”, “niechaj oby”, “tam”, “już”, “tylko”, “czy”.

The common adverbs category, which in English LSM methodology includes words such as *hardly*, *often*, etc. has to be addressed and well defined for Polish. The author of the work decided to adopt the methodology based on frequency lists made on large linguistic data sets. In order to create a list of popular Polish adverbs, the frequency list for Polish words created by the G4.19 Language Technology Group of the Wrocław University of Technology was used. The frequency lists were extracted from large corpora of texts, which include e.g. the IPI PAN corpus, the Rzeczpospolita corpus, Wikipedia (content from early 2010) and a collection of large documents downloaded from the Internet (around 1,8 billion tokens in total). The lists are available for download as txt files on the research team's website[25].
The frequency_list_orth.txt file, which contains 1,048,576 lexical units, was used to create the adverbs list. All words are annotated with a part-of-speech tag and the number of occurrences. From all words, adverbs (12,287 items) were selected and sorted by the highest number of occurrences. For the Polish version of the LSM methodology, the first 100 of the most common adverbs were used, taking into account the inflected versions of words, e.g. “bardzo” and “bardziej” (Eng. “much” and “more”) are counted as separate words.

Table 6 contains all grammatical categories and corresponding word lists applied to the Polish version of the LSM analysis.

Category	**List of items**
Auxiliary verbs	*być, bądź, bądźcie, będą, będący, będę, będzie, będziecie, będziemy, będziesz, byli, byliby, bylibyście, bylibyśmy, byliście, byliśmy, był, była, byłaby, byłabym, byłabyś, byłam, byłaś, byłby, byłbym, byłbyś, byłem, byłeś, było, byłoby, były, byłyby, byłybyście, byłybyśmy, byłyście, byłyśmy, jest, jestem, jesteś, jesteście, jesteśmy, są*
Prepositions	*na, o, w, z, za, u, ku, od, do, bez, pod, przed, nad, dla, między, przez, po, poprzez, pomiędzy, ponad, wśród, spośród, obok, około, oprócz, pomimo, zza, znad, wbrew, względem, pod względem, niby, dlaczego, poza*

[25] http://nlp.pwr.wroc.pl/en/tools-and-resources/resources/frequency-list

Category	**List of items**
Conjunctions	*a, a nawet, aby, acz, aczkolwiek, albo, albo i, bo, albowiem, ale, ani, ażeby, bądź, bo, bowiem, by, choć, chociaż, chyba że, czy, czy raczej, czyli, dlatego, gdyż, gdyby, gdybym, gdybyś, i, i to, jak, jakby, jakbym, jakbyś jednak, jednakże, jeśli, jeżeli, lecz, lub, lub też, mianowicie, mimo że, natomiast, ni, oraz, pomimo że, ponieważ, przeto, tedy, to jest, to znaczy, toteż, tudzież, tylko, więc, zaś, zatem, że, żeby, żebym, żebyś, skoro*
Particles	*no, niech, by, nawet, właśnie, bodaj, to, tak, nie, tu, lada, chyba, ci, co, niechaj oby, tam, jeszcze, już, tylko, czy, też, jakby, także, przecież, naprawdę, bym, byście, byśmy, bodaj, niech, raczej, chyba, niestety, jeszcze, już, także, również, też, bynajmniej, dopiero, przynajmniej*
Common adverbs	**List of 100 the most frequent adverbs extracted from the general word frequency list for Polish[26]:** *bardziej, bardzo, bezpośrednio, blisko, bliżej, całkowicie, chętnie, ciągle, cicho, ciężko, często, częściej, częściowo, dalej, daleko, dawniej, dawno, długo, dłużej, dobrze, dodatkowo, dokładnie, doskonale, dużo, głęboko, głośno, głównie, gwałtownie, jednocześnie, jedynie, jutro, krótko, lekko, lepiej, łatwo, łącznie, mało, mniej, mocno, nagle, najbardziej, najczęściej, najlepiej, najmniej, najwyraźniej, następnie, niedawno, niewątpliwie, niezależnie, niezwykle, nowo, obecnie, oczywiście, odpowiednio, osobiście, ostatecznie, ostatnio, ostrożnie, pewnie, pewno, początkowo, podobnie, podobno, ponownie, poważnie, później, późno, praktycznie, prawdopodobnie, prosto, rano, równocześnie, rzadko, rzeczywiście, specjalnie, spokojnie, sporo, stale, szczególnie, szczerze, szeroko, szybciej, szybko, trudno, uważnie, wcześniej, wielokrotnie, więcej, właściwie, wspólnie, wyłącznie, wyraźnie, wysoko, wyżej, zdecydowanie, zgodnie, znacznie, zupełnie, zwykle, źle*
Indefinite numbers	*ciut, dużo, gros, ile, ilekolwiek, ileś, ileż, iloma, ilu, ilukolwiek, iluś, iluż, kilka, kilkadziesiąt, kilkanaście, kilkanaściorga, kilkanaścioro, kilkaset, kilkoma, kilkorga, kilkorgiem, kilkorgu, kilkoro, kilku, kilkudziesięcioma, kilkudziesięcioro, kilkudziesięciu, kilkunastoma, kilkunastu, kilkuset, kupa, łaska, malutko, mało, najwięcej, nastoma, nastu, naście, naścioro, nie więcej, niedużo, niemało, niewiela,*

[26] The frequency list for Polish words created by the G4.19 Language Technology Group of the Wrocław University of Technology (http://nlp.pwr.wroc.pl/en/tools-and-resources/resources/frequency-list).

Category	**List of items**
	niewiele, niewieloma, niewielu, od groma, parę, parędziesiąt, parenaście, paręset, paroma, paru, parudziesięciu, parunastu, paruset, trochę, troszeczkę, tyle, tyleż, tyloma, tylu, tyluż, wiela, wielą, wiele, wieleset, wieleż, wieloma, wielu, wieluset, wieluż, więcej
Indefinite pronouns	*coś, czegoś, czemuś, coś, czymś, ktoś, kogoś, komuś, kogoś, kimś, cokolwiek, czegokolwiek, czemukolwiek, czymkolwiek, ktokolwiek, kogokolwiek, komukolwiek, kimkolwiek, jakiś, jakaś, jakieś, jacyś, jakieś, jakiegoś, jakiejś, jakiegoś, jakichś, jakichś, jakiemuś, jakiejś, jakiemuś, jakimś, jakimś, jakiegoś, jakąś, jakieś, jakichś, jakieś, jakimś, jakąś, jakimś, jakimiś, jakimiś, jakimś, jakiejś , jakimś, jakichś, jakichś, jakiś, czyjaś , jakieś, jacyś, jakieś, skądś, dokądś, kiedyś , jakoś, gdzieś, niektóry, niektóre, niektóra*
Negative pronouns	*donikąd, nic, niczego, niczemu, niczyj, nigdy, nigdzie, nijak, nikim, nikogo, nikomu, nikt, znikąd, żaden, żadna, żadną, żadne, żadnego, żadnej, żadnemu, żadni, żadnych, żadnym, żadnymi*
Generalising pronouns	*każda, każdą, każde, każdego, każdej, każdemu, każdy, każdym, wszędzie, wszyscy, wszystkich, wszystkie, wszystkiego, wszystkiemu, wszystkim, wszystko*
Personal pronouns	*ci, ciebie, cię, go, ich, im, ja, ją, je, jego, jej, jemu, mi, mną, mnie, mu , nią, nich, nie , nam, nami , niego, niej, niemu, nim, , nimi my, nas, on, ona, one nas, oni, ono, tobą, tobie, ty, wam, wami, was, wy*
Reflexive pronouns	*się, siebie, sobie, sobą*
Interrogative pronouns	*kto, kogo, komu, kim, co, czego, czemu, czym, czyj, jaką, jakich, jakie, jakiego, jakiej, jakiemu, jakim, jakimi, która, którą, które, którego, której, któremu, który, których, którym, którymi, którzy, ile, gdzie*
Relative pronouns	*jaki, jacy, jaka, jaką, jaki, jakich, jakie, jakiego, jakiej, jakiemu, jakim, jakimi, która, którą, które, którego, której, któremu, który, których, którym, którymi, którzy*
Demonstrative pronouns	*ci, owa, ową, owe, owego, owej, owemu, owi, owo, owych, owym, owymi, ów, stamtąd, stąd, ta, tacy, taka, taką, taką, taki, takich, takie, takiego, takiej, takie, takiego, takiej, takiemu, takim, takimi, tam, tamci, tamta, tamte, tamtego, tamtej, tamtemu, tamten, tamto, tamtych, tamtym, tamtymi, tą, te, tego, tej, temu, ten, tę, tędy, to, tu, tutaj, tych, tym, tymi, wtedy, sam, sama, samo*

Table 6. Grammatical categories and corresponding word lists applied to the Polish version of the LSM analysis.

1.3 Tools

In order to perform the planned analyses, a number of activities were required, including the adaptation of the research material for the purposes of this study and the annotation of texts. For this purpose, publicly available tools such as Praat, POS taggers Spacy and Concraft and Python scripts written by the author of this work were used. The following chapters describe all the tools used and the reason for and how they were used.

1.3.1 Praat

Performing the lexical convergence analysis of the dialogues from the Harmonia corpus required the preparation of the research material in the form of clear texts. The Harmonia corpus was previously transcribed and annotated for phonetic convergence analysis, which was done by the author of this work according to the instructions created by the research team of Prof. Grażyna Demenko and PhD Jolanta Bachan. The annotation was multi-level and contained special characters used to denote phonetic phenomena, speech endings, stressed syllables, etc. For this work, the transcription had to be cleared of these markings. The recordings were annotated in Praat and all files were saved as textgrids. For the purposes of this work, all textgrids were downloaded and an already existing Praat script was used to extract the transcripts. The script is called save_conversation_tiers_as_text_file and is available e.g. on GitHub[27]. In the next step, all files were collected into one txt file, in which all special characters were removed manually.

1.3.2 CLARIN-PL – Spacy

The Spacy tool offered by CLARIN-PL was used to perform automatic lemmatization and annotation with POS tags. It works in the form of a web application[28] and provides an API.

Spacy is a tool that recognizes parts of speech, proper names and performs parsing for texts in various languages. The browser version of the

[27]https://github.com/FieldDB/Praat-Scripts/blob/main/save_conversation_tiers_as_text_file.praat
[28] https://ws.clarin-pl.eu/spacy#

program allows pasting the text and uploading doc, docx, pptx, xlsx, odt, pdf, html, rt files. In the next step, the user is required to select the tool (Tagger, Parser or NER) and the language of the text (Polish, German, Russian, Spanish or English). The analysis result can be displayed on the website or downloaded to a CCL file. When using the API, the required input file format is zip.

Spacy uses machine learning and statistical models to predict which tag is most likely to apply in a given context. The trained component contains binary data that is generated by showing enough examples to the system for it to make predictions that generalise across the language.

Spacy performs automatic segmentation (words) and annotates text with the tags presented in Table 7.

Tag	**Part-of-Speech**
verb	verbs
noun	nouns
adj	adjectives
advs	adverbs
part	particles
pro	pronouns
det	determiner
adp	adpositions
aux	auxiliary verb
cconj	coordinating and correlative conjunction
intj	interjection
num	numbers
propn	proper names
punct	punctuation
sconj	subordinating conjunction
sym	special characters, symbols (e.g. $)

Table 7. List of parts of speech and their corresponding tags used in the Spacy application.

1.3.3 Multiservice NLP - Concraft-pl

Concraft-pl is a tool for tagging morphosintactics in Polish developed by the research team of Institute of Computer Science of the Polish Academy of Sciences (IPIPAN). This tool is based on the Conditional Random Fields and is Coupled with Morfeusz 2. Concraft -PL is available in the form of code[29]

[29] https://github.com/kawu/concraft-pl

and can be used through the browser interface - Multiservice[30]. Concraft-pl performs automatic text annotation and uses the tags used in the National Corpus of Polish (NKJP), which are reserved in the Table 8.

Flexeme	**Abbreviation**	**Base form**	**Example**
noun	*subst*	singular nominative	*profesor*
depreciative form	*depr*	singular nominative form of the corresponding noun	*profesor*
main numeral	*num*	inanimate masculine nominative form	*pięć, dwa*
collective numeral	*numcol*	inanimate masculine nominative form of the main numeral	*pięć, dwa*
adjective	*adj*	singular nominative masculine positive form	*polski*
ad-adjectival adjective	*adja*	singular nominative masculine positive form of the adjective	*polski*
post-prepositional adjective	*adjp*	singular nominative masculine positive form of the adjective	*polski*
predicative adjective	*adjc*	singular nominative masculine positive form of the adjective	*zdrowy, ciekawy*
adverb	*adv*	positive form	*dobrze, b ardzo*
non-3rd person pronoun	*ppron12*	singular nominative	*ja*
3rd-person pronoun	*ppron3*	singular nominative	*on*
pronoun siebie	*siebie*	accusative	*siebie*
non-past form	*fin*	infinitive	*czytać*
future być	*bedzie*	infinitive	*być*
agglutinate być	*aglt*	infinitive	*być*
l-participle	*praet*	infinitive	*czytać*
imperative	*impt*	infinitive	*czytać*
impersonal	*imps*	infinitive	*czytać*
infinitive	*inf*	infinitive	*czytać*
contemporary adv. participle	*pcon*	infinitive	*czytać*
anterior adv. participle	*pant*	infinitive	*czytać*
gerund	*ger*	infinitive	*czytać*
active adj. participle	*pact*	infinitive	*czytać*
passive adj. participle	*ppas*	infinitive	*czytać*
winien	*winien*	singular masculine form	*powinien, rad*
predicative	*pred*	the only form of that flexeme	*warto*

[30] http://multiservice.nlp.ipipan.waw.pl/pl/

Flexeme	**Abbreviation**	**Base form**	**Example**
preposition	*prep*	the non-vocalic form of that flexeme	*na, przez, w*
coordinating conjunction	*conj*	the only form of that flexeme	*oraz*
subordinating conjunction	*comp*	the only form of that flexeme	*że*
particle-adverb	*qub*	the only form of that flexeme	*nie, - że, się*
abbreviation	*brev*	the full dictionary form	*rok, i tak dalej*
bound word	*burk*	the only form of that flexeme	*trochu, oścież*
interjection	*interj*	the only form of that flexeme	*ech, kurde*
punctuation	*interp*	the only form of that flexeme	*;, ., (,]*
alien	*xxx*	the only form of that flexeme	*cool , nihil*
unknown form	*ign*	the only form of that flexeme	

Table 8. Information about base forms for all grammatical classes, as well as the abbreviations of these classes as used in the National Corpus of Polish (Przepiórkowski, 2011).

1.3.4 Python

Python scripts have been prepared to download and use tagging POS applications and to create statistics from the occurring parts of speech. CLARIN-PL offers an API for the Spacy application, which was used in the work. The script starts with downloading the appropriate libraries for navigating the operating system, manipulating folders and files, a library for using regular expressions, a library for creating .zip files and tools for using the API. Further in the script there are commands for tagging and language selection, commands for uploading data and downloading results. In addition, the script includes a function that splits one .txt file with all the text into separate files and creates a .zip file that is required by the Spacy application. The script used for POS tagging with Spacy is available in Appendix V.3.3.1. The second script counts the occurrences of individual parts and speech and prints the results in a table, which is ultimately saved in an .xls file. The script is presented in Appendix V.3.3.2.

Similar tasks are performed by the POS tagging script using a tool offered by IPIPAN, but in this case the application works in browser form, so

another method of automating text tagging was used. The prepared program also uses the selenium technology that enables the automation of repetitive activities. The script allows you to communicate with the application at the URL address, open the files used, tag and save the results. The script is presented in Appendix V.3.3.3.

1.3.5 POS taggers evaluation

1.3.5.1 Evaluation of CLARIN-PL – Spacy

After manual verification of the POS tagging results using the tool offered by the CLARIN-PL consortium, it was concluded that the results require many corrections and are not suitable for creating appropriate statistics of the occurrence of individual parts of speech in texts. In the case of verb recognition, the program took into account many adjectives (e.g. “niebieski” - Eng. “blue”, “szary” - Eng. “gray”), verbs (e.g. “leci” - Eng. “fly”, “powinnaś” - Eng. “you should”, “dzwoń” - Eng. “call”, “kupmy” - Eng. “let's buy”, etc.) adverbs (e.g. “blisko” - Eng. “close”, “pieszo” - Eng. “on foot”, etc.). The list of verbs includes nouns, particles and auxilary verbs. Similarly, in the case of adjectives, the program incorrectly tagged words in a foreign language (e.g. „carrots”) and included numerals, determiners and interjections. The adverbs included nouns and particles. In the case of the pronouns and determiners category, there were also errors in the assignment to the appropriate grammatical category, e.g. particles. In the results, it is difficult to see the rules according to which the program decides on the choice of tag, e.g. the particle “to” (Eng. “it”/ “this”) appears in virtually all grammatical categories, the pronoun “wszystko” / “wszystkie” (Eng. “everything”/ “all”) appears in both pronouns and determiners.

The tags used in the program are adapted to the needs of the English language, which may be a problem when used in Polish. There are several types of pronouns in Polish, and among them there are demonstrative pronouns that are most similar to English determiners. However, many pronouns in Polish can perform different functions, so it is difficult to decide on their correct assignment without checking the context in which they occurred.

Spacy performs lemmatization of all analysed lexical units. The results show that not all words have been reduced to their basic form, e.g. “kupmy” (Eng. “let's buy”) should appear as “kupić”/ “kupować” (Eng. “to buy”), “dzwonię” (Eng. “I’m calling”) should appear as “dzwonić” (Eng. “to call”). In these cases, the app was tagging words incorrectly. At the same time, in other cases, words that appeared in the basic form were misrecognized and lemmatized, e.g. “dzień” (Eng. “day”) was considered an archaic noun “dzienie” (Eng. “wild beehive”). The inflection of the Polish language is very extensive and some words in the inflected form have different meanings, which is an additional difficulty in the automatic morphosyntactic disambiguation of texts.

In general, comparing the results of Manual tagging and using the Spacy tool, a conclusion arises that the program omitted a significant number of units without assigning them to any category. The sum of occurrences of items in particular grammatical categories and the sum of all recognized units differ significantly. Spacy tags each word separately. It does not treat prepositional phrases or other expressions, e.g. greetings “dzień dobry” (Eng. “good morning”) as one item. Manual annotation was done in the same way. The differences in the number of items are therefore difficult to explain.

Table 9 summarises the recognized lexical items in several categories. The parts-of-speech shortlist contains generalised categories used in both manual annotation and Spacy: nouns (no proper names), verbs (no auxiliary verbs), adjectives, adverbs (in manual sum of adverbs, common adverbs and time adverbs), numeral (in manual numerals and indefinite numerals) , conjunctions (in Spacy the sum of coordinating and subordinating conjunctions), pronouns.

Part of Speech	Manual	Spacy	Difference
Nouns	11763	10047	**1716**
Verbs	11074	7837	**3237**
Adjectives	3984	4666	**-682**
Adverbs	4180	3766	**414**
Numeral	878	335	**543**
Conjunctions	7289	2180	**5109**
Pronouns	11397	2557	**8840**

Table 9. Summary of the recognized lexical items in several, generalised categories by Spacy versus manually.

The Tables 10-11 show sample results of the Spacy annotation for one dialog N04_Z05, speaker 1 and 2. Words that are categorised incorrectly are marked in yellow.

The annotation results are questionable and matching some words to certain parts of speech seems to be incorrect. On the Spacy website, there is a rating of the accuracy of individual tools, including POS tagging, and it is a rating of 0,98[31]. It can therefore be concluded that the tool is of very high quality but needs some refinement.

[31] https://spacy.io/models/pl

File	Nouns	Verbs	Adjectives	Adverbs	Part	Pronouns	Det
n04_z05_spk1	16	15	14	11	9	8	10
n04_z05_spk1	{'dywan', 'czirp', 'kratka', 'bok', 'żółto', 'pani', 'ptak', 'leci', 'niebieski', 'blisko', 'popcorn', 'obrazek', 'garnek', 'napis', 'materac', 'patelnia'}	{'myśleć', 'leżeć', 'powiedzieć', 'strzelać', 'robić', 'obserwować', 'stać', 'można', 'dziać', 'być', 'warzywa', 'stwierdzić', 'mieć', 'wiedzieć', 'kroić'}	{'carrots', 'czerwony', 'dobry', 'zielony', 'drugi', 'różowy', 'strzelający', 'błękitny', 'mały', 'wiszą', 'pierwszy', 'aha', 'jeden', 'wyraźny'}	{'ciężko', 'podobnie', 'bardzo', 'tutaj', 'centralnie', 'konkretnie', 'kula', 'jak', 'potem', 'tam', 'tak'}	{'no', 'tylko', 'chyba', 'też', 'około', 'jakby', 'nie', 'chociaż', 'że'}	{'to', 'sobie', 'się', 'wszystko', 'nic', 'on', 'ja', 'co'}	{'tym', 'taki', 'ta', 'która', 'takie', 'tego', 'ten', 'tej', 'takim', 'taka'}
n04_z05_spk2	57	25	29	11	9	10	17
n04_z05_spk2	{'pan', 'zlewa', 'blondynek', 'pipipip', 'lewa', 'pani', 'kula', 'spódnica', 'dom', 'głowa', 'okno', 'patelnia', 'poziom', 'dywan', 'materac', 'kurek', 'kuchenka', 'paska', 'ziemia', 'ogół', 'dach', 'naczynie', 'kobieta', 'bluzka', 'trzepak', 'myśl', 'środek', 'strzelba', 'otoczka', 'worek', 'raz', 'niebieski', 'kurczaczek', 'farba', 'bits', 'dół', 'obroża', 'złoto', 'obrazek', 'kawałek', 'góra', 'garnek', 'szafka', 'okej', 'różnica', 'strzała', 'zwierzątko', 'piesek', 'ptak', 'fartuch', 'strona', 'wnęter', 'który', 'mniejszą', 'twarz', 'opis', 'domek'}	{'stać', 'znaleźć', 'opisywać', 'wyglądać', 'strzelać', 'wrócić', 'żebym', 'warzywa', 'znaczyć', 'myć', 'leżeć', 'mówić', 'prowadzić', 'proponować', 'mieć', 'okej', 'kroić', 'to', 'trzepać', 'widzieć', 'być', 'porównywać', 'widać', 'okienko', 'wydawać'}	{'czerwony', 'gazowy', 'pierwszy', 'aha', 'cały', 'sam', 'dobry', 'malutki', 'pomarańczowy', 'duży', 'różny', 'namalowany', 'jeden', 'kolejny', 'różowy', 'niebieski', 'akwarelowy', 'trzeci', 'wodny', 'wystrzelone', 'prawy', 'taki', 'zielony', 'czarny', 'biały', 'lewy', 'pokazane', 'mały', 'fioletowy'}	{'późno', 'koza', 'dokładnie', 'możliwe', 'wysoko', 'kula', 'jak', 'ognie', 'gdzie', 'dobrze', 'tak'}	{'no', 'i', 'chyba', 'też', 'jakby', 'już', 'nie', 'jeszcze', 'osobno'}	{'ty', 'którym', 'tym', 'to', 'się', 'wszystko', 'wy', 'on', 'ja', 'coś'}	{'taki', 'wszystkie', 'moim', 'taką', 'które', 'moje', 'mój', 'który', 'takiego', 'swoim', 'tego', 'takie', 'jakieś', 'tej', 'takim', 'ten', 'taka'}

Table 10. An example of Spacy annotation results with errors marked in yellow.
Categories: nouns, verbs, adjectives, adverbs, part (particles), pronouns, det (determiners).

File	Adp	Aux	Cconj	Intj	Num	Propn	Punct	Sconj	Sym	Others (X)
n04_z05_spk1	8	2	3	0	1	0	0	1	0	0
n04_z05_spk1	{'na', 'obok', 'do', 'u', 'w', 'z', 'pod', 'co'}	{'to', 'jest'}	{'i', 'ale', 'a'}	set()	{'cztery'}	set()	set()	{'że'}	set()	set()
n04_z05_spk2	11	4	4	0	4	1	0	2	0	0
n04_z05_spk2	{'na', 'obok', 'koło', 'od', 'do', 'przy', 'u', 'w', 'nad', 'z', 'pod'}	{'to', 'jest', 'są', 'być'}	{'czy', 'albo', 'i', 'a'}	set()	{'cztery', 'trzy', 'dwa', 'pięć'}	{'daria'}	set()	{'bo', 'że'}	set()	set()

Table 11. An example of Spacy annotation results with errors marked in yellow. Categories: adp (adpositions), aux (auxiliary verb), cconj (coordinating and correlative conjunction), intj (interjection), num (numerals), propn (proper names), punct (punctuation), sconj (subordinating conjunction), sym (special characters, symbols (e.g. $), others.

1.3.5.2 Evaluation of Multiservice NLP - Concraft-pl

Manual analysis of the POS tagging results using the Concraft-pl tool showed errors. The program categorised pronouns as nouns, e.g. “nic” (Eng. “nothing”), “wszystko” (Eng. “everything”), particles or pronouns such as “to”, “co” (Eng. “it”, “what”) and many other lexical items that belong to other POS categories. The program incorrectly annotated personal and demonstrative pronouns by tagging most of them as adjectives. In the case of particles and interjections, the assigned categories vary, e.g. the particle „OK” has been assigned to both nouns, interjections, adjectives and adverbs. Similarly, the particle or adjective “dobra” (Eng. “good”) appeared in the list of nouns and adverbs. In the case of numerals and conjunctions, the vast majority were correctly tagged. The program included indefinite numbers in the correct category, however, ordinal numbers and the word “jeden” (Eng. “one”) were categorised as adjectives.

Similar to the results obtained by Spacy, the most errors are observed in annotation of pronouns and particles. Concraft-pl included them in various grammatical categories, the vast majority were tagged incorrectly. The Tables 13-15 show sample results of tagging lemmatized lexical units by Concraft-pl. Words that were clearly assigned to the incorrect category are marked with yellow.

Concraft uses a different annotation methodology than Spacy, includes more tags, and takes into account different grammatical forms, tense, perfective and imperfective verbs. Based on the results alone (Tables 13-15), it is difficult to assess the correctness of the categorization of individual verbs, because they occur in lemmatized form. A thorough analysis of the original texts and results is needed.

Concraft-pl performed the task of lemmatization and tagging, however, due to numerous errors, the results will not be used for the analysis of lexical convergence in this work. In addition, the annotation methodology used makes the analysis in accordance with the LSM methodology difficult.

Table 12 summarises the recognized lexical items in several categories. All categories falling under one specific part of speech have been

summed up accordingly. The differences in the number of tagged lexical units are significant.

	Manual	**Concraft**	**Difference**
Nouns	11763	16387	**-4624**
Verbs	11074	13164	**-2090**
Adjectives	3984	8629	**-4645**
Adverbs	4180	6428	**-2248**
Numeral	878	493	**385**
Conjunctions	7289	2503	**4786**
Pronouns	11397	2425	**8972**

Table 12. Summary of the recognized lexical items in several, generalised categories by Concraft-pl versus manually.

File	subts	num	adj	adv	ppron12	ppron3	siebie
n04_z05_spk1	24	2	27	24	3	1	1
n04_z05_spk1	['warzywo', 'obrazek', 'napis', 'carrots', 'patelnia', 'garnek', 'popcorn', 'czirp', 'dywan', 'pani', 'nic', 'co', 'pani', 'kratka', 'wszystko', 'co', 'materac', 'bok', 'dywan', 'to', 'dywan', 'ptak', 'kula', 'wszystko']	['cztery', 'cztery']	['błękitny', 'taki', 'mały', 'taki', 'niewyraźny', 'ten', 'strzelający', 'ten', 'ten', 'ten', 'pierwszy', 'ten', 'taki', 'czerwony', 'drugi', 'taki', 'zielony', 'taki', 'różowy', 'niebieski', 'taki', 'ten', 'taki', 'ten', 'jeden', 'który', 'jeden']	['podobnie', 'ciężko', 'tak', 'tak', 'tam', 'zielono', 'żółto', 'potem', 'tutaj', 'tak', 'tak', 'tak', 'tak', 'bardzo', 'tak', 'jak', 'tak', 'konkretnie', 'centralnie', 'tak', 'blisko', 'potem', 'tak', 'tak']	['ja', 'ja', 'ja']	['on']	['siebie']
n04_z05_spk2	86	12	75	19	10	5	0
n04_z05_spk2	['Daria', 'obrazek', 'opis', 'swoje', 'strona', 'dół', 'obrazek', 'kawałek', 'dom', 'ogół', 'obrazek', 'farba', 'dół', 'to', 'domek', 'kawałek', 'dach', 'otoczka', 'środek', 'wnętrze', 'dom', 'kobieta', 'blondynka', 'bluzka', 'spódnica', 'fartuch', 'zlew', 'szafka', 'szafka', 'naczynie', 'poziom', 'twarz', 'głowa', 'okno', 'warzywo', 'szafka', 'worek', 'bits', 'różnica', 'strona', 'kuchenka', 'dół', 'okienko', 'góra', 'kurek', 'garnek', 'patelnia', 'coś', 'to', 'patelnia', 'domek', 'zwierzątko', 'piesek', 'obroża', 'kurczaczek', 'pipipip', 'różnica', 'kurek', 'strona', 'koza', 'trzepak', 'pani', 'dywan', 'pasek', 'to', 'dywan', 'myśl', 'to', 'to', 'pani', 'dywan', 'pasek', 'różnica', 'dywan', 'ziemia', 'dywan', 'materac', 'coś', 'to', 'strona', 'dół', 'obrazek', 'kawałek', 'dom', 'to', 'pan']	['dwa', 'pięć', 'trzy', 'trzy', 'cztery', 'cztery', 'trzy', 'trzy', 'cztery', 'trzy', 'cztery', 'trzy']	['mój', 'mój', 'okej', 'prawy', 'mój', 'cały', 'taki', 'wodny', 'akwarelowy', 'biały', 'niebieski', 'ten', 'który', 'czerwony', 'zielony', 'ten', 'różowy', 'pomarańczowy', 'fioletowy', 'który', 'zielony', 'okej', 'ten', 'ten', 'pierwszy', 'okej', 'prawy', 'taki', 'gazowy', 'zielony', 'złoty', 'duży', 'czerwony', 'czarny', 'taki', 'mały', 'taki', 'mały', 'czerwony', 'możliwy', 'okej', 'ten', 'różny', 'biały', 'mały', 'czerwony', 'który', 'kolejny', 'lewy', 'jeden', 'taki', 'zielony', 'jakiś', 'czerwony', 'duży', 'jeden', 'taki', 'malutki', 'wszystek', 'taki', 'sam', 'mój', 'niebieski', 'ten', 'okej', 'dobry', 'jeden', 'trzeci', 'taki', 'różowy', 'zielony', 'taki', 'okej', 'prawy', 'mój']	['dobrze', 'dokładnie', 'dobrze', 'późno', 'dobrze', 'gdzie', 'tak', 'tak', 'późno', 'osobno', 'tak', 'lewo', 'dobrze', 'jak', 'tak', 'tak', 'obok', 'tak', 'wysoko']	['ty', 'ja', 'ja', 'ja', 'ja', 'ja', 'ja', 'ja', 'ja', 'wy']	['on', 'on', 'on', 'on', 'on']	-

Table 13. An example of Concraft-pl assignment results to grammatical categories with errors marked in yellow. Categories: subst (noun), num (main numeral), adj (adjective), adv (adverb), ppron12 (non-3rd person pronoun), ppron3 (n3rd person pronoun), siebie (pronoun siebie).

File	fin	bedzie	aglt	praet	inf	ppas	pred	prep	comp	interj
n04_z05_spk1	21	0	0	1	3	0	4	11	3	1
n04_z05_spk1	['mieć', 'kroić', 'być', 'myśleć', 'być', 'mieć', 'wisieć', 'robić', 'stać', 'wiedzieć', 'być', 'być', 'stać', 'obserwować', 'dziać', 'być', 'leżeć', 'wiedzieć', 'być', 'lecieć']			['strzelać']	['stwierdzić', 'powiedzieć', 'powiedzieć']		['można', 'to', 'to', 'to']	['u', 'z', 'w', 'u', 'obok', 'z', 'na', 'w', 'z', 'pod', 'do']	['chociaż', 'że', 'że']	['dobra']
n04_z05_spk2	55	2	4	3	1	2	3	39	5	0
n04_z05_spk2	['proponować', 'być', 'być', 'być', 'być', 'wrócić', 'być', 'mieć', 'być', 'być', 'stać', 'być', 'myć', 'być', 'kroić', 'leżeć', 'mieć', 'mieć', 'być', 'być', 'mieć', 'być', 'mieć', 'mieć', 'znaczyć', 'wydawać', 'być', 'być', 'mówić', 'widzieć', 'mieć', 'prowadzić', 'mieć', 'być', 'być', 'być', 'stać', 'trzepać', 'mieć', 'mieć', 'być', 'być', 'być', 'wyglądać', 'mieć', 'trzepać', 'mieć', 'mieć', 'mieć', 'leżeć', 'leżeć', 'mieć', 'być', 'być']	['być', 'być']	['być', 'być', 'być', 'być']	['opisywać', 'porównywać', 'mieć']	['znaleźć']	['namalować', 'pokazać']	['to', 'widać', 'widać']	['z', 'z', 'na', 'na', 'w', 'na', 'do', 'w', 'w', 'przy', 'pod', 'na', 'u', 'koło', 'z', 'na', 'do', 'koło', 'na', 'od', 'z', 'nad', 'z', 'nad', 'przy', 'w', 'na', 'u', 'w', 'przy', 'u', 'na', 'na', 'u', 'nad', 'z', 'na', 'na', 'nad']	['że', 'żeby', 'że', 'że', 'bo']	

Table 14. An example of Concraft-pl assignment of results to grammatical categories with errors marked in yellow. Categories: fin (verb in non-past form), bedzie (future być), aglt (agglutinate być) , praet (l-participle), inf (infinitive), ppas (passive adj. participle), pred (predicative), comp (subordinating conjunction), interj (interjection).

File	impt	imps	pcon	pant	ger	pact	winien	qub	brev	burk	interp	xxx	ign
n04_z05_spk1	0	0	0	0	0	0	0	0	0	0	0	0	0
n04_z05_spk1													
n04_z05_spk2	0	0	0	0	0	0	0	0	0	0	0	0	0
n04_z05_spk2													

Table 15. An example of Concraft-pl assignment results to grammatical categories (zero results). Categories: impt (imperative), imps (impersonal), pcon (contemporary adv. participle), pant (anterior adv. participle), ger (gerund), pact (active ad.v participle), winien (winien), qub (particle-adverb), brev (abbreviation), burk (bound word), interp (punctuation), xxx (alien), ign (unknown form).

1.4 Manual POS annotation

Available POS tagging tools for Polish use tags dedicated to English-language materials, as in the case of Spacy, and tags used in NKJP, as in Concraft-pl. For the purposes of this work, it is required to recognize verbs and nouns (content words) as well as auxiliary verbs, prepositions, conjunctions, particles, common adverbs, indefinite numbers, indefinite pronouns, negative pronouns, generalising pronouns, personal pronouns, reflexive pronouns, interrogative pronouns, relative pronouns, demonstrative pronouns (non-content words) required to calculate Language Style Matching factors. Due to these specific requirements and the unsatisfactory quality of the available tools, it was decided to manually annotate the corpus.

The POS annotation was made in a spreadsheet. A frequency list for words was created from the entire corpus and a corresponding tag was added to each lexical unit. For the purposes of the study, a list of 23 tags defining individual lexical units was created. Parts of speech such as adjectives, adverbs, verbs, nouns, numbers, conjunctions, pronouns, particles, prepositions are listed. For some parts of speech, an additional division was introduced due to functions:

- pronouns (possessive, personal, reflexive, interrogative, indefinite, demonstrative, negative, relative, generalising);
- numerals (indefinite and other);
- adverbs (time-related, common adverbs, other);
- verbs (auxiliary verbs and other).

In addition, a separate tag has been assigned for proper names and greetings and farewells. The „other" category includes all kinds of onomatopoeia (e.g. „pipipi" and „chirp" - sounds imitating a bird's chirp) and foreign words (e.g. „beets" and „carrots", which were written in the picture used in Task 5). In Table 16 the complete list of tags used for manual annotation is presented.

Tag	**Part of Speech**	**Additional information/ Examples**
adj	Adjectives	e.g. „niebieski", „ładny"
adv	Adverbs	e.g. „niebezpiecznie", „ładnie"
advtime	Adverbs time	e.g. „wieczorem", „jutro"
aux	Auxilary verb	„być" in all conjugations

Tag	**Part of Speech**	**Additional information/ Examples**
cconj	Conjunctions	e.g. „a", „i", „więc", „tylko"
comadv	Common adverb	Top 100 the most frequently used adverbs based on the frequency list for Polish words created by the G4.19 Language Technology Group of the Wrocław University of Technology (see Chapter III. 1.2.4.1, Table 6).
dempro	Demonstrative pronouns	see Chapter III. 1.2.4.1, Table 6
genpro	Generalizing pronoun	see Chapter III. 1.2.4.1, Table 6
indepro	Indefinite pronoun	see Chapter III. 1.2.4.1, Table 6
indnum	Indefinite number	see Chapter III. 1.2.4.1, Table 6
intj	Interjection	e.g. „ach", „oj"
intpro	Interogative pronoun	see Chapter III. 1.2.4.1, Table 6
negpro	Negative pronouns	see Chapter III. 1.2.4.1, Table 6
noun	Nouns	e.g. „warzywa", „kobieta", „bezpieczeństwo"
num	Numbers	cardinal numbers, ordinal numbers, collective numbers, fractional numbers, e.g. "pięć", "stu", "million"
other	Other	e.g. „chirp", „pipipi", „teges"
part	Particle	see Chapter III. 1.2.4.1, Table 6
perspro	Personal pronoun	see Chapter III. 1.2.4.1, Table 6
posspro	Possessive pronoun	see Chapter III. 1.2.4.1, Table 6
prepos	Prepositions	see Chapter III. 1.2.4.1, Table 6
propname	Proper name	e.g. interlocutors' names, street names etc.
refpro	Reflexive pronoun	e.g. „się", „sobie"
verb	Verbs	e.g. „widzę", „kupię", „warto", „można"
welfar	Welcome/farewell	e.g. „cześć", „hej", „dzień dobry"

Table 16. List of tags used for manual annotation.

The annotation was made on unaltered, unlemmatized material. The tagged words had exactly the same form as they appeared in the recordings. The Polish language is characterised by rich inflection and many ambiguities. One word in different contexts can perform different functions and change the meaning which applies to all parts of speech and inflections. For this reason, the context of the occurrence was regularly checked and an appropriate tag was added based on a thorough analysis. Figure 3 presents the fragment of the frequency list of words created from the Harmonia corpus. The first column contains information with the code of the correct statement. In the next column you can see a list of words in inflected forms in which they appeared in the text. Next is a column with information about the number of occurrences of a given word in the same inflected form. The fourth column contains a tag corresponding to the grammatical category to which the word

has been assigned. An exemplary fragment of the annotated corpus is presented in Figure 3.

File	Word	Occurances	
n04_z05_dpa_spk1	mam	2	verb
n04_z05_dpa_spk1	kroi	1	verb
n04_z05_dpa_spk1	warzywa	1	noun
n04_z05_dpa_spk1	stwierdzić	1	verb
n04_z05_dpa_spk1	jest	3	aux
n04_z05_dpa_spk1	myślę	1	verb
n04_z05_dpa_spk1	robi	1	verb
n04_z05_dpa_spk1	stoi	2	verb
n04_z05_dpa_spk1	wiesz	1	verb
n04_z05_dpa_spk1	można	1	verb
n04_z05_dpa_spk1	powiedzieć	2	verb
n04_z05_dpa_spk1	obserwuje	1	verb
n04_z05_dpa_spk1	dzieje	1	verb
n04_z05_dpa_spk1	leży	1	verb
n04_z05_dpa_spk1	wiem	1	verb
n04_z05_dpa_spk1	strzelał	1	verb
n04_z05_dpa_spk1	obrazek	1	noun
n04_z05_dpa_spk1	napisem	1	noun
n04_z05_dpa_spk1	patelnię	1	noun
n04_z05_dpa_spk1	garnku	1	noun
n04_z05_dpa_spk1	popcorn	1	noun
n04_z05_dpa_spk1	czirp	1	other
n04_z05_dpa_spk1	dywany	1	noun
n04_z05_dpa_spk1	pani	2	noun
n04_z05_dpa_spk1	żółto	1	adj
n04_z05_dpa_spk1	niebieski	1	adj
n04_z05_dpa_spk1	kratkę	1	noun
n04_z05_dpa_spk1	materac	1	noun
n04_z05_dpa_spk1	boku	1	noun

Figure 3. Screenshot of the frequency list created from the Harmonia corpus with POS tags added manually.

The texts were divided into the sum of one speaker's utterances in one recording. Within one such piece, there were repetitions of the same words in the same inflected form, which brought a given lexical unit to one position on the list. The total number of words to be tagged consisted of 44911 items. The total number of lexical items in the corpus (including repetitions) was 79093 items. Table 17 presents a summary of the recognized POS in the entire corpus.

Tag	**Number of tags in the whole corpus**
adj	3984
adv	1888
advtime	333
aux	3068
cconj	7289
comadv	1959
dempro	3917
genpro	302
indepro	952

Tag	Number of tags in the whole corpus
indnum	249
intj	235
intpro	1533
negpro	169
noun	11763
num	629
other	98
part	14731
perspro	2425
propname	2367
posspro	404
prepos	7451
refpro	1695
verb	11074
welfar	578

Table 17. Number of tagged items in the entire corpus.

For the purposes of the analyses, it was necessary to count the occurrences of particular categories in particular recordings, taking into account the speaker. For this purpose, the SUMIFS function was used. In this way, a table was created containing the number of occurrences of individual parts of speech in the speeches of each interlocutor in the dialogues. This material was further used to analyse formality, lexical convergence and LSM factor.

2 Lexical convergence results

All dialogues were divided into statements of individual speakers, which allowed for a thorough analysis of changes in the length of statements and the number of words spoken by the same speakers in different tasks and with different interlocutors (assigned partner vs. Teacher). The results also took into account statistics for entire dialogues in individual tasks.

2.1 Formality level

The analysis of the transcripts of the recordings shows that the interlocutors adjust the level of formality to the partner. In the dialogues conducted by the students, the level of language formality and polite forms used was lower than in the dialogues conducted between the students and the Teacher. In the dialogues conducted by the students, there is a difference in the formality of

the language between the tasks. In those in which students were to play roles, more formal phrases, greetings and farewells are used. In dialogues in which students have to perform a task but do not play a specific role (they act as themselves) the level of formality is the lowest, the language is the most free.

In Task 5, the vast majority of dialogue starts with „ok", „dobra", „no dobra" (Eng. "OK", "alright", "well alright") or the greeting or introduction is omitted altogether. Some of the interlocutors started with „halo" (Eng. "hello"), which in Polish functions as an exclamation used to summon or attract attention, often used in telephone conversations[32]. Only in one dialogue did the interlocutors greet each other with the word „cześć" (Eng. "hello"). In Task 6, in only one dialogue the interlocutors started the conversation with "cześć" (Eng. "hello"), in all others the first words were "dzień dobry" (Eng. "good morning") or "halo dzień dobry" (Eng. "hello good morning"). The same greetings occurred in Task 7, where the talkers were to act out the same conversation only in reversed roles.

In Tasks 8-11, all greetings are "dzień dobry" (Eng. "good morning"), in a few cases the interlocutors added "halo" and "dryn dryn" - the onomatopoeia of a ringing telephone.

In Tasks 12-13, all dialogues were less formal, starting with "cześć" (Eng. "hello") or "hej" (Eng. "hi") or omitting the greeting altogether.

In conversations with the Teacher, Tasks 14-16, the form of greeting was defined by relationships in real life - in some dialogues the greeting is "cześć" (Eng. "hello") and "dzień dobry" (Eng. "good morning") in several cases. In conversations in which participants address each other per "ty" (Eng. "you") without a polite form, "cześć" is used. In conversations in which the interlocutors call each other as "Pan" (Eng. "Mister") or "Pani" (Eng. "Mrs"), the dialogues begin with "dzień dobry". In many cases, the greeting is also omitted, regardless of the polite or casual form used later.

In Task 5, all the interlocutors use casual forms ("ty" – Eng. "you"), in Tasks 6-7, in the case of one pair, there is a free form, in the remaining dialogues, the interlocutors use the polite form, addressing each other as

[32] according to the Polish language dictionary (https://sjp.pl/)

"Pan" (Eng. "Mister") or "Pani" (Eng. "Mrs") In Tasks 8-11, all interlocutors use the polite form. However, in Tasks 12-13 there is only casual form, direct phrases ("ty").

Figure 4 shows a chart that summarises how many times the interlocutors used the "cześć"/ "hej", "dzień dobry" or omitted the greeting in dialogues at all in each task.

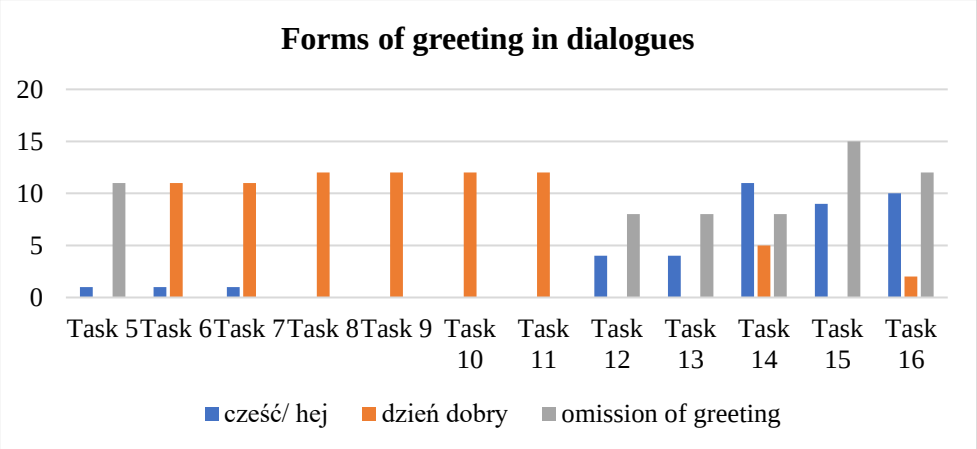


Figure 4. Forms of greeting in dialogues in each task informal "cześć" or "hej", more formal "dzień dobry" or omission.

The results of the analysis of the polite forms used show adaptation to the communicative situation (the scene played) and, consequently, to the role played by the interlocutors in the dialogues, as well as adaptation to the interlocutor in the dialogues whose participants were supposed to speak on their behalf.

2.2 Lexical choices

2.2.1 Lexical choices in Task 5

Task 5 was the diapix task. Each participant had one version of a picture and the purpose was to find the differences by discussing and describing them. In these dialogues, there is a noticeable convergence in the naming of individual elements visible in the pictures. If in a given dialogue one of the partners used one word, usually the conversation partner used the same word. Interlocutors used various terms for birds in the sky – "ptaki", "łabędzie", "kaczki" (Eng. "birds", "swans", "ducks"). Similarly, the character shooting at them was referred to as "pan", "facet", "koleś", "mężczyzna", "strzelec" (Eng.

"mister", "guy", "dude", "man", "shooter"). Terms for shots fired at flying birds include, for example "kula", "pocisk", "nabój" (Eng. "bullet", "round", "gunshot"). The person dressed in yellow was referred to as "kobieta", "pani" (Eng. "woman", "lady"). Hanging materials were referred to as "kołdry", "koce", "pokrowce", "szmaty", "szmatki" (Eng. "duvets", "blankets", "covers", "rags", "cloths"). The animal shown on the left side of the picture is called "koza" (Eng. "goat") (N04), "kozioł" (Eng. "buck") (in N07, N09, N10, Z13, N14, N15) or "baranek" (Eng. "lamb") (N06, N08), or omitted (Z11, Z12). In these cases, only one term was used for this object in a given dialogue, the interlocutors did not introduce new terms.

The pictures also show a dog, which has faeces drawn next to it in one of the versions. Different approaches to naming the faces and the activity that dog performs are used. In N04, N12 this element of the picture was omitted. In N08, the interlocutors established that both versions of the picture show a dog and did not elaborate on its surroundings. In the other recordings, faeces were referred to as "kupa", "odchody", "niespodzianka", "coś" (Eng. "poo", "feces", "surprise", "something"). In several conversations, there are expressions in the dialogues that can be treated as strategies for describing this element of the picture indirectly, e.g. N06_Z05, N15_Z05.

In most dialogues, when one term is used for a given object, the interlocutor uses the same word and does not introduce new terms. In a few cases, it can be observed that the interlocutors somehow negotiate the names of the objects and end up using the same terms. Table 18 presents a few examples of fragments of dialogues in which the duplication of concepts (e.g. N06_05 and N12_05) and finding a consistent use of terms for objects (e.g. N09_05 and N14_05) are noticeable.

Source recording	**Fragment of a dialogue**	**Translation to English**
N06_Z05	SPK1: *no butów to aż tak nie widzę ale ale pomyślmy że tak jest dobra dalej mamy wywieszone te jakieś koce dywany* SPK2: *znaczy po po po po lewej stronie znaczy wie pan jak to u mnie wygląda są jest linka z trzema powiedzmy tymi koc kocodywanami*	SPK1: *well, I don't see shoes that much but let's just think it's good, we still have some blankets carpets hanging* SPK2: *I mean on the left side, you know what it looks like in my picture, there is a cord with, say, three blanketycarpets* (…)

Source recording	Fragment of a dialogue	Translation to English
	(…) SPK2: *tak mam i mam też pieska poniżej tych kurczaków* SPK1: *czy robi on coś specyficznego stoi bo mój robi kupę ewidentnie* SPK2: *no to ten nie robi nic charakterystycznego*	SPK2: *yes I do and I also have a dog below these chickens* SPK1: *is he doing something specific, because mine is pooping evidently* SPK2: *well, this one doesn't do anything distinctive*
N09_05	SPK1: *ile wisi tych szmat* SPK2: *chyba trzy takie duże kołdry i coś co wygląda jak ręcznik albo szmatka jakaś* SPK1: *jakiego koloru są kołdry* SPK2: *zielona taka żółta i różowa* SPK1: *zielona żółta różowa ja mam to znaczy już nie uwzględniając tego czym one są patrząc z lewej mam takie taki zielony nie wiem dywan czy szmatka czy cokolwiek zielono czerwone z frędzlami*	SPK1: *how many of these rags are hanging* SPK2: *I think three big duvets and something that looks like a towel or some cloth* SPK1: *what color are the duvets* SPK2: *green kind of yellow and pink* SPK1: *green yellow pink I have it, not taking into account what they are looking from the left I have this green I don't know carpet or cloth or whatever green red with fringes*
N12_Z05	SPK2: *co widzisz na twoim obrazku jest tam jest tam strzelec jest niebo* SPK1: *jest jest strzelec niebo kaczki jakiś domek pranie* SPK1: *strzela jednym nabojem* SPK2: *a u mnie strzela dwoma nabojami za zarazem czyli ma dubeltówkę taką podwójną*	SPK2: *what you see in your picture is there is a shooter there is the sky* SPK1: *there is is shooter sky ducks some house laundry* SPK1: *fires one gunshots* SPK2: *and mine shoots two gunshots at a time, so it has a double shotgun*
N14_Z05	SPK2: *mężczyzna oddał dwa strzały i widać jak kule lecą w stronę tych ptaków są dwa powiedzmy rozbłyski po wystrzale z tego jego z tej jego strzelby* SPK1: *poczekaj ale widzisz nabój i rozbłysk czy* SPK2: *widzę dwa rozbłyski i dwa naboje w powietrzu* SPK1: *a nie to u mnie jest słuchaj u mnie jest jeden nabój i jeden rozbłysk*	SPK2: *the man fired two shots and you can see the bullets flying towards these birds there are two, let's say, flashes after firing his shotgun* SPK1: *wait but you see a gunshots and a flash or* SPK2: *I see two flashes and two gunshots in the air* SPK1: *and that's not what I have, listen, I have one gunshots and one flare*
N15_Z05	SPK1: *tak a nie wiem czy widzisz pieska* SPK2: *też widzę pieska z czerwoną obrożą* SPK1: *a czy ten piesek coś robi u ciebie* SPK2: *patrzy się w ziemię* SPK1: *aha no bo mój stoi i właśnie coś z siebie wydalił chyba tak tam to wygląda* SPK2: *możliwe że mój również* SPK1: *ale widzisz coś tam obok niego*	SPK1: *yes and I don't know if you can see the dog* SPK2: *I also see a dog with a red collar* SPK1: *and is this dog doing anything at your picture* SPK2: *stares at the ground* SPK1: *aha, because mine is standing and has just expelled something, I think it looks like that* SPK2: *maybe mine too* SPK1: *but can you see something there next to it*

Source recording	Fragment of a dialogue	Translation to English
	SPK2: *możliwe że jest pod nim jakaś kałuża albo cień trudno mi to osądzić i zweryfikować dlatego może poszukajmy trzeciej różnicy w innym miejscu żebyśmy byli* SPK1: *dobrze ale wiesz co to to jest na pewno różnica mi się wydaje że skoro nie widzisz nic obok tego pieska*	SPK2: *it is possible that there is a puddle or shadow under it, it is difficult for me to judge and verify it, so maybe we should look for the third difference in another place so that we are* SPK1: *ok but you know what it is for sure the difference seems to me that since you can't see anything next to this dog*

Table 18. Examples of lexical mimicry tendencies in Task 5.

2.2.2 Lexical choices in Tasks 6-7

In Tasks 6 and 7, the interviewees had maps at their disposal and had to use them to explain the way from the train station to the hotel. In many dialogues, one person explained the route, and the other was limited to single confirmations in the form of "tak", "rozumiem" (Eng. "yes", "I see"). The dialogue N08_07 was the shortest one. One interlocutor explained the route, the other finally confirmed and thanked for the information with the words "ok dziękuję bardzo" (Eng. "ok thank you very much"). In several cases, the person playing the role of a tourist chose to repeat the instructions given by the person playing the hotel receptionist. Several dialogues also consisted of exchanging information, regular checking, reassurance and repeating fragments of instructions. In these cases, the interlocutors repeated fragments of previously heard cues, usually in exactly the same manner. In the dialogues conducted within Task 6 and 7 the names of streets and objects that were written on the map were used.
Table 19 presents examples of duplication of keywords and structures in Tasks 6-7.

Source recording	Fragment of a dialogue	Translation to English
N09_Z07	SPK1: *i skręcić w ulicę Olimpijską* SPK2: *w Olimpijską to będzie trzeci zakręt na prawo* SPK1: *tak trzeci zakręt tak to będzie trzeci zakręt na prawo później proszę później będzie kolejne takie dość specyficzne skrzyżowanie z ulicą tak naprawdę Jaśminową i proszę wtedy skręcić w ulicę Dębową* SPK2: *good*	SPK1: *and turn into Olimpijska Street* SPK2: *into Olimpijska Street it will be the third turn to the right* SPK1*: yes, third turn yes, it will be the third turn to the right later, please, later there will be another quite specific intersection with Jaśminowa Street and then please turn into Dębowa Street* SPK2: *into Dębowa*

Source recording	Fragment of a dialogue	Translation to English
	SPK2: *w Dębową dobrze*	
N04_Z06	SPK2: *dobrze to jesteśmy mniej więcej przy Garbarach tak* SPK1: *tak tak już Alei Niepodległości i Garbary* SPK2: *Garbary i musisz przejść jeszcze dalej prosto aż do ulicy Piastowskiej tak*	SPK2: *okay, we're more or less at Garbary, yes* SPK1: *yes yes, Aleja Niepodległości and Garbary* SPK2: *Garbary and you have to go even further straight up to Piastowska Street, yes*

Table 19. Examples of lexical mimicry tendencies in Tasks 6-7.

2.2.3 Lexical choices in Tasks 8-11

In Tasks 8-11, the interlocutors were asked to exchange information about events in two circumstances. In the content of the task instructions, participants read „Twoim zadaniem jest jednak udzielenie informacji o wydarzeniach i ciekawych miejscach w mieście i przekonanie rozmówcy, by skorzystał z Twoich 3 propozycji", „Wybierasz najbezpieczniejszą propozycję." (Eng. „Your task is to provide information about events and interesting places in the city and convince the interlocutor to use your 3 proposals", „You choose the safest proposal."). The word "propozycja" (Eng. "proposal") was further used by interlocutors in Z08 (N04, N06, N08, N09, N10, N11, N12, N14), Z09 (N06, N08, N09, N15), Z10 (N04, N07, N08, N10, N11, N12, N15), Z11 (N04, N08, N09, N15). Participants in the recordings used phrases such as "chciałbym prosić o propozycję", "mam dla pana propozycje" (Eng. "I would like to ask for a proposal", "I have proposals for you"). During the presentation of the proposal, the interlocutors usually start using the word "opcja" (Eng."option"), e.g. "oraz trzecia opcja może taka niecodzienna ale no pozwolę sobie zaproponować otóż taras mojego domu" (Eng. „and the third option may be so unusual, but let me suggest the terrace of my house"). In none of the dialogues does the word "opcja" appear at the beginning of the dialogue, only in the middle and at the end. In Tasks 8-9, this word appears 6 times. In Tasks 10-11, this word is used 16 times, usually with the adjective "bezpieczny" (Eng. "safe") e.g. N06_Z10.

There are some patterns in the choice of words in the recordings. In one recording (N06_Z08) it can be noticed that the information provider adapts the terms to the caller. A person impersonating an information office

employee uses the terms “posłuchać muzyki”, “pójść do kina” (Eng. “listen to music”, “going to the cinema”), to which the tourist replies using the phrases “wydarzenia muzyczne lub jakieś filmowe” (Eng. “music events or some film events”). The new structure is used by the interlocutor in the next utterance. In most dialogues, the interlocutors use the Polish word “wydarzenie” (Eng. “event”), but in the dialogues N11_Z10 and N12_09, the English equivalent of “event” is used. Another example is the use of the word “kiczowaty” (Eng. “kitschy”) in N10_Z10, where one person describes modern art as such and another person uses the same word. Table 20 shows examples of lexical mimicry tendencies in Task 8-11.

Source recording	**Fragment of a dialogue**	**Translation to English**
N06_Z08	SPK1: *dzisiejszy wieczór już już się tym zajmuję może pan ja w tym momencie otworzyłem sobie foldery właśnie przeglądam może pan przy okazji powiedzieć co pana interesuje czy lubi pan na przykład spokojnie spędzić czas posłuchać muzyki może pójść do kina* SPK2: *myślę że interesowałyby mnie wydarzenia muzyczne lub jakieś filmowe* SPK1: *wydarzenia muzyczne lub filmowe dobrze tutaj znalazłem ofertę Kina Rialto dzisiaj mamy środę i z tego co jest napisane w folderze w środę w Kinie Rialto można udać się na specyficzny rodzaj wydarzenia na tak zwany środa z Klasykiem (...)*	SPK1: *I'm taking care of it, tonight, maybe I've just opened my folders, I'm going through it, can you tell me what you're interested in, for example, do you like to spend time quietly listening to music, maybe go to the cinema* SPK2: *I think I'd be interested in music events or some film events* SPK1: *music or film events I found the Rialto Cinema offer well here today is Wednesday and from what is written in the folder on Wednesday at the Rialto Cinema you can go to a specific type of event on the so-called Wednesday with the Classic (...)*
N06_Z10	SPK2: *może jakieś miejsce nie wiem w którym jest ochrona jakieś wydarzenie zamknięte* SPK1: *wydarzenie zamknięte wydarzenie zamknięte już sprawdzam w folderze co dzisiaj Poznań ma nam do zaoferowania tak i widzę rzeczywiście dzisiaj w Klubie Czekolada odbywa się koncert didżeja (...) i wie pan co* SPK2: *może to będzie najbezpieczniejsza opcja* SPK1: *tak tak jest to najbezpieczniejsza opcja i bardzo gorąco ją w tym wypadku polecam mam nadzieję że za parę dni sytuacja już będzie opanowana*	SPK2: *maybe some place I don't know where there is security some closed event* SPK1: *closed event closed event I’m checking in the folder what Poznań has to offer us today yes and I can see that a DJ concert is taking place today in the Czekolada Club (...) and you know what* SPK2: *maybe this will be the safest option* SPK1: *yes yes it is the safest option and I highly recommend it in this case I hope that in a few days the situation will be under control then call me I will definitely help*

Source recording	Fragment of a dialogue	Translation to English
	wtedy niech pan dzwoni na pewno pomogę	
N07_Z10	SPK2: *jest to jakaś jest to jakieś rozwiązanie czy wystawę czy czy coś ciekawego* SPK1: *wie pan ostatnio w Muzeum Archeologicznym bo zakładam że skoro wystawę to właśnie właściwie błędnie zakładam mówimy o wystawie sztuki czy jakiejś historycznej o muzeum* SPK2: *może wystawa sztuki* SPK1: *wystawa sztuki wie pan myślę że Muzeum Narodowe jest zawsze takim jeżeli interesuje się pan sztuką bo zakładam że skoro pan mówi że że chciałby to zobaczyć to zakładam że trochę tak jest Muzeum Narodowe ma obecnie bardzo ciekawą wystawę impresjonistów z Ameryki Południowej (...)* SPK2: *jakieś inne możliwości mimo wszystko wiem że nie jest pan chętny do udzielania* SPK1: *ja to nie tyle jestem chętny wie pan po prostu boję się że coś inne propozycje no wie pan (...)*	SPK2: *is it some kind of solution or an exhibition or something interesting* SPK1: *you know recently at the Archaeological Museum because I assume that since the exhibition is actually an incorrect assumption, we are talking about an art exhibition or a historical one about the museum* SPK2: *maybe an art exhibition* SPK1: *art exhibition you know I think the National Museum is always like that if you are interested in art because I assume that since you say that you would like to see it I assume that it is a bit like that The National Museum currently has a very interesting exhibition of impressionists from South America (...)* SPK2: *any other possibilities, after all, I know that you are not willing to give* SPK1: *I'm not so much willing, you know, I'm just afraid that there are other proposals, you know (...)*
N08_Z10	SPK2: *dzień dobry mam dzisiaj wolny wieczór i chciałbym wyjść gdzieś troszeczkę się rozerwać na miasto* SPK1: *dzień dobry a czy ma pan świadomość że dzisiaj odbyły się zamachy terrorystyczne* SPK2: *tak mam z mam tego świadomość ale jestem już tak zmęczony po pracy że muszę się rozerwać* SPK1: *no policja i służby zachęcają obywateli do zostania w domach więc proszę się naprawdę zastanowić czy to jest dobry pomysł dzisiejszej akurat nocy wychodzić* SPK2: *no dobrze to co co ma pan w takim razie do zaoferowania mi* SPK1: *naj jeśli mam być szczery z panem najbezpieczniejszą opcją jest po prostu zostać w domu ale jeśli koniecznie pan chce gdzieś wyjść to sugeruję miejsca gdzie nie zbiera się specjalnie dużo ludzi chociażby Muzeum Figur Woskowych (...)*	SPK2: *good morning, I'm free tonight and I'd like to go out and have some fun in the city* SPK1: *good morning, are you aware that there have been terrorist attacks today* SPK2: *yes I'm aware of that but I'm so tired after work that I need to unwind* SPK1: *well, the police and services encourage citizens to stay at home so please really think about whether it's a good idea to go out tonight* SPK2: *well then what do you have to offer me then* SPK1: *if I'm being honest with you, the safest option is to just stay at home but if you absolutely want to go out, I suggest places that don't gather too many people, like the Wax Museum, (...)* SPK2: *something else I see that there are a lot of attractions in this city, but I will decide here I will decide on the Wax Museum here I feel that this is the safest option from the proposal presented here also thank you very much for your help*

Source recording	Fragment of a dialogue	Translation to English
	SPK2: *coś jeszcze no widzę że bardzo tutaj dużo atrakcji w tym mieście jest to jednak się tutaj zdecyduję zdecyduję na na Muzeum Figur Woskowych tutaj czuję że to jest najbezpieczniejsza opcja z propozycja z tutaj z pana przedstawionych także bardzo dziękuję za pomoc* SPK1: *atrakcji jest sporo dziękuję*	SPK1: *there are many attractions, thank you*
N10_Z10	SPK2: *dobrze a czy jeśli chodzi o ten ostatni wernisaż to to chodzi o jakąś taką sztukę nowoczesną czy po prostu zbiera się wszystkie kiczowate* SPK1: *to to jest właśnie zbiór kiczowatych przedmiotów związanych ze sferą sacrum podobno będzie bardzo interesująco jeszcze nie miałam okazji widzieć ale klienci niektórzy byli i bardzo polecali*	SPK2: *ok, and when it comes to the last vernissage, is it about some kind of modern art or is it just all kitschy* SPK1: *this is a collection of kitschy objects related to the sphere of the sacred supposedly it will be very interesting I haven't had a chance to see it yet but some customers have been there and highly recommended it*

Table 20. Examples of lexical mimicry tendencies in Task 8-11.

2.2.4 Lexical choices in Tasks 12-13

In Task 12, participants were asked to express a positive opinion about a picture of a concentration camp made of Lego bricks. These dialogues show different approaches to treating the installation as a work of art, an exhibition or as a publicly available set of bricks. In the dialogues that the interlocutors recognized the set as an artistic expression, the conversation focuses mainly on the description of installation, its elements, colours and opinions and interpretations. In the dialogues in which the participants considered the set to be a generally available toy, the conversations concerned their educational and historical values as well as their use in playing and teaching children.

Analysing the content words (nouns and verbs) of the conversations it is difficult to notice the tendencies to mime expressions and terms. In each subsequent statement, the interlocutors present new observations and new arguments to explain the positive reception of art. Interlocutors, when exchanging observations, often use confirmations, e.g. “faktycznie”, “właśnie”, “dokładnie” (Eng. “actually”, “quite so”, “exactly”). In general, there is a preference for using specific phrases by speakers, but the tendency

is that everyone uses their own kind of confirmation, rarely repeating the one used by the conversation partner. In one recording (N14_Z12), one person tends to repeat the same particles and adverbs and use the same nouns multiple times in one utterance. The second person seems to mime the style of speech using a similar manner of repetition. In all the recordings in this task, the art is most often described as "ciekawa", "interesująca" (Eng. „interesting"). Table 21 shows examples of lexical mimicry in Task 12.

Source recording	**Fragment of a dialogue**	**Translation to English**
N04_Z12	SPK1: *tak porozmawiali i tak no właśnie myślę że pod względem takim historyczno-edukacyjnym to będzie jak najbardziej potrzebne* SPK2: *tak faktycznie cały ten obraz tej makiety jest przerażający prawda ale to chyba to chyba właśnie* SPK1: *ale to chyba zamierzony cel żeby właśnie to żeby zrozumieć te czasy* SPK2: *tak dokładnie a widzisz te ludziki* SPK1: *nie bo to jest no właśnie widzę straszne* SPK2: *no odzwierciedlają* SPK1: *właśnie sytuację tak właśnie* SPK2: *to tych biednych ludzi tak tak dokładnie*	SPK1: *that's how they talked and quite so that's how I think that in terms of historical and educational aspects it will be most necessary* SPK2: *yes actually, the whole picture of this mock-up is really scary, but that's probably it quite so* SPK1: *but I guess that's the intended purpose quite so to understand these times* SPK2: *yes, exactly and you see these men* SPK1: *no, because that's quite so I see is terrible* SPK2: *are not reflected* SPK1: *quite so the situation just like that quite so* SPK2: *It's those poor people so exactly*
N14_Z12	SPK2: *muszę powiedzieć że bardzo bardzo bardzo mi się spodobała ta sztuka bardzo odważne odważne niebanalne no wreszcie wreszcie wreszcie coś nowego no to* SPK1: *tak tak ale to takie takie mocne uderzenie* SPK2: *mocne mocne widać że tu artysta się nie boi poruszać tematów kontrowersyjnych lub takich właściwie które by się komuś mogły wydawać kontrowersyjne bo tutaj prawda żadnych kontrowersji nie ma to jest sztuka nie nie powinna być ograniczana* SPK1: *tak tak ale ale jeszcze tak zwróć uwagę na to że no te szkielety to nie wiem czy to jest to jest genialny pomysł że to już jakby tam scena egzekucji ale ci ludzie są już szkieletami zanim jeszcze ta*	SPK2: *I have to say that I liked this play very very very much brave brave original finally finally something new so* SPK1: *yes yes but it's it's such a strong hit* SPK2: *strong strong you can see that here the artist is not afraid to raise controversial topics or those that might seem controversial to someone because here the truth has no controversies, this is art, it should not be limited* SPK1: *yes yes but but also note that these skeletons, I don't know if they are, it's a brilliant idea that it's like an execution scene, but these people are already skeletons before the execution takes place, they're gone there's no there's no going back they already they already know it's great* SPK2: *that's that that's that that's great I admit that I didn't pay attention*

Source recording	Fragment of a dialogue	Translation to English
	egzekucja się odbędzie to oni już nie ma nie ma odwrotu oni już oni już wiedzą to jest rewelacyjne SPK2: *no to jest to jest to jest rewelacyjne przyznam że nie zwróciłem na to uwagi że tak tak martwi za życia to rzeczywiście fantastycznie to* SPK1: *tak tak ale ta ta kolorystyka taka uboga ale jednocześnie to tak to tak dużo wyraża i ja ja myślę że tu bardzo tak bardzo silne emocje przez tą kolorystykę wyrażone* SPK2: *i tak tak tak bardzo fajnie kontrastują i i i ci strażnicy i te szkielety o których wspomniałeś no zwróć uwagę że zarówno szkielety jak i strażnicy jakby są taką masą to znaczy oni mają nie mają indywidualnych rysów twarzy no to też myślę że ma tutaj znaczenie że to jest taka masa może może może wtłoczona w ten system który się tutaj pojawił no oni jakby spełniają swoją rolę lub lub lub lub są jakby obiektem tej roli którą spełnia ta druga strona* SPK1: *tak tak tak i bardzo dobrze bardzo dobrze że że po prostu że autor nie bał się poruszyć tego tematu bo jednak jednak trzeba o tym rozmawiać to jest to jest rzecz której nie wolno unikać* SPK2: *tak tak no fantastyczne kolory jeszcze raz powiem tu ładnie gra tą barwą no też zwraca uwagę no tak myślę że to jest takie odwołanie też do do zabawy w wojnę jak to się czasem mówi no to* SPK1: *tak dokładnie dokładnie ale jeszcze kontrastowość tych strażników i szkieletów czarni strażnicy i i białe szkielety to no rewelacja po prostu rewelacyjne* SPK2: *rewelacyjne rewelacyjne a zauważ że gdyby sięgnąć pod tą powłokę cielesności albo właściwie nie cielesności tylko pod tą wierzchnią warstwę tych strażników to byśmy tam z pewnością odnaleźli takie same szkielety oni też są martwi tylko tego nie widać na pierwszy rzut oka fantastyczna sztuka fantastyczna kupię dzieciom takie klocki jeśli jakaś firma zdecyduje*	*that it's so so dead when alive it's really fantastic* SPK1: *yes yes but this this color scheme is so poor but at the same time it expresses so much so much and I I think that there are very so very much strong emotions expressed by this color scheme* SPK2: *and yes yes yes, they contrast so well and and and those guards and those skeletons you mentioned, notice that both the skeletons and the guards are like a mass, that is, they don't have individual facial features, so I also think it matters here that there is such a mass maybe maybe maybe squeezed into this system that appeared here well they kind of fulfill their role or or or or are like an object of this role that the other side plays* SPK1: *yes yes yes and very good very good that that simply that the author was not afraid to raise this topic because however however it is necessary to talk about it, this is this is something that must not be avoided* SPK2: *yes yes, fantastic colors, let me say it again, it plays nicely with this color, it also draws attention, yes, I think it is also a reference to playing war, as it is sometimes said, well* SPK1: *yes exactly exactly but also the contrast of these guards and skeletons the black guards and the white skeletons are amazing just amazing* SPK2: *sensational sensational and notice that if we reached under this layer of corporeality or actually not corporeality but only under this top layer of these guards, we would certainly find the same skeletons there they are dead too, only you can't see it at first glance fantastic fantastic art I will buy for children such blocks if a company decides to release them and I will talk about this game of colors and this game of meanings great great* SPK1: *yes yes yes yes true true really sensational approach*

Source recording	Fragment of a dialogue	Translation to English
	się je wypuścić i będę rozmawiał o tej grze barw i tej grze znaczeń kapitalne kapitalne SPK1: *tak tak tak tak prawda prawda rewelacyjne naprawdę podejście*	

Table 21. Examples of lexical mimicry tendencies Task 12.

In Task 13, participants were asked to criticise the installation of Lego bricks. As in the previous task, the participants considered the set with the concentration camp to be modern art or a toy for children, and depending on the approach, the conversations went differently. In the dialogues in which the participants considered the blocks to be an artistic object, the talks focus on criticism of the installation itself, the creator and the Lego company. In conversations in which the set was considered a toy, the content mainly concerned the negative consequences that may result from the use of it by children.

Analysing the dialogue tests from this task, again, it is difficult to find patterns of alignment in the use of similar content words, but there are trends in particles, adverbs and some syntactic similarities. In Z04_Z13, one interlocutor begins the statement with “nie wiem”, (Eng. “I don't know”, which is repeated in subsequent turns. Similarly, there is the phrase “mi się wydaje” (Eng. “it seems to me”) which is later used by the interlocutor. Similar tendency can be observed with other terms such as “po prostu” (Eng. “simply”). In N16_Z13 you can see some cooperation, where one person finishes the sentence of the previous one and the interlocutor repeats the proposed phrase. Table 22 shows examples of lexical mimicry in Task 13.

Source recording	Fragment of a dialogue	Translation to English
N04_Z13	SPK2*: przy ścianie przecież to jest potworne jaki rodzic swojemu dziecku coś takiego kupi* SPK1*: nie wiem właśnie* SPK2*: nie wiem naprawdę nie wiem w jakim celu oni coś takiego stworzyli* SPK1*: nie wiem to jedynie dla ludzi dorosłych chociaż i tak w szkole w liceum*	SPK2: *next to the wall, it's monstrous what parent will buy something like that for their child* SPK1: *I don't know exactly* SPK2: *I don't know, I really don't know why they created something like that* SPK1: *I don't know, it's only for adults, although in high school anyway*

Source recording	Fragment of a dialogue	Translation to English
	SPK2*: no ale dorośli się nie bawią klockami nie* SPK1*: nie wiem czy osobom z liceum można by też to dać* SPK2*: ja myślę że to jest w ogóle bez sensu tworzenie czegoś takiego przecież w liceum ludzie jeżdżą na wycieczki na przykład do Oświęcimia i oglądają te te budynki faktycznie jak one wyglądały a klocki Lego* SPK1*: nie wiem chyba żeby żeby rzeczywiście tak ja myślę że to będzie zniechęcać tych licealistów do historii po pierwsze a po drugie coś mi się wydaje że te klocki nie będą już tak popularne przez to* SPK2*: też mi się tak wydaje masz rację*	SPK2: *well, adults don't play with blocks, no* SPK1: *I don't know if high school people could give it too* SPK2: *I think it makes no sense to create something like that, after all, in high school people go on trips to, for example, Oświęcim and see these buildings, what they really looked like and Lego bricks* SPK1: *I don't know if it's true, I think that it will discourage these high school students from history, firstly, and secondly, it seems to me that these blocks will not be so popular anymore because of this* SPK2: *it seems to me too, you're right*
N16_Z13	SPK2: *znaczy już pomijając walory estetyczne tak no agresja jest po prostu na świecie i i wolałabym żeby moje dziecko po prostu jak naj później się* SPK1: *no właśnie znaczy po prostu żeby się nauczyło pewnie jakoś tam z tym radzić nie no* SPK2: *radzić tak natomiast nie żeby ona nie była agresywna żeby nie nie uważała że to jest rzecz normalna że trzeba być agresywnym bo sobie człowiek w ten sposób radzi w życiu no agresją sobie człowiek nie poradzi w życiu tak* SPK1: *no nie agresja wzbudza agresję także* SPK2: *no dokładnie także uważam że producent miał po prostu* SPK1: *fatalny pomysł* SPK2: *fatalny pomysł ja nie wiem co o czym myślał*	SPK2: *I mean, apart from aesthetic values, yes, aggression is simply in the world and I would prefer my child simply to just recover as soon as possible* SPK1: *well, it just means that it should learn to deal with it somehow, no* SPK2: *deal with yes, but no, so that she is not aggressive, so that she does not think that it is normal that you have to be aggressive because this is how a man copes in life, no man can cope with aggression in life yes* SPK1: *well, not aggression causes aggression too* SPK2: *well exactly I also think that the author had simply* SPK1: *a terrible idea* SPK2: *terrible idea, I don't know what he was thinking*

Table 22. Examples of lexical mimicry tendencies in Task 13.

2.2.5 Lexical choices in Tasks 14-15

In Tasks 14-15, the interlocutors were again asked to express positive and negative opinions about art, but this time in dialogues with the Teacher. In general, the conversations were similar, the interlocutors shared their opinions and exchanged further observations. There is a general tendency to add further arguments and observations, which is associated with new content

words in each subsequent utterance of the interlocutors. At the lexical level, the adaptation is mainly noticeable in the use of affirmations and interjections, e.g. N09_Z14_P2, N06_Z15_P2. Interlocutors also tend to repeat the same adjectives to describe art, e.g. N09_Z14_P1. However, there are a few examples where lexical mimicry is observable in nouns, e.g. N06_Z15_P1.

In Task 14, the most frequently used terms describing piece of art are "świetne" (Eng. "great") – used 57 times, "niesamowite" (Eng. "amazing") – used 27 times, "wspaniałe" (Eng. "wonderful") – used 17 times. The conversation partners described their impressions under the influence of the presented art as "zachwycona"/ "zachwycony" (Eng. "delighted") – used 13 times. In Task 15, the most frequently used terms describing piece of art are "straszne" (Eng. "horrible") – used 21 times, "oburzające" (Eng. "outrageous") – used 9 times, "okropne" (Eng. "terrible") – used 5 times. The conversation partners described their impressions under the influence of the presented art as "oburzona"/ "oburzony" (Eng. "indignant") – used 19 times. Table 23 shows examples of shared words and phrases in dialogues in Tasks 14-15.

Source recording	**Fragment of a dialogue**	**Translation to English**
N09_Z14_P2	SPK1: *tak tak dokładnie w ogóle Libera był tak sprytny że te ogrodzenia które są wokoło też są pod napięciem żeby oddać historię tak jak było naprawdę i nie wiem czy pani jeżeli pani ogląda w tej chwili widzi pani po prawej stronie tam nawet jeden ludzik wisi na na tym ogrodzeniu widzi pani no niech po lewej stronie* SPK2: *tak tak nie zauważyłam może rzeczywiście może to naprawdę będzie uczyć dzieci życia* SPK1: *nie no no dokładnie życia i przeżycia prawda* SPK2: *dokładnie* SPK1: *i śmierdzi no trzy tematy w jednym ale dobrze że pani poruszyła też tą korporację bo no niesamowite naprawdę muszę o tym porozmawiać z bratem który pracuje w*	SPK1: *yes, yes exactly, Libera was so clever that the fences that are around are also electrified to tell the story as it really was and I don't know if if you are watching right now, you can see on the right side there is even one man hanging on on this fence you can see on the left side* SPK2: *yes yes, I didn't notice, maybe indeed, maybe it will really teach children about life* SPK1: *no, well, exactly life and experience, right* SPK2: *exactly* SPK1: *and it smells like three topics in one but it's good that you also raised this corporation because it's amazing I really need to talk about it with my brother who works in a corporation maybe I'll buy him such blocks* SPK2: *I think here he could see a lot of his everyday life* SPK1: *yes yes exactly, only these buildings are now built upwards rather than along right*

Source recording	Fragment of a dialogue	Translation to English
	korporacji może kupię mu takie klocki SPK2: *ja myślę że tutaj mógłby zobaczyć sporo ze swojego codziennego życia* SPK1: *tak tak dokładnie tylko te budynki teraz są raczej budowane wzwyż niż wzdłuż prawda*	
N09_Z14_P1	SPK1: *tak ja ja też byłam pierwszego dnia po na otwarciu no po prostu niesamowite* SPK2: *ja też no ale taki tłum że pewnie dlatego się nie spotkałyśmy ja się zgadzam niesamowite* SPK1: *tak na pewno czyż to nie była wspaniała wystawa tutaj przeniesienie historii w takie realia zabawy prawda no no niesamowite* SPK2: *dla mnie to było bardzo takie celne przedstawienie rzeczywistości obozowej* SPK1: *tak wie pani że to już jest w sklepach dla dzieci* SPK2: *naprawdę i dzieci tak się mogą tym bawić* SPK1: *tak tak mogą budować swoje własne konstrukcje naprawdę no to jest niesamowite* SPK2: *no to prawda to prawda*	SPK1: *yes, I was there the first day after the opening, just amazing* SPK2: me *too, but such a crowd that's probably why we haven't met, I agree, amazing* SPK1: *yes, for sure, wasn't it a great exhibition here, transferring history into such realities of fun, it's true, it's amazing* SPK2: f*or me, it was a very accurate depiction of camp reality* SPK1: y*es, you know that it is already in stores for children* SPK2: *really, and children can play with it like that* SPK1: *yes yes they can build their own structures really well that's amazing* SPK2: *that's true, that's true*
N06_Z15_P1	SPK1: *wiesz co na w takiej małej galerii na Różanej ja mieszkam na Wildzie i tam akurat przechodziłam i tam to wystawiają i to jeszcze wiesz na Różanej jest szkoła obok te dzieciaki patrzą przez patrzą przez te okna no i ja nie wiem nie wybudowali sobie cały taki obóz kilka kilka tych budynków plus plus jakieś tam takie wiesz te budynki z piecami z zagazowane zagazowane nie wiem jak by tu wiesz jak to się nazywa* SPK2: *te komory gazowe* SPK1: *a właśnie tak tak komory gazowe i ogólnie jest pełno takich zbiorowych grobów* SPK2: *o rany zbiorowych grobów nawet* SPK1: *tak i to wszystko jest w oknie właśnie w tej małej galerii na Różanej koło szkoły*	SPK1: *you know what, in a small gallery in Różana I live in Wilda and I was just passing by there and they exhibit it there and you know that there is a school next door in Różana these kids look through these windows and I don't know they didn't build a whole place like that the camp a few a few of these buildings plus plus some stuff you know those buildings with stoves gassed gassed I don't know how you know what it's called* SPK2: *those gas chambers* SPK1: *that's right, gas chambers and mass graves in general* SPK2: *oh my god mass graves even* SPK1: *Yes, and it's all in the window in this small gallery in Różana near the school* SPK2: *near the school, no I at all, I'm generally a little bit surprised that this crap ended up in a small gallery in Różana, once the whole of Polska*

Source recording	Fragment of a dialogue	Translation to English
	SPK2: *koło szkoły nie ja w ogóle jestem generalnie trochę mnie to nie dziwi że to paskudztwo w ogóle wylądowało w małej galerii na Różanej kiedyś trąbiła o tym cała Polska Wyborcza wiadomo nasi*	*Wyborcza trumpeted about it, we know*
N06_Z15_P2	SPK2: *no nie wiem myślałem że to tylko jakaś forma pseudo sztuki zrobiona dla reklamy ale jeśli się okazuje że to wyszło do sklepów trafiło to ktoś naprawdę pomysł na marketing miał świetny* SPK1: *nie no ma być cała seria z obozami z Korei Północnej oraz łagrami z Rosji nie wiem jakie tam masz jakieś* SPK2: *nie wiem na co jeszcze ci ludzie wpadną po prostu obozy świata serio* SPK1: *nie wiem nie wiem w ogóle ja mam zamiar napisać jakieś pismo do Smyka żeby to w wycofali no nie będę tam dobra sztuka sztuką niech to sobie będzie w tych*	SPK2: *I don't know, I thought it was just some form of pseudo art made for advertising but if it turns out that it went to stores then someone really had a great idea for marketing* SPK1: *no, there's supposed to be a whole series with camps from North Korea and gulags from Russia, I don't know what you have there* SPK2: *I don't know what else these people will come up with, just world camps seriously* SPK1: *I don't know, I don't know at all I'm going to write a letter to Smyk to withdraw it well I won't be there good art let it be in these*

Table 23. Examples of lexical mimicry tendencies in Tasks 14-15.

2.2.6 Lexical choices in Task 16

In Task 16, the interlocutors were to disagree on the assessment of the controversial performance. Conversations were held between the students and the Teacher. There is a higher degree of lexical mimicry in these conversations than in previous tasks. The partners present their arguments, which the interlocutors refer to in subsequent statements using the same terms and expressions.

In these dialogues, there are many more patterns of repetition of content words, interlocutors focus on the description of the performance and refer to related issues from their general knowledge. Terms such as "bluźnierstwo"/ "bluźniercy" (Eng. "blasphemy"/ „blasphemers"), "prowokacja" (Eng. "provocation") and many references to religious feelings are used here. Opinions are less personal and more related to the general public, on which interactional partners map their feelings and try to justify them. Interlocutors give examples and refer to related events and facts, and

exchange views and arguments referring to previous statements of the interlocutor. There are rhetorical questions and phrases here, which often contain fragments of the partner's statement. Table 24 shows examples of shared phrases and keywords.

Source recording	**Fragment of a dialogue**	**Translation to English**
N04_Z16_P1	SPK1: *no co też pani mówi przecież to jakie to bluźnierstwo jak ona mogła się ukrzyżować* SPK2: *ale jakie bluźnierstwo a pewno miała konkretny konkretną myśl ideę do przekazania to nie jest nic złego ale to jest głupota to jest prawdziwy przykład zacofania i płytkiego myślenia naprawdę Madonna miała konkretne pomysły do przekazania tym co zrobiła i przecież nie wszyscy po koncercie wyszli i rzucali w jej autobus jajkami tylko jakaś część mała grupa osób a cała reszta rozeszła się* SPK1: *nie ale o dobrze o właśnie tych bluźnierców tych bluźnierców którzy w ogóle nie ja nie wiem co to byli za ludzie (...)*	SPK1: *well, what are you saying, it's blasphemy how she could crucify herself* SPK2: *but what blasphemy she definitely had a specific idea to convey it's not bad but it's stupidity it's a real example of backwardness and shallow thinking Madonna really had specific ideas to convey what she did and after the concert not everyone went out and threw only some small group of people on her bus and all the rest dispersed* SPK1: *no but oh well about those blasphemers those blasphemers who are not at all I don't know what kind of people they were (...)*
N04_Z16_P2	SPK1: *no może ja jestem już starsza no nie wiem ale nie no ale naprawdę nie no nie może pani nie widziała ja byłam tam na tym koncercie i ona tu sobie śpiewała o miłości i tak dalej a następnie weszła i ją tancerze półnadzy przykuli do tego tutaj krzyża takiego oświetlonego super fajnie i ja nie wiem i sobie tam śpiewała nie wiadomo o czym nie wiem jeszcze* SPK2: *no no o miłości śpiewała* SPK1: *no jakiej miłości tutaj gdzie bóg na krzyżu umiera za nas no tak no z miłości żeby nas odkupić a ona sobie tutaj o miłości nie wiem do chyba faceta i to jeszcze a przepraszam nie wiem czy pani wie ale przecież ona chodziła kiedyś z Jesusem który się pisze Jezus i był czterdzieści lat młodszy nie wiem to może po rozstaniu słyszała pani może po się rozstali i dlatego postanowiła się ukrzyżować* SPK2: *Jesus no tak tak no widocznie bardzo cierpiała i to*	SPK1: *well, maybe I'm older, I don't know, but no, but really no, well, maybe you didn't see I was there at that concert and she was singing about love and so on and then she came in and half-naked dancers chained her to this here the cross is so illuminated it's super cool and I don't know and she was singing there I don't know what I don't know yet* SPK2: *well, she sang about love* SPK1: *what kind of love here, where god on the cross dies for us, yes, out of love to redeem us and she here about love I don't know for a guy, I don't know and then oh sorry I don't know if you know but she used to go out with Jesus who it's spelled Jesus and he was forty years younger I don't know maybe you heard after they broke up maybe they broke up and that's why she decided to crucify herself* SPK2: *Jesus, yes, yes, she must have suffered a lot and it's obvious, but one has to be really talented to stand on the stage like this, actually hang*

Source recording	Fragment of a dialogue	Translation to English
	widać ale ale to trzeba być naprawdę uzdolnionym żeby tak stać na scenie właściwie tutaj wisieć na krzyżu mieć ukrzyżowany tutaj ręce związane i żeby tak śpiewać tak żeby tak śpiewać i czysto	*on the cross here, have your hands tied here and sing like that, sing like that and carry a tune*
N07_Z16_P1	SPK2: *oczywiście że pan ale pan bóg czy ja wiem czy naprawdę powinniśmy rozpatrywać to zaraz w kategoriach łamania praw boskich no przecież to jest zwykły performens no Madonna no o to jest część koncertu to jest wizerunek sceniczny* SPK1: *nie jaki performens no w to było ale to było to musiało być zrobione celowo skoro ona zrobiła to w Polsce no ale co ona chciała tu przekazać ja nie wiem co* SPK2: *oczywiście że było zrobione celowo ona to robi to samo w każdym kraju robi to pewnie w Hiszpanii która też jest przecież mocno wierząca* SPK1: *nie ja tu nie słyszałam bo gdyby to było zrobione w Hiszpanii to na pewno byśmy o tym wiedzieli byśmy to ocenzurowali przecież my i partia rządząca nie dopuściłaby takiego performensu tutaj w Polsce*	SPK2: *of course you, but god do I know whether we should really consider it in terms of breaking the God's laws, after all, this is just a performance no Madonna, this is part of the concert, this is a stage image* SPK1: *no what performance, but it was, it must have been done on purpose, since she did it in Poland, but what did she want to convey here, I don't know what* SPK2: *of course it was done on purpose she does it in every country she probably does it in Spain which is also a strong believer* SPK1: *no, I haven't heard of it, because if it was done in Spain, we would have known about it, we would have censored it, after all, we and the ruling party wouldn't allow such a performance here in Poland*

Table 24. Examples of lexical mimicry tendencies in Task 16.

2.2.7 Overall assessment of lexical choices

An overall analysis of the lexical choices scores shows some variation in vocabulary choice between different tasks. The tasks in which the most adjectives occurred were Tasks 5, 8, 9, 12 and 14. Interviewees used the most adverbs in Tasks 7 and 8 and slightly less in Tasks 6, 9 and 10. The most verbs appeared in Tasks 7-11, similarly to nouns. The number of pronouns varied with the decreasing number of nouns in the tasks and reached the highest frequency in Tasks 12, 13, 16. Prepositions appeared proportionally the most often in map tasks, i.e. Tasks 6 and 7. Particles reached the highest frequency in Tasks 5, 12-16. Conjunctions occurred in all recordings in a similar percentage between 8-11%. Figure 5 shows the average percentage of each part of speech in each task.

When analysing the results of the occurrence of individual POS in the tasks, there is a clear difference in the use of nouns in Tasks 6-11 and Tasks 5, 12-16. In tasks with the map and the tourist information office, the interviewees used the most content words (nouns and verbs). In the diapix and provocative, emotion-inducing task, there were significantly fewer nouns and more pronouns. The share of particles in the dialogues is similarly distributed - there are the least of them in Tasks 6-11, much more in the others. The chart in Figure 5 shows the average percentages of occurrence of individual parts of speech in individual tasks.

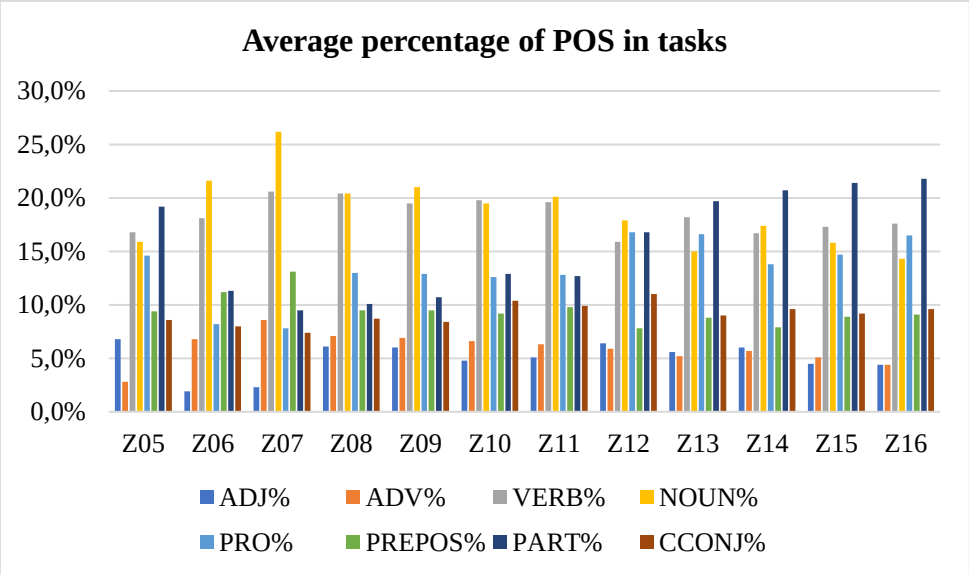


Figure 5. Average percentage of adjectives, adverbs, verbs, nouns, pronouns, prepositions, particles, and conjunctions in each task.

The analysis of the percentage occurrence of pronouns of various functional categories and indefinite numerals showed the greatest differences in the number of demonstrative, personal and interrogative pronouns. The number of demonstrative pronouns is highest in Tasks 12-15 and lowest in Tasks 6-7. The percentage of generalising pronouns was the highest in Tasks 12-13, slightly less in Tasks 10, 11, 15, 16. Indefinite pronouns were the highest in Tasks 10 and 11, and interrogative in Tasks 8,9, 12 and 13. Negative pronouns were used relatively most commonly in Tasks 5, 10, 13, and 16. The number of personal pronouns reached the highest percentage in Task 15 and slightly lower in Task 5, as did possessive pronouns. Reflexive pronouns were used most often in Tasks 8 and 9, and indefinite numerals in

Tasks 7-10. Figure 6 shows the average percentage of occurrence of individual pronouns and indefinite numerals in all tasks.

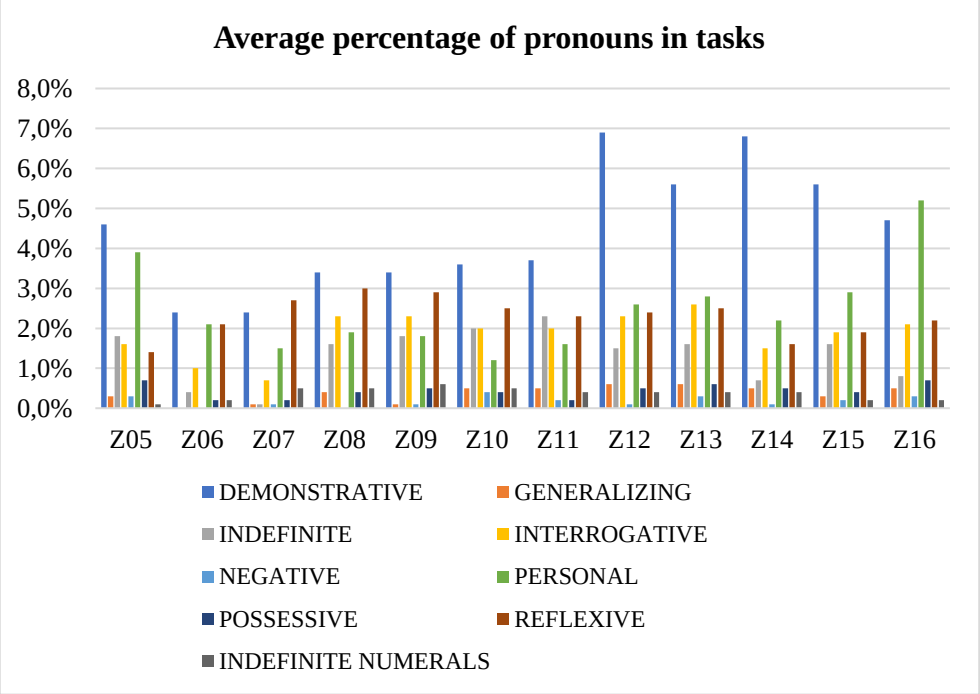


Figure 6. Average percentage of demonstrative, generalising, indefinite, interrogative, negative, personal, possessive, reflexive pronouns.

2.3 Collaborative effort

Collaborative effort is the number of words uttered during the conversations. The results show that the number of words spoken, and thus the length of the utterance, varies for the same speaker depending on the task and the conversation partner. No significant differences in the number of words spoken between women and men were noticed. There is a noticeable increase in conversations that evoke emotions, especially in dialogues where the interlocutors do not agree with their views.

2.3.1 Results by task

The average length of the dialogues (length of both interlocutors' statements) was 382 in Task 5, 252 in Task 6 and 287 in Task 7. In tasks 8-11 these values were higher and ranged between 429 in Task 8 and 345 in Task 11. Similar results are observable in tasks 12 and 13, in which only students participated. Tasks 14-16, in which the Teacher also participated, there is a visible increase

in spoken words, especially in Task 16. The chart in Figure 7 shows the results of measuring the average number of words spoken in dialogues in individual tasks.

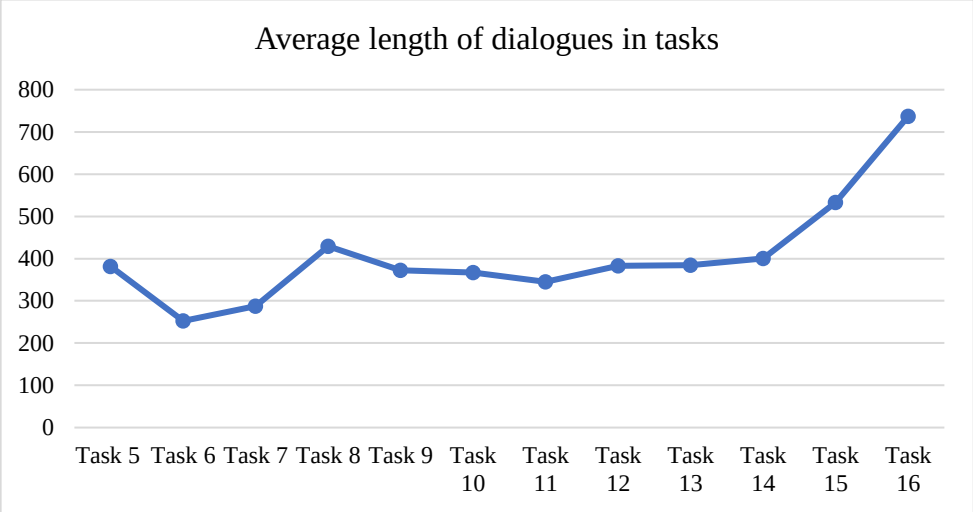
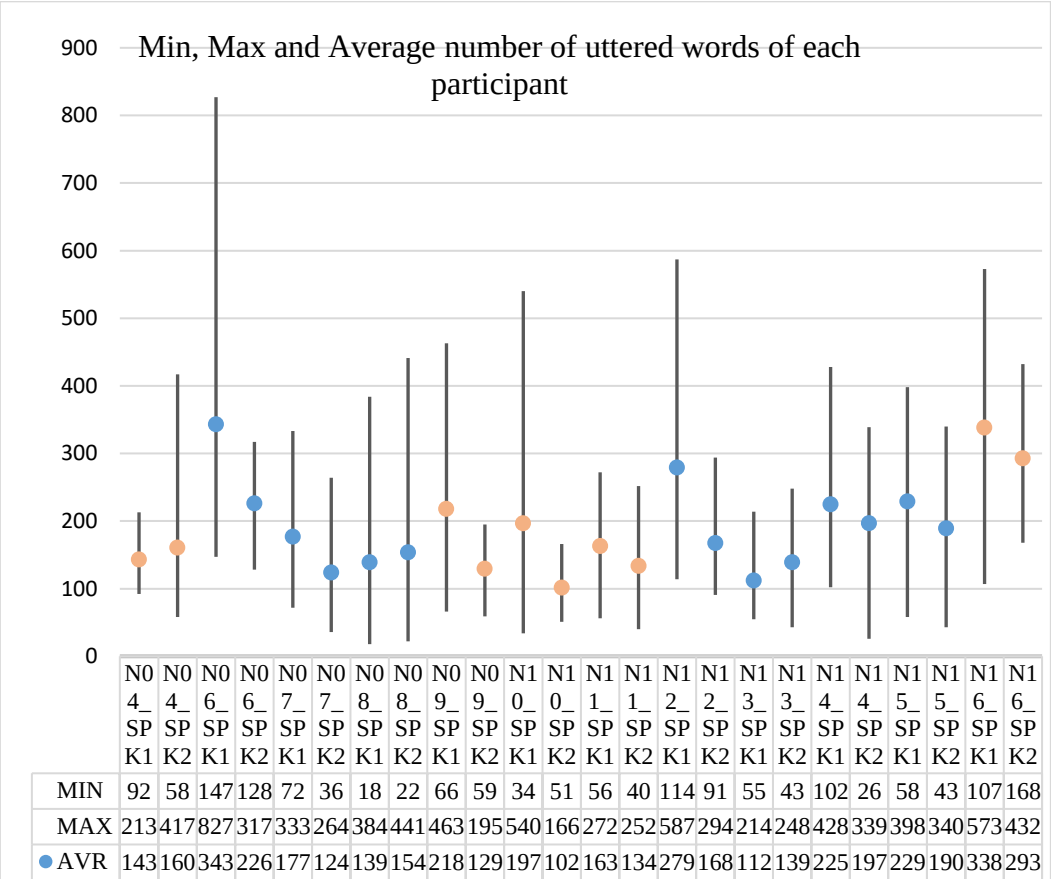


Figure 7. Average length of dialogues (number of uttered words).

2.3.2 Results by speaker

The average length of one person's utterance in the Harmonia corpus dialogues was 235 words. The longest utterance occurred in dialogue N06, in Task 8, and included 827 spoken words. The lowest value was observed in the statement N14_SPK2 in Task 7. Generally, the lowest values are observed in Tasks 6 and 7, where one person explains the way and the other person is often limited to single confirmations and thanks. Clearly, the most words were uttered by the speakers in Task 16. The results also show trends in the length of individual dialogues – the shortest dialogues were created by the interlocutors in N13 and the longest in N16. The person who uttered the most words on average was speaker 1 in N16 and the least SPK1 in N13. No significant differences were observed between the average lengths of speeches by women (average 188 words) and men (average 193 words). When analysing the results of the collaborative effort individually, one can notice the tendencies of individual people to utter words. Measurements of average utterance lengths range from 102 to 343. Figure 8 shows the minimum, maximum, and average word counts for each speaker (except the

Teacher) in each recording. Orange dots indicate female interlocutors and blue dots indicate male interlocutors.



	N04_SPK1	N04_SPK2	N06_SPK1	N06_SPK2	N07_SPK1	N07_SPK2	N08_SPK1	N08_SPK2	N09_SPK1	N09_SPK2	N10_SPK1	N10_SPK2	N11_SPK1	N11_SPK2	N12_SPK1	N12_SPK2	N13_SPK1	N13_SPK2	N14_SPK1	N14_SPK2	N15_SPK1	N15_SPK2	N16_SPK1	N16_SPK2
MIN	92	58	147	128	72	36	18	22	66	59	34	51	56	40	114	91	55	43	102	26	58	43	107	168
MAX	213	417	827	317	333	264	384	441	463	195	540	166	272	252	587	294	214	248	428	339	398	340	573	432
AVR	143	160	343	226	177	124	139	154	218	129	197	102	163	134	279	168	112	139	225	197	229	190	338	293

Figure 8. Min, max and average values of speech length (number of words) of each participant (without Teacher).

The Teacher uttered an average of 319 words, the least in Task 14 (average 222) and the most in Task 16 (average 418). The dialogue in which the Teacher said the most words was with N06_SPK1 in Task 16, and the least with N11_SPK2 in Task 14. There is a general trend towards an increase in the number of words spoken in Task 16 in most of the dialogue and the decrease in Task 14. Figure 9 shows a summary of the number of words spoken by the Teacher in the individual dialogues in Tasks 14-16.

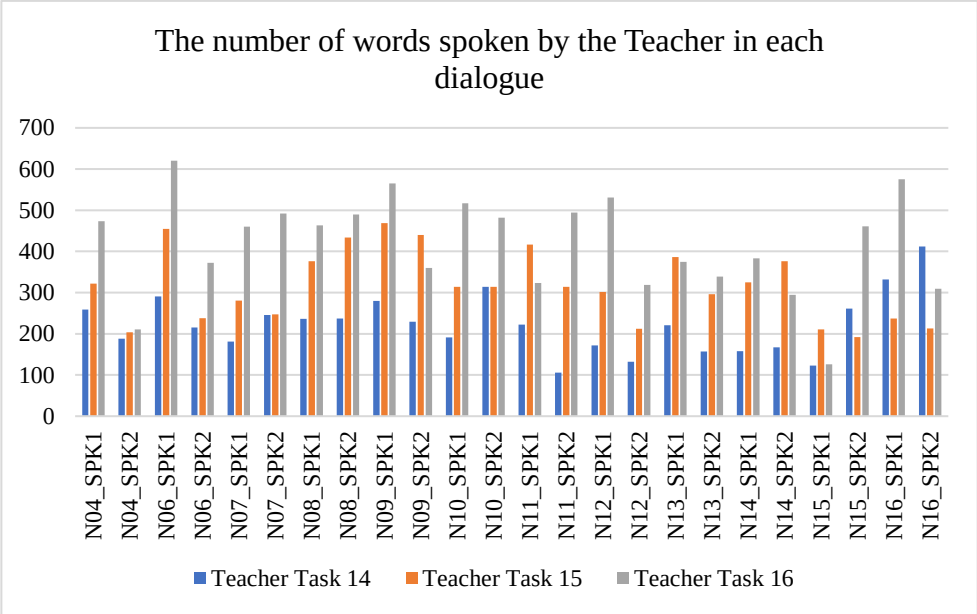


Figure 9. The number of words spoken by the Teacher in each dialogue.

Tasks 12 and 14 consisted of a conversation about a controversial piece of art that the interlocutors were supposed to commend in agreement. In Task 12, the dialogues were shorter, but the average number of words spoken per person was slightly higher than in the dialogues with the Teacher. Interlocutors in Task 12 uttered an average of 184 words, in Task 14 an average of 173 words, while the Teacher averaged 222 words. Figure 10 presents a chart of numbers of words spoken in Tasks 12 and 14 by all the speakers.

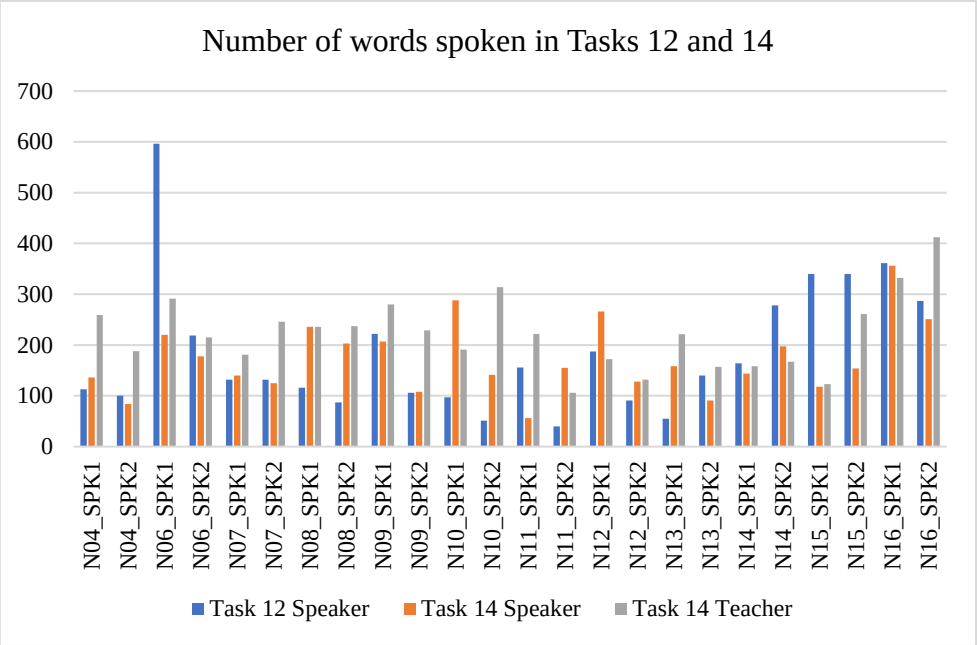


Figure 10. Number of words spoken in Tasks 12 and 14 by all the speakers.

Tasks 13 and 15 consisted in joint, unanimous criticism of controversial art by the interlocutors. In the dialogues of the participants, the average number of words spoken was 192. In the dialogues with the Teacher, the participants uttered an average of 221 words and the Teacher 316. The chart below shows the exact measurement results for Tasks 13 and 15. Figure 11 presents a chart of numbers of words spoken in Tasks 13 and 15 by all the speakers.

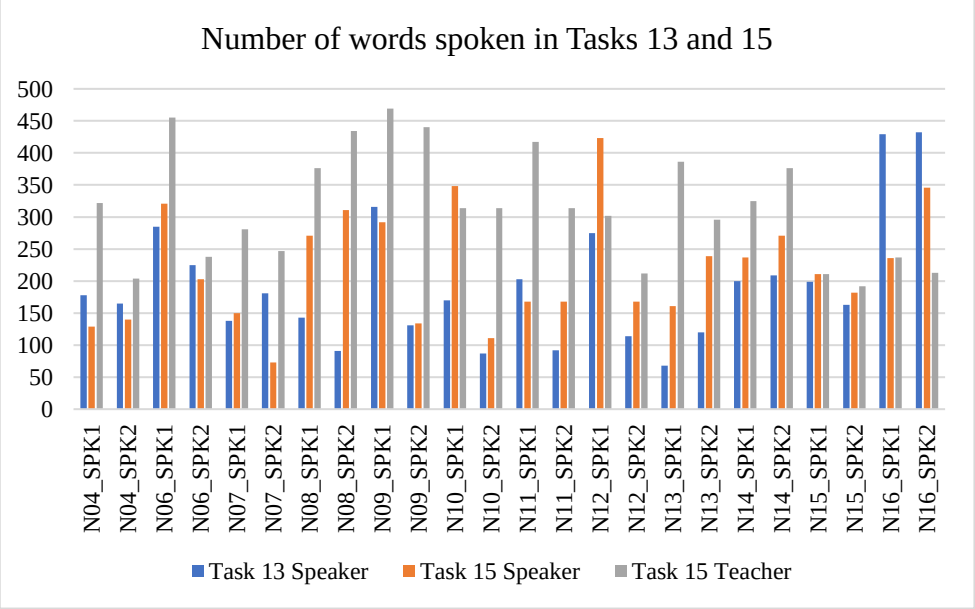


Figure 11. Number of words spoken in Tasks 13 and 15 by all the speakers.

In Task 16, where the interlocutors disagreed about the controversial performance, the dialogues lasted the longest and the average number of words uttered by the participants and the Teacher was the highest. The chart in

Figure 12 shows the exact results of these measurements.

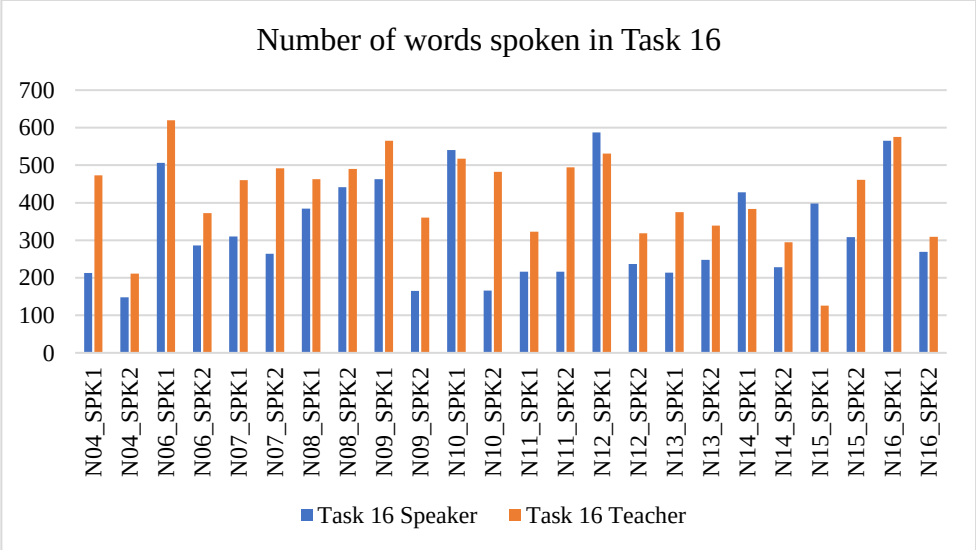


Figure 12. Number of words spoken in Tasks 16 by all the speakers.

2.4 Language Style Matching

In accordance with the LSM methodology, it is necessary to calculate the occurrences of individual types of non-content words and to calculate the percentage of their occurrence in relation to the entire utterance. In the next step, for each grammatical category, the LSM factor should be calculated separately, the result of which is between 0 and 1. The closer the result is to 1, the higher the level of lexical alignment.

All calculations were performed in Excel. The number of occurrences of particular grammatical categories was obtained from the manually annotated corpus. In the first step, LSM values were calculated for individual subcategories, i.e. auxiliary verbs, conjunctions, common adverbs, particles and each kind of pronoun as described in Chapter 4.1. The full results are shown in Appendix V.5. Based on the results of LSM factors for individual categories in each dialogue, overall LSM factors for all dialogues were calculated. Table 25 shows the calculation results for all dialogs. Coloured cells in the table indicate higher values (close to red) and lower

values (close to green). The colours of the cells with the recording number indicate the sex of the interlocutors: orange - women, blue - men.

	Cooperation, common goal			Expression, persuasiveness				Provoc. arousing emotions		Provocative, arousing emotions (students with Teacher)					
	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12	Task 13	Task 14	Task 15	Task 16	Task 14	Task 15	Task 16
N04	0,57	0,54	0,69	0,76	0,51	0,63	0,66	0,67	0,61	0,75	0,87	0,73	0,65	0,63	0,7
N06	0,65	0,72	0,78	0,82	0,7	0,69	0,8	0,72	0,81	0,66	0,7	0,77	0,77	0,78	0,71
N07	0,69	0,75	0,65	0,6	0,56	0,66	0,58	0,62	0,7	0,79	0,69	0,78	0,82	0,5	0,63
N08	0,65	0,68	0,82	0,49	0,57	0,68	0,65	0,65	0,6	0,66	0,71	0,75	0,82	0,67	0,7
N09	0,6	0,48	0,58	0,6	0,63	0,76	0,59	0,77	0,69	0,63	0,73	0,8	0,76	0,54	0,7
N10	0,59	0,63	0,65	0,74	0,63	0,6	0,69	0,62	0,7	0,65	0,71	0,86	0,62	0,73	0,74
N11	0,68	0,46	0,66	0,63	0,56	0,82	0,6	0,54	0,73	0,64	0,68	0,68	0,67	0,65	0,78
N12	0,57	0,6	0,76	0,64	0,77	0,67	0,61	0,71	0,81	0,65	0,88	0,78	0,82	0,7	0,67
N13	0,69	0,71	0,63	0,53	0,7	0,6	0,64	0,65	0,54	0,73	0,65	0,75	0,56	0,66	0,75
N14	0,8	0,63	0,62	0,53	0,61	0,57	0,75	0,7	0,66	0,7	0,78	0,61	0,76	0,73	0,7
N15	0,68	0,74	0,47	0,55	0,65	0,56	0,71	0,79	0,73	0,76	0,6	0,7	0,58	0,68	0,57
N16	0,73	0,59	0,65	0,64	0,64	0,59	0,73	0,7	0,86	0,72	0,67	0,69	0,67	0,7	0,68

Table 25. Summary of LSM factor scores for all dialogues.

The results of the LSM factor in the dialogues of the Harmony corpus range from 0,46 (N11_Z06) to 0,88 (N12_Z15). The lowest values of LSM factor (below 0,5) appeared in two dialogues of female pairs in Task 6 and two dialogues of male pairs in Tasks 7 and 8. The highest values (above 0,85) of LSM factor occurred in dialogues between students and students with the Teacher, where all interlocutors were female. These values appeared in Tasks 13, 14, 16, that is, in conversations on controversial topics. It is worth noting that in these tasks speakers were supposed to agreeably criticise, express a negative opinion (Tasks 13-14) and disagree on the assessment of a controversial performance (Tasks 16).

In Tasks 5-7, which were focused on cooperation, achieving a common goal, the LSM factor scores ranged from 0,46 to 0,82. The highest average values occur in N06, N07, N08, slightly lower in N13, N14, N15. The lowest average LSM factor appeared in N09, N04, N11. In these dialogues, the difference between the level of lexical mimicry between female

and male interlocutors is observable. Table 26 presents detailed results and average LSM factors for each pair in Tasks 5-7.

	Task 5	**Task 6**	**Task 7**	**AVR**
N04	0,57	0,54	0,69	0,600
N06	0,65	0,72	0,78	0,717
N07	0,69	0,75	0,65	0,697
N08	0,65	0,68	0,82	0,717
N09	0,6	0,48	0,58	0,553
N10	0,59	0,63	0,65	0,623
N11	0,68	0,46	0,66	0,600
N12	0,57	0,6	0,76	0,643
N13	0,69	0,71	0,63	0,677
N14	0,8	0,63	0,62	0,683
N15	0,68	0,74	0,47	0,630
N16	0,73	0,59	0,65	0,657

Table 26. Detailed results and average LSM factors for each pair in Tasks 5-7.

In Tasks 8-11, which were designed to evoke expressiveness and persuasiveness, the lowest score was 0,51 and the highest 0,82. The average LSM factor scores for each pair ranged from 0,598 to 0,753. In this group of tasks, there are also noticeable differences in the dialogues of male and female couples. The LSM factor in dialogues between female pairs was slightly higher than in male pairs, but one male pair showed a significantly higher score (N06) than in the other cases. Table 27 shows detailed results and average LSM factors for each pair in Tasks 8-11.

	Task 8	**Task 9**	**Task 10**	**Task 11**	**AVR**
N04	0,76	0,51	0,63	0,66	0,64
N06	0,82	0,7	0,69	0,8	0,753
N07	0,6	0,56	0,66	0,58	0,6
N08	0,49	0,57	0,68	0,65	0,598
N09	0,6	0,63	0,76	0,59	0,645
N10	0,74	0,63	0,6	0,69	0,665
N11	0,63	0,56	0,82	0,6	0,653
N12	0,64	0,77	0,67	0,61	0,673
N13	0,53	0,7	0,6	0,64	0,618
N14	0,53	0,61	0,57	0,75	0,615
N15	0,55	0,65	0,56	0,71	0,618
N16	0,64	0,64	0,59	0,73	0,65

Table 27. Detailed results and average LSM factors for each pair in Tasks 8-11.

In Tasks 12 and 13, the interlocutors were to agree on the assessment of a controversial play, express approval in Task 12 and criticise in Task 13. In these tasks, the lowest LSM factor score is 0,54 and the highest is 0,86. The highest average scores (above 0,75) appeared in N06, N12, N15 and N16. The lowest scores (below 0,65) occurred in N04, N08, N11, N13. Table 28 shows detailed results and average LSM factors for each pair.

	Task 12	**Task 13**	**AVR**
N04	0,67	0,61	0,64
N06	0,72	0,81	0,765
N07	0,62	0,7	0,66
N08	0,65	0,6	0,625
N09	0,77	0,69	0,73
N10	0,62	0,7	0,66
N11	0,54	0,73	0,635
N12	0,71	0,81	0,76
N13	0,65	0,54	0,595
N14	0,7	0,66	0,68
N15	0,79	0,73	0,76
N16	0,7	0,86	0,78

Table 28. Detailed results and average LSM factors for each pair.

Tasks 14-16 were held with the participation of the Teacher, which means that there were female and female-male pairs. In these dialogues, the interlocutors again had to agree and express approval and criticism of the art, and in the last task - to express inconsistent views on performance. In the first group of dialogues, the lowest LSM factor was 0,6 and the highest 0,88. The highest averages for the indicator in these dialogues (above 0,75) appeared in N04, N07, N10, N12. The lowest average occurred in N11 and amounted to 0,667. Table 29 shows detailed results and average LSM factors for each pair in Task 14-16 (pair 1).

	Task 14	**Task 15**	**Task 16**	**AVR**
N04	0,75	0,87	0,73	0,783
N06	0,66	0,7	0,77	0,71
N07	0,79	0,69	0,78	0,753
N08	0,66	0,71	0,75	0,707
N09	0,63	0,73	0,8	0,72
N10	0,65	0,71	0,86	0,74
N11	0,64	0,68	0,68	0,667
N12	0,65	0,88	0,78	0,77

	Task 14	Task 15	Task 16	AVR
N13	0,73	0,65	0,75	0,71
N14	0,7	0,78	0,61	0,697
N15	0,76	0,6	0,7	0,687
N16	0,72	0,67	0,69	0,693

Table 29. Detailed results and average LSM factors for each pair (Pair 1).

In the second part of the dialogues in Tasks 14-16, the lowest LSM factor value was 0,5 and the highest 0,82. The highest average (above 0,75) appeared in one dialogue - N06, and the lowest (below 0,65) in dialogue N15. Table 30 shows detailed results and average LSM factors for each pair in Task 14-16 (pair 2).

	Task 14	Task 15	Task 16	AVR
N04	0,65	0,63	0,7	0,66
N06	0,77	0,78	0,71	0,753
N07	0,82	0,5	0,63	0,65
N08	0,82	0,67	0,7	0,73
N09	0,76	0,54	0,7	0,667
N10	0,62	0,73	0,74	0,697
N11	0,67	0,65	0,78	0,7
N12	0,82	0,7	0,67	0,73
N13	0,56	0,66	0,75	0,657
N14	0,76	0,73	0,7	0,73
N15	0,58	0,68	0,57	0,61
N16	0,67	0,7	0,68	0,683

Table 30. Detailed results and average LSM factors for each pair (Pair 2).

The average LSM factor scores on each task range from 0,611 to 0,717. The lowest mean values (below 0,65) occurred in Tasks 6-10, and the highest (above 0,7) in Tasks 13, 14, 16. On average, the highest level of the LSM factor was achieved in student-student dialogues in which they were supposed to criticise a work of art together (Task 13). The lowest value of the LSM factor was in the task where one speaker impersonated a tourist and the other presented suggestions for events in the city. Also a low, but slightly higher level of LSM factor occurred in Tasks 10-11, which were also based on a conversation between a tourist and an information worker, only in these dialogues there was an added element of danger - a high risk of a terrorist attack in the city. Table 31 presents the average LSM factor results in each task.

Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12	Task 13	Task 14	Task 15	Task 16
0,675	0,623	0,638	0,611	0,648	0,633	0,675	0,674	0,717	0,702	0,683	0,707

Table 31. Average LSM factor results in each Task.

Table 32 shows the averages calculated for the LSM factor scores in Tasks 5-13, where only students were engaged in conversations, and for Tasks 14-16, where students spoke to the Teacher. In this table, conditional formatting using colours was done for each row separately to show how the LSM level changed with different conversation partners. There is a tendency to increase the level of lexical mimicry in dialogues conducted with the Teacher in comparison to the average measurements from the results for students' dialogues. Table 32 shows average LSM factor results for Tasks 5-13 (student dialogues) and for Tasks 14-16 (student-teacher dialogues).

	Tasks 5-13	Tasks 14-16 pair 1	Tasks 14-16 pair 2
N04	0,627	0,783	0,660
N06	0,743	0,710	0,753
N07	0,646	0,753	0,650
N08	0,643	0,707	0,730
N09	0,633	0,720	0,667
N10	0,650	0,740	0,697
N11	0,631	0,667	0,700
N12	0,682	0,770	0,730
N13	0,632	0,710	0,657
N14	0,652	0,697	0,730
N15	0,653	0,687	0,610
N16	0,681	0,693	0,683

Table 32. Average LSM factor results for Tasks 5-13 (student dialogues) and for Tasks 14-16 (student-teacher dialogues).

Summarising the results of LSM factor measurements and calculations, the analysis showed a higher level of lexical mimicry in tasks related to controversy and expressing opinions than in information providing and common goal oriented tasks. Based on the results, it can also be concluded that in student-teacher dialogues the level of LSM factor was on average higher than in student-student dialogues. There were no significant differences in LSM factor scores for male and female interactional pairs.

3 Summary

The **formality level** analysis showed that the participants of the recordings changed their communication behaviour in this regard depending on the task and the interlocutor. The results of the convergence analysis in terms of the adaptation of polite forms show convergence, especially in Tasks 6-11, in which the interlocutors adapted the forms to those appropriate for strangers. In Task 5, they most often omitted the greeting altogether. In Tasks 12-13, the interlocutors more often omitted the greeting, but in several dialogues the free form was used to start the conversation. In the tasks with the Teacher, the interviewees used both casual and more formal forms of greeting. In a few cases, the greeting was also omitted. It can be concluded that the participants here used the forms they use in conversations with the Teacher or adapted them to those used by the Teacher at the beginning of the conversation.

In terms of **collaborative effort**, the shortest dialogues were conducted in Tasks 6-7 and the longest in Tasks 16. In Tasks 5, 8-14, the average lengths of dialogues were similar. In Task 15, the number of words spoken increased. From these data, it can be concluded that the conversations with the Teacher lasted slightly longer than the student-student conversations. However, the topic of the talks is important – in the case of common criticism to piece of art, the talks were longer than in the case of common positive opinion. The longest conversations concerned a provocative performance, about which the interlocutors had different opinions. It can therefore be assumed that conflict talks last longer than consensual talks.

The results of this part of study show that the interlocutors adapt their communication behaviour at the lexical level to the communication situation. Convergence within the **lexical choices** is evident in all tasks at different levels. In Task 5, the interactional partners used the most nouns and adjectives and adapted the descriptions of objects visible in the picture. Once used, the term was usually repeated in a dialogue. It can be concluded that the interlocutors made assumptions about objects and their descriptions, which allowed effective and quick performance of the task. In Tasks 6-7, the interlocutors showed a tendency to duplicate expressions defining the

directions and manner of movement. In terms of content words, proper names that were described on the map occurred in the highest frequency. Interlocutors used them in their instructions and confirmations. In Tasks 8-11, it is clear that the interlocutors created a distance appropriate for strangers and used polite forms and formal greetings and farewells. In terms of **lexical mimicry**, there were not many similarities, but in a few cases the interlocutors adapted the same terms. It can be concluded that information providers imitated information seekers more often. In Tasks 12-13, the number of particles, conjunctions, pronouns used has increased compared to previous tasks. The interlocutors duplicated the forms of confirmation, expressing consent with the interlocutor. Particles and pronouns increased significantly in Tasks 14-16, especially compared to Tasks 6-11. In the dialogues in which the interlocutors agreed with each other in their assessment of the provocative piece of art, a lower level of lexical mimicry was noticed than in the dialogues in which the participants had different opinions. The dialogue in the recordings of Tasks 12-15 was based on drawing conclusions together and adding new observations and ideas that matched the shared opinion. In Task 16, there were more repetitions of content words between interlocutors. In these conversations, the partners often referred to previous statements and tried to present counterarguments to defend their opinion or to convince the interlocutor to their views.

The demanding challenge was to adapt and apply the **Language Style Matching** methodology in Polish, which was done in this work. The results of calculating the LSM factor for individual dialogues are consistent with the results of measuring the occurrence of individual parts of speech and non-content words. The results of the LSM calculations suggest that the highest level of lexical mimicry was achieved in Tasks 13, 14, 16. It can therefore be concluded that students adjusted their communication behaviour more in dialogues with other students in the case of joint opposition to art, and in conversations with the Teacher in the case of joint praise art. The lowest LSM factor values occurred in Tasks 6-10, which can be explained by

the higher frequency of occurrence of proper names, nouns and verbs than non-content words.

In general, the results of the study are satisfactory and give a general picture of communication behaviour in dialogues in Polish in various simulated communication situations. The research material used in this work was appropriate and made it possible to achieve the intended goals and test theses. However, it is assumed that recordings of natural, unstimulated conversations in various communication situations would provide more reliable, real-life results. This type of linguistic material is very difficult to obtain, especially if it was supposed to be dialogues of the same partners in different situations on different topics. The Harmonia corpus made it possible to study communication behaviour in dialogues of 12 pairs in various situations, and additionally each speaker with another person (Teacher). Among the currently available language resources in Polish, no better material for the research conducted in this work has been found.

The methodology and parameters used were appropriate both to achieve the assumed goals. The analysis of the formality of language was a relatively simple measure that yielded unambiguous results. In this study, the level of formality was measured by the type of greeting and farewell and the form in which interlocutors addressed each other. This type of research can be extended to additional, more complex elements such as colloquial versus formal vocabulary used, colloquial interjections, type of affirmations used (e.g. “no” vs. “tak”, “dobra” vs. “dobrze”, etc.). The results of the collaborative effort suggest that in controversial dialogues where the interlocutors exchange views, the conversations last longer than in the case of goal-oriented dialogues. The results are interpretable, but they were probably more useful when combined with additional measurements. For example, it would be useful to check the number and length of each speaker's turns in dialogues. This type of measurement would provide information on whether the engagement is symmetrical, whether someone is dominating, interrupting the interlocutor, etc. The analysis of lexical choices is very important in this type of research, and both the approach based on counting

the used parts of speech and the manual analysis of lexical mimicry in dialogues yielded interesting results. A valuable supplement would be an analysis of the level of matching in terms of language syntax. Language Style Matching is a good source of information on lexical mimicry achieved with simple calculations. The problem here may be the annotation of the texts on the basis of which such an analysis should be carried out. However, these tools can certainly be used for the first annotation, which will require correction in the next step. Also, they can be adapted to non-content words annotation. Taking into account the popularity of the use of the LSM methodology for texts in English and the possibilities it brings in terms of drawing conclusions about psychological and sociological processes, LSM should also be used more widely in research on Polish texts.

IV Final summary and conclusions

The main theses of this dissertation, as presented in the first chapter, are:

- **lexical convergence in dialogues in Polish occurs in lexical choice at the part of speech level,**
- **the intensity of convergence and the level observable in the choice of vocabulary of parts of speech varies depending on the communicative situation,**
- **the interlocutors adjust the level of formality in the language to their own role, the interlocutor and the communication situation,**
- **dialogues on provocative, emotional issues last longer on average than goal-oriented dialogues.**

The study made it possible to confirm the theses. In the part on lexical choices and the calculation of LSM factors, it proved that the level of lexical mimicry changes in specific tasks and with a partner (student-student, student-teacher dialogues). The intensity of convergence was evident in all tasks at different levels in the measures used in this study. The study of the level of formality of language showed clear results that the interlocutors use the forms relevant to the relation with interactional partner, level of familiarity and adjust the forms in tasks where the assumptions on this issue change.

The analysis of lexical convergence in the Harmonia corpus was observed, as demonstrated in this work. Based on the results of the research, it can be concluded that speakers adapt their language to communication situations. Knowledge about the range and parameters as well as circumstances and intensity within which speech and language is adjusted may enable understanding psychological and sociological processes occurring during interaction. The study described in this dissertation gives clear results as to the preference of using part-of-speech vocabulary in various simulated communication situations. This, and similar, broader research on lexical choices can therefore be the basis for modelling lexical choices in the subtask of natural language generation that involves the choice of the content words in a synthesised text and speech. Further research on convergence may

reveal more details about the levels and areas of lexical mimicry in conversations. For Polish, studies in the area of convergence are needed to determine communication alignment in various age groups, also mixed, in dialogues of people in various work, family, official and non-official relations etc. The Harmonia corpus is a set of recordings and transcriptions that have been based on scenarios which means that the dialogues are acted out by the participants. This might have influenced participants' behaviour. In order to verify it and discover trends and patterns of behaviour in natural conversations, similar research should be carried out on real-life material.

Assuming the goal of creating an intelligent dialogue system based on verbal communication in Polish, changes in the linguistic and paralinguistic layer of communication in different situations and between different conversation partners should be processed and simulated. Knowledge of how conversation partners adjust the lexis and prosody of speech will enable modelling of natural dialogues. In addition, research in the field of human-computer communication is of great significance. Nowadays, nearly everyone uses modern technologies, smartphones and computers as well as chatbots and intelligent voice interfaces. To achieve a high level of quality of this type of solution, it has to be understood how a person treats a computer in a communication situation. In other words, what are the user's expectations towards the system in terms of politeness, interjections, frequency and type of phrases in the phatic function etc. Such research should shed light on how language generating systems should adapt the style to the topic, circumstances and the interlocutor so that the person involved wants to have that kind of conversation. This type of adaptation and naturalness of dialogues will increase the quality of the dialogue systems as well as the services based on them.

The subject of convergence is an interdisciplinary research problem, the understanding of which may give the opportunity to develop other scientific areas. One of the potential applications is the possibility of using this knowledge in modern systems based on language technology. Currently, these systems are at a very advanced level for English and it can be assumed

that they will reach a similar level for other languages, including Polish. Effective verbal communication has certain characteristics that will have to be ensured in order for advanced, intelligent systems using natural language for human-computer communication to be accepted and commonly used. Research in the field of communication alignment is necessary, valuable and useful not only for linguistics but also for other fields of science. Further research is especially needed for the Polish language, which is currently scarce. In-depth knowledge of communication alignment at the linguistic and paralinguistic level will contribute to the development of modern speech technologies in Polish.

References

Acoustical Society of America, S. S. (1994). *American National Standard Acoustical Terminology.* Melville, NY: Acoustical Society of America.

Agerri, R., Agirre, E., Aldabe, I., Aranberri, N., Arriola, J. M., Atutxa, A., Iruskieta, M. (2021). *Report on the state of the art in Language Technology and Language-centric AI* . Dublin: European Language Equality Consortium.

Aldwairi, M., & Alwahedi, A. (2018). Detecting Fake News in Social Media Networks. *Procedia Computer Science*, pp. 215-222.

Amit, D., Sturt, P., & Keller, F. (2005). Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 827–834). Association for Computational Linguistics.

Andreeva, B., Grażyna, D., Wolska, M., Möbius, B., Zimmerer, F., Jügler, J., Trouvain, J. (2014). Comparison of Pitch Range and Pitch Variation in Slavic and Germanic Languages. *Proc. 7th International Conference on Speech Prosody* , (pp. 776-780). Dublin.

Aronsson, K., Jönsson, L., & Linell, P. (1987). The Courtroom Hearing as a Middle Ground: Speech Accommodation by Lawyers and Defendants. *Journal of Language and Social Psychology*, pp. 99 – 115.

Bachan, J. (2022). Phonetic Convergence in a Prototype Dialogue System. In M. Ekpenyong, & I. Udoh, *Current Issues in Descriptive Linguistics and Digital Humanities* (pp. 705–719). Singapore: Springer.

Bachan, J., Owsianny, M., & Demenko, G. (2017). Creation of a Dialogue Corpus for Automatic Analysis of Phonetic Convergence. Poznań: Language and Technology Conference.

Bachan, J., Owsianny, M., & Demenko, G. (2020). The Harmonia Corpus – A Dialogue Corpus for Automatic Analysis of Phonetic Convergence. In Z. Vetulani, P. Paroubek, & M. Kubis, *Human Language Technology. Challenges for Computer Science and Linguistics.* Springer.

Bailly, G., & Martin, A. (2014). Assessing objective characterizations of phonetic convergence. *Interspeech 2014 - 15th Annual Conference of the International* (pp. 19-29). Singapor: ISCA.

Bangerter, A., Mayor, E., & Knutsen, D. (2020). Lexical entertainment without conceptual pacts? Revisiting the matching task. *Journal of Memory and Language*, pp. 104-129.

Bańko, M. (2022). *Polszczyzna na co dzień.* Wydawnictwo Naukowe PWN.

Bender, E. M., Sag, I., & Wasow, T. (1999). *Syntactic Theory: A Formal Introduction.* Cambridge University Press.

Bone, D., Li, M., Black, M. P., & Narayanana, S. S. (2014, March). Intoxicated Speech Detection: A Fusion Framework with Speaker-Normalized Hierarchical Functionals and GMM Supervectors. *Computer Speech & Language*, pp. 375-391.

Bowen, J. D., Winczewski, L. A., & Collins, N. L. (2017). Language style matching in romantic partners' conflict and support interactions. *Journal of Language and Social Psychology*, pp. 263–286.

Bradshaw, A. R., & McGettigan, C. (2021). Convergence in voice fundamental frequency during synchronous speech. *PLoS One (eCollection 2021)*.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, pp. B13-B25.

Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., & Wardyński, A. (2012). KPWr: Towards a Free Corpus of Polish. *Proceedings of LREC'12*.

Brown, R. (1973). Development of the first language in the human species. *Am. Psychol. 28*, pp. 97–106.

Brownell, C. A. (1988). Combinatorial Skills: Converging Developments over the Second Year. *Child Development*, pp. 675-685.

Campbell, N., & Scherer, S. (2010). Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with respect to Turn-taking Activity . *Interspeech* (pp. 2546-2549). Makuhari: ISCA.

Carrick, T., Rashid, A., & Taylor, P. J. (2016). Mimicry in Online Conversations: An Exploratory Study of Linguistic Analysis Techniques. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* San Francisco.

Chomsky, N. (1957). *Syntactic Structures* . The Hague: Mouton Publishers.

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

Colby, K. M., Weber, S., & Hilf, F. D. (1971). Artificial Paranoia. *Artificial Intelligence 2* , pp. 1-25.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research 12* , pp. 2493-2537.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics.* Oxford: Blackwell Publishing Ltd.

de Jong, M., Theune, M., & Hofs, D. (2008). Politeness and Alignment in Dialogues with a Virtual Guide. *Applied Physics Letters*, pp. 207-214.

de Looze, C., Oertel, C., Rauzy, S., & Campbell, N. (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*.

Demenko, G. (2015). *Korpusowe badania języka mówionego.* Warszawa: EXIT.

Demenko, G. (2020). Phonetic convergence evaluation. In G. Demenko, *Convergence in Spoken Dialogues in View of Speech Technology Application* (pp. 60-100). Warszawa: EXIT.

Demenko, G. (2020). Prospects for the development of dialogue systems. In G. Demenko, *Phonetic Convergence in Spoken Dialogues in View of Speech Technology Applications* (pp. 137-141). Warszawa: EXIT.

Demenko, G., Grocholewski, S., Klessa, K., Ogórkiewicz, J., Lange, M., Śledziński, D., & N., C. (2008). Jurisdic–Polish Speech Database for taking dictation of legal texts. *Proceedings of the Sixth International Language Resources and*

Evaluation. Marrakech, Morocco: European Language Resources Association (ELRA).

Dingwall, W. O. (1979). Perspectives in Neurolinguistics and Psycholinguistics. In H. Whitaker, & H. A. Whitaker, *Studies in Neurolinguistics - Perspectives in Neurolinguistics and Psycholinguistics* (pp. 1-95). Academic Press.

Donohue, W. A., & Liang, Y. J. (2011). Transformative Linguistic Styles in Divorce Mediation. *Communication Arts & Sciences Building*, pp. 200-218.

Doyle, G., & Frank, M. C. (2016). Investigating the Sources of Linguistic Alignment in Conversation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 526–536.

Elhami, A. (2020). Communication Accommodation Theory: A Brief Review of the Literature. *Journal of Advances in Education and Philosophy*, pp. 192-200.

Gallois, C., Ogay, T., & Giles, H. (2005). Communication Accommodation Theory: A Look Back and a Look Ahead. In B. Gudykunst, *Theorizing about intercultural communication* (pp. 121-148). Sage: Thousand Oaks.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, pp. 181-218.

Gessinger, I., Schweitzer, A., Andreeva, B., Raveh, E., Möbius, B., & Steiner, I. (2018). Convergence of Pitch Accents in a Shadowing Task. *Proc. Speech Prosody 2018*, (pp. 225-229). Poznań.

Gibbon, D. (2004). *First steps in corpus building for linguistics and technology.* Lisbon: SALTMIL.

Gibbon, D., & Borchardt, N. (2007). Computational lexicography: a training programme for language documentation in West Africa. In B. Mbah, & E. Mbah, *Linguistics in History: Essays in honour of P.A. Nwachukwu.* Nsukka, Nigeria: University of Nigeria Press.

Gibbon, D., & Lüngen, H. (2000). Speech Lexica and Consistent Multilingual Vocabularies. In W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation. Artificial Intelligence* (pp. 296–307). Berlin: Springer.

Gibbon, D., Moore, R., & Winski, R. (1998). *Spoken Language System and Corpus Design* . Berlin, New York: Mouton de Gruyter.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, pp. 87-109.

Giles, H. (2008). Communication accommodation theory. In L. A. Baxter, & D. O. Braithewaite, *Engaging theories in interpersonal communication: Multiple perspectives* (pp. 161–173). Sage Publications, Inc.

Giles, H., & Ogay, T. (2007). Communication Accommodation Theory. In B. Whaley, & W. Samter, *Explaining communication: Contemporary theories and exemplars* (pp. 293-310). Mahwah: Lawrence Erlbaum.

Giles, H., & Soliz, J. (2014). Communication accommodation theory: a situated framework for interpersonal, family, and intergroup dynamics. *Engaging Interpersonal Theories*, pp. 159-167.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, N. Coupland, & J. Coupland, *Contexts of accommodation: Developments in applied*

sociolinguistics (pp. 1–68). Cambridge: Cambridge University Press; Editions de la Maison des Sciences de l'Homme.

Gold, E., French, P., & Harrison, P. (2013). Clicking behavior as a possible speaker discriminant in English. *Journal of the International Phonetic Association*, pp. 339-349.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language Style Matching as a Predictor of Social Dynamics in Small Groups. *Communication Research*, pp. 3-19.

Greeley, H. P., Berg, J., Friets, E., Wilson, J., Greenough, G., Picone, J., . . . Nesthus, T. (2007). Fatigue estimation using voice analysis. *Behavior Research Methods*, pp. 610-619.

Gregory, S. W., & Webster, S. (1996). A Nonverbal Signal in Voices of Interview Partners Effectively Predicts Communication Accommodation and Social Status Perceptions. *Journal of Personality and Social Psychology*, pp. 1231-1240.

Grzegorczykowa, R. (1984). *Gramatyka współczesnego języka polskiego. Morfologia.* Warszawa: Wydawnictwo Naukowe PWN.

Heath, J. S. (2017). *Causes and Consequences of Convergence.* Retrieved from UC Berkeley Electronic Theses and Dissertations: https://escholarship.org/uc/item/7ft935nw

Hernandez-Castro, J., & L. Roberts, D. (2015). Automatic detection of potentially illegal online sales of elephant ivory via data mining. *Computer Science*.

Hlavchevaa, Y., Glavcheva, M., Bobicev, V., & Kanishcheva, O. (2020). Language-Independent Features for Authorship Attribution on Ukrainian Texts. *CEUR Workshop Proceedings.* Kyiv.

Howes, C., Healey, P. G., & Purver, M. (2010). Tracking Lexical and Syntactic Alignment in Conversation. *Proceedings of the Annual Meeting of the Cognitive Science Society, 32(32)*, pp. 2004-2009.

Hutchins, W. (1995). MACHINE TRANSLATION: A BRIEF HISTORY. In R. Asher, & E. Koerner, *Concise history of the language sciences: from the Sumerians to the cognitivists* (pp. 431-445). Oxford: Pergamon Press.

Ignas, M., & Ziółko, B. (2013). Baza nagrań mowy emocjonalnej. *Studia Informatica*, pp. 67-77.

Igras, M., Ziółko, B., & Jadczyk, T. (2012). Audiowizualna baza nagrań mowy polskiej. *Studia Informatica* , pp. 163-172.

Ireland, M., Slatcher, R., Eastwick, P., Scissors, L., Finkel, E., & Pennebaker, J. (2011). Language Style Matching Predicts Relationship Initiation and Stability. *Psychological science*, pp. 39-44.

Jakobson, R. (1960). Linguistics and Poetics. In T. S. (Ed.), *Style in Language* (pp. 350-377). Cambridge: Massachusetts Institute of Technology Press.

Jankowska, K., Kuczmarski, T., & Demenko, G. (2021). Phonetic convergence in the shadowing for natural and synthesized speech in Polish. *Lingua Posnaniensis, 62*(2), 7-17. Retrieved from https://sciendo.com/issue/LINPO/62/2

Jassem, W. (2003). Polish consonants. Illustration of the IPA. *Journal of the International Phonetic Association* , pp. 103 - 107.

Jiménez-Hernández, M. (2016). A tutorial to extract the pitch in speech signals using autocorrelation. *Open Journal of Technology & Engineering Disciplines*, pp. 1-10.

Joseph D., M., Arthur N., W., Ruth G., M., & George, S. (1968). Speech and silence behavior in clinical psychotherapy and its laboratory correlates. *Research in psychotherapy*, pp. 347–394.

Karpiński, M., Czoska, A., Jarmołowicz-Nowikow, E., & Juszczyk, K. (2018). Aspects of gestural alignment in task-oriented dialogues. *Cognitive Studies/ Études cognitives*, pp. 1-17.

Karpiński, M., Klessa, K., & Czoska, A. (2014). Local and global convergence in the temporal domain in Polish task-oriented dialogue. *Proceedings of the International Conference on Speech Prosody*.

Kay, M. (1979). Functional Grammar. *Berkeley Linguistics Society: Proceedings of the Annual Meeting.*

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, pp. 107–128.

Kimura, D. (1993). *Oxford psychology series, No. 20. Neuromotor mechanisms in human communication.* Oxford: Oxford University Press.

Kootstra, G. J., Dijkstra, T., & van Hell, J. G. (2020). Interactive Alignment and Lexical Triggering of Code-Switching in Bilingual Dialogue. *Front. Psychol. Sec. Language Sciences*.

Kousidis, S., Dorran, D., Wang, Y., Vaughan, B., Cullen, C., Campbell, D., . . . Coyle, E. (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. *Interspeech* (pp. 1692-1695). Brisbane: ISCA.

Kovacs, B., & Kleinbaum, A. M. (2019). Language-Style Similarity and Social Networks. *Psychological Science*, pp. 1-12.

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior*, pp. 145–162.

Levit, M., Huber, R., Batliner, A., & Noeth, E. (2001). Use of Prosodic Speech Characteristics for Automated Detection of Alcohol Intoxication. *Prosody*.

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Interspeech* (pp. 3081-3084). Florence: ISCA.

Lewandowska-Tomaszczyk, B. (2011). Nowe wyzwania w jakościowej i ilościowej metodologii analizy języka. *Bulletin de La Société Polonaise de Linguistique LXVII*, pp. 141-165.

Lewandowska-Tomaszczyk, B., & A. Wilson, P. (2009). Culture Based Conceptions of Emotion in Polish and English. *PALC 2009: Explorations across Languages and Corpora* (pp. 229-239). Łódź: Peter Lang: Frankfurt a. Main.

Lewandowski, N., & Jilka, M. (2019, May 15). Phonetic Convergence, Language Talent, Personality and Attention. *Frontiers in Communication*.

Li, A.-j., & Yin, Z.-g. (2007). Standardization of Speech Corpus. *Data Science Journal*, pp. 806-812.

Lia, N., & Dash Wu, D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, pp. 354-368.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012, November/December). Behavior Matching in Multimodal Communication Is Synchronized. *Cognitive Science*, pp. 1404-1426.

Lubaszewski, W., Moskal, B., Pisarek, P., & Rokicka, T. (1996). *Komputerowy słownik odmiany wyrazów trudnych.* Kraków.

Łyskawa, P., Maddeaux, R., Melara, E., & Nagy, N. (2016). Heritage Speakers Follow All the Rules: Language Contact and Convergence in Polish Devoicing. *Heritage Language Journal*, pp. 219-244.

Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, pp. 419-426.

Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, pp. 419-426.

Marmot, M. (2005). *Social Determinants of Health: The Solid Facts* . Oxford: Oxford University Press.

Matarazzo, J. D., & Wiens, A. N. (1972). *The interview: Research on its anatomy and structure.* Chicago: Aldine-Atherton.

Mazarweh, S. (2010). *Fillmore Case Grammar: Introduction to the Theory.* GRIN Verlag.

McDonald, D. D., & Pustejovsky, J. D. (1985). A Computational Theory of Prose Style for Natural Language Generation. *EACL '85: Proceedings of the second conference on European chapter of the Association for Computational Linguistics*, (pp. 187–193).

McGarva, A. R., & Warner, R. M. (2003). Attraction and Social Coordination: Mutual Entrainment of Vocal Activity Rhythms. *Journal of Psycholinguistic Research* , pp. 335–354.

McKeown, K. R. (1985). Discourse Strategies for Generating Natural-Langugage Text. *Artificial Intelligence* , pp. 1-41 .

Miller, G. A. (1995, November). WordNet: a lexical database for English. *Communications of the ACM*, pp. 39-41.

Mirzaiyan, A., Parvaresh, V., Hashemian, M., & Saeedi, M. (2010). Convergence and Divergence in Telephone Conversations: A Case of Persian . *International Journal of Social Sciences*, pp. 199-205.

Misiek, T., Favre, B., & Fourtassi, A. (2020). Development of Multi-level Linguistic Alignment in Child-Adult Conversations. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 54–58.

Nagórko, A. (2007). *Zarys gramatyki polskiej*. Warszawa: PWN.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, pp. 790–804.

Ogrodniczuk, M. (2018). Polish Parliamentary Corpus. In D. Fišer, M. Eskevich, & F. de Jong, *Proceedings of the LREC 2018 Workshop ParlaCLARIN:*

Creating and Using Parliamentary Corpora (pp. 15-19). European Language Resources Association.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Acoustical Society of America*, pp. 2382–2393.

Pardo, J. S., Pellegrino, E., Dellwo, V., & Möbius, B. (2022). Special issue: Vocal accommodation in speech communication. *Journal of Phonetics,* pp. 1-9.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, pp. 637–659.

Pardo, J., Gibbons, R., Suppes, A., & Krauss, R. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, pp. 190-197.

Pęzik, P. (2012). Wyszukiwarka PELCRA dla danych NKJP. In A. Przepiórkowski, M. Bańko, R. Górski, & B. Lewandowska-Tomaszczyk, *Narodowy Korpus Języka Polskiego.* PWN.

Pickering, M. J., & Branigan, H. P. (1998). The Representation of Verbs: Evidence from Syntactic Priming in Language Production. *Journal of Memory and Language*, pp. 633-651.

Pickering, M. J., & Garrod, S. (2004, April). The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, pp. 212-225.

Pickering, M. J., & Garrod, S. (2004, April). Toward a mechanistic psychology of dialogue. *Cambridge University Press*, pp. 169-190.

Pitts, M. J., & Giles, H. (2010). Social psychology and personal relationships: Accommodation and relational influence across time and contexts. In D. Matsumoto, *APA handbook of interpersonal communication* (pp. 3-16). Walter de Gruyter & Co.

Placiński, M. (2019). Interactive alignment in Polish: A CMC-based study. *Beyond Philology*, pp. 45-76.

Polański, K. (1999). *Encyklopedia językoznawstwa ogólnego.* Wrocław : OSSOLINEUM.

Polański, K. (1999). *Encyklopedia Językoznawstwa Ogólnego.* Wrocław: Ossolineum.

Przepiórkowski, A. (2011, 10 2). *Ściągawka do Narodowego Korpusu Języka Polskiego* Retrieved from NKJP: http://nkjp.pl/poliqarp/help/ense2.html?fbclid=IwAR0MYMIograsC55sO50B3LZQ0hPDSvSm98UxRQUWzsw5rBL_hFlusuhoHNE

Przepiórkowski, A., Bańko, M., Górski, R. L., Lewandowska-Tomaszczyk, B., Marek, Ł., & Pęzik, P. (2011). National Corpus of Polish. *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, (pp. 259–263). Poznań.

Purpura, S., Schwanda, V., Williams, K., Stubler, W., & Sengers, P. (2011). Fit4life: the design of a persuasive technology promoting healthy behavior and ideal weight. *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 423–432.

PWN, W. N. (n.d.). *Słownik Języka Polskiego.* Retrieved from https://sjp.pwn.pl/

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Journal of the Society for General Systems Research*, pp. 410-430.

Rains, S. A. (2016). Language Style Matching as a Predictor of Perceived Social Support in ComputerMediated Interaction Among Individuals Coping With Illness. *Communication Research*, pp. 694–712.

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, pp. 29-46.

Reitter, D., Keller, F., & Moore, J. D. (2006). Computational modelling of structural priming in dialogue. *Proceedings of the Human Language Technology Conference of the NAACL* (pp. 121-124). New York City: Association for Computational Linguistics.

Richardson, B. H., Taylor, P. J., Snook, B., Conchie, S. M., & Bennell, C. (2014). Language style matching and police interrogation outcomes. *Law and Human Behavior*, pp. 357–366.

Roses, L. K., Brito, J. C., & Filho, G. J. (2015). Conversational Competences Model For Information Technology And Business Strategic Alignment. *Journal of Information Systems and Technology Management*, pp. 125-144 .

Sadeghian, R., Schaffer, D., & Zahorian, S. A. (2021). Towards an Automatic Speech-Based Diagnostic Test for Alzheimer's Disease. *Frontiers in Computer Science*.

Saloni, Z. (1974). *Jak pisać wypracowania.* Wydawnictwa Szkolne i Pedagogiczne.

Schank, R. (1969). A conceptual dependency parser for natural language. *Proceedings of the 1969 conference on Computational linguistics*, (pp. 1-3). Sång-Säby, Sweden.

Schneider, S., Ramirez-Aristizabal, A. G., Gavilan, C., & Kello, C. T. (2020). Complexity matching and lexical matching in monolingual and bilingual conversations. *Bilingualism: Language and Cognition*, pp. 845 - 857.

Schötz, S. (2007). Acoustic Analysis of Adult Speaker Age. In C. Müller, *Speaker Classification I, Lecture Notes in Computer Science* (pp. 88-107). Lund: Springer.

Schweitzer, A., & Lewandowski, N. (2013). Convergence of Articulation Rate in Spontaneous Speech. *INTERSPEECH 2013* (pp. 525-529). Lyon, France: ISCA.

Scollon, R., & Scollon, S. B. (1980, December). Linguistics: Linguistic Convergence: An Ethnography of Speaking at Fort Chipewyan, Alberta. *American Anthropologist*, pp. 874-875.

Sharma, E., & De Choudhury, M. (2018). Mental Health Support and its Relationship to Linguistic Accommodation in Online Communities. *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1-13.

Shaw, H., Taylor, P., Conchie, S., & Ellis, D. (2019, March 06). *Language Style Matching : A Comprehensive List of Articles and Tools.* Retrieved from PsyArXiv: https://psyarxiv.com/yz4br/

Solanki, V. J. (2017, June). Brains in dialogue: investigating accommodation in live conversational speech for both speech and EEG data. *PhD Thesis*. Glasgow.

Street Jr., R. L. (1984). Speech Convergence and Speech Evaluation in Fact-Finding Interviews . *Human Communication Research*, pp. 139–169.

Šturm, P., & Volín, J. (2023). Occurrence and Duration of Pauses in Relation to Speech Tempo and Structural Organization in Two Speech Genres. *Languages*, pp. 23-41.

Surya Gunawan, T., Fahreza Alghifari, M., Arman Morshidi, M., & Kartiwi, M. (2018). A Review on Emotion Recognition Algorithms using Speech Analysis. *Indonesian Journal of Electrical Engineering and Informatics*, pp. 12-20.

Szczepankowska, I. (2012). O semantyce zaimków. *Białostockie Archiwum Językowe*, pp. 275-292.

Szymański, M., & Bachan, J. (2012). Zgodność anotacji segmentalnej i prozodycznej w polskiej bazie Jurisdict. *Speech and Language Technology*.

Tet-Mei Fung, K., Chuah, K.-M., & Ting, S.-H. (2020). Gender differences in computer-mediated communication: a case study on Malaysian millennials. *Humanities & Social Sciences Reviews*, pp. 426-433 .

Tkaczewski, D. (2008). Český národní korpus - internetowe źródło standaryzacji i weryfi kacji języka czeskiego oraz nowoczesne narzędzie dydaktyczne. *Bohemistyka*, pp. 363 - 378.

Toribio, A. J. (2004). Convergence as an optimization strategy in bilingual speech: Evidence from code-switching. *Bilingualism: Language and Cognition*, pp. 165 - 17.

Tschacher, W., Rees, G. M., & Ramseyer, F. (2014, November). Nonverbal synchrony and affect in dyadic interactions. *Front. Psychol.*

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of Native-and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles. *Language and Speech*, pp. 510-540.

Van Eynde, F., & Gibbon, D. (2000). *Processing, Lexicon Development for Speech and Language.* Springer.

Vinciarelli, A., Esposito, A., André, E., Bonin, F., Chetouani, M., Cohn, J. F., . . . Potamianos, A. (2015). Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions. *Cognitive Computation* , pp. 397-413.

Wagner, M. A., Broersma, M., McQueen, J. M., Dhaene, S., & Lemhöfer, K. (2021). Phonetic convergence to non-native speech: Acoustic and perceptual evidence. *Journal of Phonetics*, pp. 1-20.

Wang, Y., Reitter, D., & Yen, J. (2014). Linguistic Adaptation in Conversation Threads: Analyzing Alignment in Online Health Communities. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 55-62). Baltimore: Association for Computational Linguistics.

Weizenbaum, J. (1966). ELIZA--A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 36-35.

Welbers, K., & de Nooy, W. (2014). Stylistic Accommodation on an Internet Forum as Bonding: Do Posters Adapt to the Style of Their Peers? *American Behavioral Scientist*, pp. 1361–1375.

Wierzbicka-Piotrowska, E. (2011). *Polskie zaimki nieokreślone: wybrane zagadnienia semantyczne, syntaktyczne i pragmatyczne.* Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

Wilks, Y., & Fass, D. (1992). The preference semantics family. *Computers & Mathematics with Applications*, pp. 205-221.

Winograd, T. (1971). *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language.* MIT .

Woods, W. (1978). Semantics and Quantification in Natural Language Question Answering. *Advances in Computers*, pp. 1-87.

Xu, Y., & Reitter, D. (2015). An Evaluation and Comparison of Linguistic Alignment Measures. *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 58-67). Denver: Association for Computational Linguistics.

Xu, Y., & Reitter, D. (2016). Convergence of Syntactic Complexity in Conversation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 443–448.

Zając, M. (2015). *Phonetic convergence in the speech of Polish learners.* Łódź: Uniwersytet Łódzki, Wydział Filologiczny.

Zasina, A. (2018). Językoznawstwo korpusowe. Empiryczne podejście w badaniach humanistycznych. *Dziennikarstwo i media 9*, 169-178.

List of Figures

List of Tables

V Appendices

1 Linguistic corpora

The table below gives an overview of existing language corpora in different languages. The list was prepared by the author of this work.

Name	**Description**	**Source**
American National Corpus	15 million words of contemporary American English with automatically-produced annotations for a variety of linguistic phenomena. Texts of all genres and transcripts of spoken data produced from 1990 onward. All data and annotations are fully open and unrestricted for any use.	http://www.anc.org/
Collins Birmingham University International Language Database (COBUILD)	An analytical database of English with over 4.5 billion words. It contains written material from websites, newspapers, magazines and books published around the world, and spoken material from radio, TV and everyday conversations.	https://collins.co.uk/pages/elt-cobuild-reference
British National Corpus	BNC contains 100 million words of text from a wide range of genres (e.g. spoken, fiction, magazines, newspapers and academic). It was originally created by Oxford University press in the 1980s - early 1990s.	https://www.english-corpora.org/bnc/
Bergen Corpus of London Teenage Language (COLT)	The first large English Corpus focusing on the speech of teenagers, collected in 1993 and consists of the spoken language of 13 to 17-year-old teenagers from London. It contains about 500 000 words transcribed orthographically and word-class tagged. COLT is now available through the CLARIN[33] infrastructure.	http://korpus.uib.no/icame/clarin/
Brown Corpus	It contains over 1 million words (500 samples of 2000+ words each) of running text of edited English prose printed in the USA in 1961.	http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/index.html
Corpus of Contemporary American English (COCA)	The corpus contains more than one billion words of text (25+ million words each year 1990-2019) from eight genres: spoken, fiction, popular magazines, newspapers, academic texts, and (with the update in March 2020): TV and Movies subtitles, blogs, and other web pages.	https://www.english-corpora.org/coca/
Georgetown University Multilayer corpus (GUM)	GUM is an open source multilayer corpus of annotated Web texts from four text types. It contains 4 million tokens and features a large number of high-quality automatic annotation layers, including dependency trees, non-named entity annotations, coreference resolution, and discourse trees in Rhetorical Structure Theory.	https://corpling.uis.georgetown.edu/gum/

[33] https://www.clarin.eu/

Name	**Description**	**Source**
Google Books Ngram Corpus	It is an online search engine that charts the frequencies of any set of search strings using a yearly count of n-grams found in sources printed between 1500 and 2019 in Google's text corpora in English, Chinese, French, German, Hebrew, Italian, Russian, or Spanish.	https://books.google.com/ngrams
International Corpus of English	The texts in the corpus date from 1990 or later and contain a total of approximately 1 million words. The corpus is built of written and spoken materials in national or regional varieties of English (Canada, East Africa, Great Britain, Hong Kong, India, Ireland, Jamaica, New Zealand, Nigeria (written), The Philippines, Singapore, Sri Lanka (written), USA (written)).	http://ice-corpora.net/ice/index.html
Oxford English Corpus	The corpus contains billions of words taken from written examples of English from around the world. The material is mainly collected from the Internet and from printed texts, such as academic journals, literary novels, everyday newspapers, magazines and from Hansard to the language of chatrooms, emails, and weblogs.	https://www.sketchengine.eu/oxford-english-corpus/
Relationship and Entity Extraction Evaluation Dataset (RE3D)	This entity's dataset was the output of a project aimed to create a 'gold standard' carried out by Aleph Insights and Committed Software on behalf of the Defence Science and Technology Laboratory (Dstl). The dataset was constructed using documents and structured schemas that were relevant to the defence and security analysis domain.	https://github.com/dstl/re3d
Santa Barbara Corpus of Spoken American English	The corpus is based on a large body of recordings of naturally occurring spoken interaction from all over the United States. It includes transcriptions, audio, and timestamps which correlate transcription and audio at the level of individual intonation units.	https://www.linguistics.ucsb.edu/research/santa-barbara-corpus
Scottish Corpus of Texts & Speech	The database contains a large electronic corpora of written and spoken texts for the languages of Scotland. It has reached a total of nearly 4.6 million words of text, with audio recordings to accompany many of the spoken texts.	https://www.scottishcorpus.ac.uk/
CETENFolha	CETENFolha (Corpus of Extracts of Electronic Texts NILC/Folha de S. Paulo) is a corpus of about 24 thousand words in Brazilian Portuguese.	https://www.linguateca.pt/cetenfolha/index_info.html
The Corpus of Electronic Texts	The Corpus of Electronic Texts, is Ireland's longest running Humanities Computing project. It contains over 18 million words, in over 1,600 contemporary and historical documents from many areas, including literature and the other arts.	https://www.ucc.ie/en/research-sites/celt/
Google Books Ngram Corpus	Google Ngram Viewer is an online search engine that charts the frequencies of any set of search strings using a yearly count of n-grams found in sources printed between 1500 and 2019 in Google's text corpora in English (British and	https://books.google.com/ngrams

Name	Description	Source
	American), Chinese, French, German, Hebrew, Italian, Russian and Spanish.	
The Georgian Language Corpus	GLC is a project of the Institute of Linguistic Studies of Ilia State University created in 2009-2015. The corpus contains over 100 000 000 word-forms and consists of the monolingual and bilingual sub-corpora. The monolingual sub-corpus comprises Old and Middle Georgian and New and Modern Georgian sections.	https://iliauni.edu.ge/en/iliauni/institutebi-451/lingvistur-kvlevata-centri-467/qartuli-jesturi-enis-korpusi
Thesaurus Linguae Graecae	TLG is a Special Research Program at the University of California, Irvine. Online TLG contains over 110 million words from 10,000 works associated with 4,000 authors and it is constantly updated and improved with new features and texts.	http://stephanus.tlg.uci.edu/
Eastern Armenian National Corpus (EANC)	EANC is a comprehensive linguistic database of annotated texts in Standard Eastern Armenian (SEA), the language spoken in the Republic of Armenia. It contains about 110 million tokens and is equipped with a search engine for making complex lexical morphological queries.	http://www.eanc.net/
Corpus of Academic Lithuanian (CorALit)	Corpus of Academic Lithuanian was compiled at the University of Vilnius. IT is a specialised synchronic corpus of written Lithuanian compiled in accordance with the modern theory and practice of corpus compilation and following the TEI P5 text encoding guidelines. The corpus includes academic texts published in 1999-2009 and consists of about 9 million words.	http://coralit.lt/en/node/18
Reference Corpus of Contemporary Portuguese (CRPC)	An electronic corpus of Portuguese (Europe, Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, S. Tome and Principe, Goa, Macao and East-Timor) containing 311,4 million words and covers written texts such as literary, newspaper, technical, etc.	http://clul.ulisboa.pt/en/projeto/crpc-reference-corpus-contemporary-portuguese
Turkish National Corpus	TNC is a balanced, 50 mln word corpus of contemporary Turkish containing samples of textual data of various genres covering a period of 1990-2013. Apart from written sources, 2% of TNC are transcriptions from spoken data which involves spontaneous, every day conversations and speeches collected in particular communicative settings.	https://www.tnc.org.tr/
TS Corpus	Corpus of Turkish aimed at developing Natural Language Processing tools and compiling linguistic datasets free of charge for academic studies and research. All the 14 published corpora serves a dataset of over 1.3 billion tokens derived from various sources (online newspapers, forums, social media, academic papers etc.).	https://tscorpus.com/
MacMorpho	A corpus of Brazilian Portuguese texts annotated with part-of-speech tags. The corpus is available for download split into train (76%), development (4%) and test (20%) sections.	http://www.nilc.icmc.usp.br/macmorpho/

Name	**Description**	**Source**
Belarusian N-korpus	Belarisian N-corpus with dictionary properties definition file of the Belarusian NooJ module. It contains circa 50000 texts which gives 30000000 tokens taken from fiction, newspapers, journals and online resources.	https://bnkorpus.info/
Russian National Corpus	Russian language corpus incorporating over 300 million words. It is a reference system based on a collection of Russian texts in electronic form.	https://ruscorpora.ru/old/en/index.html
General Internet Corpus of Russian	GICR is a megacorpus (more than 15 GT) created with a fully automated technology of collecting and tagging texts from the Russian Internet.	http://www.webcorpora.ru/en/
Ukrainian Language Corpus	uaTenTen is a Ukrainian corpus made up of texts collected from the Internet. The corpus belongs to the TenTen corpus family which is a set of the web corpora built using the same method with a target size 10+ billion words. Sketch Engine currently provides access to TenTen corpora in more than 30 languages.	https://www.sketchengine.eu/corpora-and-languages/ukrainian-text-corpora/
Araneum Russicum	Araneum Russicum is a Russian Web Corpus crawled in 2013. There are two versions of the corpus available: Araneum Russicum Maius: 1,200,001,911 tokens, 850,194,623 unmarked words and Araneum Russicum Minus: 120,139,611 tokens, 90,809,716 unmarked words.	http://ucts.uniba.sk/aranea_about/_russicum.html
Bulgarian National Corpus	The Bulgarian National corpus consists of a Bulgarian part and 47 parallel corpora. The Bulgarian part includes about 1.2 billion words in over 240 000 text samples. The materials in the Corpus reflect the state of the Bulgarian language, mainly written, from 1945 until present.	https://dcl.bas.bg/bulnc/en/
Croatian Language Corpus	Croatian Language Corpus was built by sampling spontaneous conversations among 617 speakers from all Croatian counties, and it comprises more than 250 000 tokens and more than 100 000 types. Data for the corpus were collected from 2010 to 2012, from 2014 to 2015 and during 2016.	https://ca.talkbank.org/access/Croatian.html
GOS Slovenian Corpus	GOS is a corpus of spoken Slovene that includes the transcripts of ~120 hours of speech from radio and TV shows, school lessons and lectures, private conversations between friends or within the family, work meetings, consultations, conversations in buying and selling situations, etc. All recordings are transcribed in two versions – with pronunciation-based spelling and with standardised spelling – and it comprises over one million words.	http://eng.slovenscina.eu/korpusi/gos
Czech National Corpus	CNC is a large electronic corpus of written and spoken Czech language, used for teaching and research purposes in corpus linguistics. The CNC collaborates with over 200 researchers and students (for spoken data acquisition), 270	https://korpus.cz/clarin

Name	Description	Source
	publishers (as text providers), and other similar research projects.	
National Corpus of Polish	National Corpus of Polish is over 1 billion words, of which a 300-million word subcorpus has been carefully balanced, and a manually-annotated 1-million corpus has been released under an open licence.	http://nkjp.pl/index.php?page=0&lang=1
German Reference Corpus (DeReKo)	The Mannheim German Reference Corpus (DeReKo) is a contemporary written corpus with over 46.9 billion words of electronic corpora with written German texts from today and the recent past.	https://www1.ids-mannheim.de/s/corpus-linguistics/projects/corpus-development.html?L=1

Table 33. An overview of existing language corpora in different languages.

2 Recording scenarios

The following tables show the recording scenarios that were used for the Harmonia corpus recordings. Please note that the author of this work is not the author of the recording scenarios.

2.1 Polish version (original)

Task	Speaker 1	Speaker 2
Zadanie 1: Powtórz zdanie	Powtórz zdanie „Jola lubi lody", akcentując odpowiednio te wyrazy, które usłyszysz w nagraniu.	
Zadanie 2: Przeczytaj	Przeczytaj poniższy dialog z podziałem na role. Czyta jedna osoba. Reporter: Dzień dobry! Gwiazda: Cześć! R: Gratuluję wydania nowej płyty. G: Dziękuję, nie było lekko. R: Czekaliśmy na nową płytę 2 lata. G: Tak, w tym czasie dużo koncertowałem i tworzyłem nowe utwory. R: Czy jesteś zadowolony z rezultatu? G: O, tak! Bardzo! Ale najważniejszą ocenę wystawiają zawsze słuchacze. R: Krytycy już ocenili Twoją płytę, nominując ją do Fryderyka w kategorii album roku. G: To niesamowite wyróżnienie. R: Spodziewasz się, że wygrasz? Konkurencja jest silna. G: Już sama nominacja jest dla mnie nagrodą. R: Co chciałbyś powiedzieć swoim fanom? G: Bardzo dziękuję za Wasze listy i pozytywną energię, którą mnie obdarzacie na koncertach. R: To my dziękujemy za Twoją muzykę i wywiad. G: Było mi bardzo miło. Do zobaczenia!	

Task	Speaker 1	Speaker 2
Zadanie 3: Powtórz	Powtarzaj po nauczycielce jak najdokładniej frazę po frazie z poprzedniego dialogu.	
Zadanie 4:Bezludna wyspa	Wraz z partnerem zastanów się, co zabrać ze sobą na bezludną wyspę, aby przeżyć. Możecie zabrać wspólnie 5 przedmiotów z listy: telewizor, lornetka, zapałki, gwoździe, mydło, ulubiony miś, materac, nóż, benzyna, namiot, długopis, miska, książka, młotek, latawiec	
Zadanie 5: Znajdź różnice	Współpracując z partnerem, znajdźcie 3 różnice pomiędzy obrazkami.	
Zadanie 6: Gdzie jest hotel?	Turysta: Wyobraź sobie, że dotarłeś pociągiem do nieznanego miasta. Znajdujesz się na Dworcu Głównym i dzwonisz do hotelu, w którym masz się zatrzymać, aby uzyskać informację, jak do niego dotrzeć. W posiadaniu masz mapę, która pomoże Ci tam trafić.	Recepcjonista: Wyobraź sobie, że pracujesz jako recepcjonista w hotelu. Dzwoni do Ciebie gość hotelowy. Udziel mu informacji, o które prosi. W posiadaniu masz mapę miasta.
Zadanie 7: Gdzie jest hotel?	Recepcjonista: Wyobraź sobie, że pracujesz jako recepcjonista w hotelu. Dzwoni do Ciebie gość hotelowy. Udziel mu informacji, o które prosi. W posiadaniu masz mapę miasta.	Turysta: Wyobraź sobie, że dotarłeś pociągiem do nieznanego miasta. Znajdujesz się na Dworcu Głównym i dzwonisz do hotelu, w którym masz się zatrzymać, aby uzyskać informację, jak do niego dotrzeć. W posiadaniu masz mapę, która pomoże Ci tam trafić.
Zadanie 8: Ekspresja w dialogu	Informator: Pracujesz w informacji turystycznej dużego miasta i właśnie dowiedziałeś się, że Twoje biuro jest najlepsze. Twoim zadaniem jest udzielenie informacji o wydarzeniach i ciekawych miejscach w mieście i przekonanie rozmówcy, by skorzystał choć z jednej z Twoich 3 propozycji. Jeśli przekonasz rozmówcę, to dostaniesz od szefa wysoka nagrodę.	Imprezowicz: Masz dzisiaj wolny wieczór i chcesz wyjść z domu, aby trochę się rozerwać. Dzwonisz do informacji turystycznej dużego miasta, aby dowiedzieć się, jakie atrakcje miasto oferuje na dzisiejszy wieczór. Wybierasz jedną z propozycji.
Zadanie 9: Ekspresja w dialogu	Imprezowicz: Masz dzisiaj wolny wieczór i chcesz wyjść z domu, aby trochę się rozerwać. Dzwonisz do informacji turystycznej dużego miasta, aby dowiedzieć się, jakie atrakcje miasto oferuje na dzisiejszy wieczór.	Informator: Pracujesz w informacji turystycznej dużego miasta i właśnie dowiedziałeś się, że Twoje biuro jest najlepsze. Twoim zadaniem jest udzielenie informacji o wydarzeniach i ciekawych miejscach w mieście i przekonanie

Task	Speaker 1	Speaker 2
	Wybierasz jedną z propozycji.	rozmówcy, by skorzystał choć z jednej z Twoich 3 propozycji. Jeśli przekonasz rozmówcę, to dostaniesz od szefa wysoka nagrodę.
Zadanie 10: Ekspresja w dialogu	Informator: Pracujesz w informacji turystycznej dużego miasta i właśnie dowiedziałeś się, że w Twoim mieście są zamachy terrorystyczne. Twoim zadaniem jest jednak udzielenie informacji o wydarzeniach i ciekawych miejscach w mieście i przekonanie rozmówcy, by skorzystał z Twoich 3 propozycji, które wydają Ci się bezpieczne.	Imprezowicz: Masz dzisiaj wolny wieczór i pomimo niebezpieczeństwa zamachów terrorystycznych chcesz wyjść z domu, aby trochę się rozerwać. Dzwonisz do informacji turystycznej dużego miasta, aby dowiedzieć się, jakie atrakcje miasto oferuje na dzisiejszy wieczór . Wybierasz najbezpieczniejszą propozycję.
Zadanie 11: Ekspresja w dialogu	Imprezowicz: Masz dzisiaj wolny wieczór i pomimo niebezpieczeństwa zamachów terrorystycznych chcesz wyjść z domu, aby trochę się rozerwać. Dzwonisz do informacji turystycznej dużego miasta, aby dowiedzieć się, jakie atrakcje miasto oferuje na dzisiejszy wieczór. Wybierasz najbezpieczniejszą propozycję.	Informator: Pracujesz w informacji turystycznej dużego miasta i właśnie dowiedziałeś się, że w Twoim mieście są zamachy terrorystyczne. Twoim zadaniem jest jednak udzielenie informacji o wydarzeniach i ciekawych miejscach w mieście i przekonanie rozmówcy, by skorzystał z Twoich 3 propozycji, które wydają Ci się bezpieczne.
Zadanie 12: Prowokacje w sztuce	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z partnerem i wymień opinie na ich temat. Oboje macie być zgodni i pochwalać tę formę sztuki.	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z partnerem i wymień opinie na ich temat. Oboje macie być zgodni i pochwalać tę formę sztuki.
Zadanie 13: Prowokacje w sztuce	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z partnerem i wymień opinie na ich temat. Oboje macie być zgodni i być przeciwnikami tej wystawy.	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z partnerem i wymień opinie na ich temat. Oboje macie być zgodni i być przeciwnikami tej wystawy.
Zadanie 14: Prowokacje w sztuce (dialog z Nauczycielką)	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z nauczycielką i wymień opinie na ich temat. Oboje	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z nauczycielką i wymień opinie na ich temat. Oboje

Task	Speaker 1	Speaker 2
	macie być zgodni i pochwalać tę formę sztuki.	macie być zgodni i pochwalać tę formę sztuki.
Zadanie 15: Prowokacje w sztuce (dialog z Nauczycielką)	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z nauczycielką i wymień opinie na ich temat. Oboje macie być zgodni i być przeciwnikami tej wystawy.	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z nauczycielką i wymień opinie na ich temat. Oboje macie być zgodni i być przeciwnikami tej wystawy.
Zadanie 16: Prowokacje w sztuce (dialog z Nauczycielką)	Rozmowa zwolennika i przeciwnika prowokacji w sztuce. Każdy stoi przy swoim. Zwolennik: Jesteś fanem sztuki współczesnej, lubisz prowokacje i kochasz Madonnę. Porozmawiaj z nauczycielką i wymień opinie na temat zaprezentowanej fotografii.	Rozmowa zwolennika i przeciwnika prowokacji w sztuce. Każdy stoi przy swoim. Zwolennik: Jesteś fanem sztuki współczesnej, lubisz prowokacje i kochasz Madonnę. Porozmawiaj z nauczycielką i wymień opinie na temat zaprezentowanej fotografii.

Table 34. The speakers' instructions from the Harmonia corpus recordings' scenarios - Polish version (original).

2.2 English translation

Task	Speaker 1	Speaker 2
Task 1: Repeat the sentence	Repeat the sentence „Jola likes ice cream", emphasizing the words you hear in the recording.	
Task 2: Read	Read the role-playing dialogue below. One person is reading. Reporter: Good morning! Star: Hi! R: Congratulations on the release of your new album. G: Thank you, it wasn't easy. R: We've been waiting for a new album for 2 years. G: Yes, at that time I toured a lot and created new songs. R: Are you satisfied with the result? G: Oh, yes! Very! But the most important evaluation is always made by the listeners. R: Critics have already rated your album, nominating it for Fryderyk in the album of the year category. G: It's an amazing honor. R: Do you expect to win? The competition is strong. G: The nomination itself is a reward for me. R: What would you like to say to your fans? G: Thank you very much for your letters and the positive energy you give me at concerts. R: We thank you for your music and interview. G: I was very pleased. See you!	
Task 3: Repeat.	Repeat after the teacher as precisely as possible phrase by phrase from the previous dialogue.	
Task 4: Desert Island	With your partner, discuss what would be necessary on a desert island to survive. You can take 5 items from the list: TV set, binoculars, matches, nails, soap, favorite bear,	

Task	Speaker 1	Speaker 2
	mattress, knife, gasoline, tent, pen, bowl, book, hammer, kite	
Task 5: Find the differences	Work with your partner to find 3 differences between the pictures.	Work with your partner to find 3 differences between the pictures.
Task 6: Where is the hotel?	Tourist: Imagine that you have arrived in an unfamiliar city by train. You are at the Central Railway Station and you call the hotel where you are to stay to get information on how to get there. You have a map to help you get there.	Receptionist: Imagine that you work as a receptionist in a hotel. A hotel guest is calling you. Give him the information he asks for. You have a city map.
Task 7: Where is the hotel?	Receptionist: Imagine that you work as a receptionist in a hotel. A hotel guest is calling you. Give him the information he asks for. You have a city map in your possession.	Tourist: Imagine that you have arrived in an unfamiliar city by train. You are at the Central Railway Station and you call the hotel where you are to stay to get information on how to get there. You have a map to help you get there.
Task 8: Expression in dialogue	Information provider: You work in a tourist information office in a big city and you just found out that your office is the best. Your task is to provide information about events and interesting places in the city and convince the interlocutor to take advantage of at least one of your 3 suggestions. If you convince the interlocutor, you will get a high reward from the boss.	Party person: You're free tonight and want to get out of the house to have some fun. You call the tourist information of a large city to find out what attractions the city has to offer for tonight. You choose one of the proposals.
Task 9: Expression in dialogue	Party person: You're free tonight and want to get out of the house to have some fun. You call the tourist information of a large city to find out what attractions the city has to offer for tonight. You choose one of the proposals.	Information provider: You work in a tourist information office in a big city and you just found out that your office is the best. Your task is to provide information about events and interesting places in the city and convince the interlocutor to take advantage of at least one of your 3 suggestions. If you convince the interlocutor, you will get a high reward from the boss.
Task 10: Expression in dialogue	Information provider: You work in the tourist information of a large city and you have just learned	Party person: You have the evening off tonight, and despite the threat of terrorist attacks, you want to get out

Task	Speaker 1	Speaker 2
	that there are terrorist attacks in your city. However, your task is to provide information about events and interesting places in the city and convince the interlocutor to use your 3 suggestions that seem safe to you.	of the house to have some fun. You call the tourist information of a large city to find out what attractions the city has to offer tonight. You choose the safest offer.
Task 11: Expression in dialogue	Party person: You have the evening off tonight, and despite the threat of terrorist attacks, you want to get out of the house to have some fun. You call the tourist information of a large city to find out what attractions the city has to offer for tonight. You choose the safest	Information provider: You work in the tourist information of a large city and you have just learned that there are terrorist attacks in your city. However, your task is to provide information about events and interesting places in the city and convince the interlocutor to use your 3 suggestions that seem safe to you.
Task 12: Art provocations	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z partnerem i wymień opinie na ich temat. Oboje macie być zgodni i pochwalać tę formę sztuki.	Co sądzisz o przedstawionych 3 obrazkach? Porozmawiaj z partnerem i wymień opinie na ich temat. Oboje macie być zgodni i pochwalać tę formę sztuki.
Task 13: Art provocations	What do you think about these 3 pictures? Talk to your partner and exchange opinions about them. You are both to agree and approve of this art form.	What do you think about these 3 pictures? Talk to your partner and exchange opinions about them. You are both to agree and approve of this art form.
Task 14: Provocations in Art (dialogue with the Teacher)	What do you think about these 3 pictures? Talk to the teacher and exchange opinions about them. You are both to agree and approve of this art form.	What do you think about these 3 pictures? Talk to the teacher and exchange opinions about them. You are both to agree and approve of this art form.
Task 15: Art provocations (dialogue with the Teacher)	What do you think about these 3 pictures? Talk to the teacher and exchange opinions about them. Both of you are to agree and be opponents of this exhibition.	What do you think about these 3 pictures? Talk to the teacher and exchange opinions about them. Both of you are to agree and be opponents of this exhibition.
Task 16: Art provocations (dialogue with the Teacher)	Conversation between the supporter and opponent of provocation in art. Everyone stands by their own. Supporter: You are a fan of contemporary art, you like provocation and you love Madonna. Talk to the	Conversation between the supporter and opponent of provocation in art. Everyone stands by their own. Supporter: You are a fan of contemporary art, you like provocation and you love Madonna. Talk to the

Task	Speaker 1	Speaker 2
	teacher and exchange opinions on the presented photo.	teacher and exchange opinions on the presented photo.

Table 35. The speakers' instructions from the Harmonia corpus recordings' scenarios - English version (translation).

3 Python scripts

The following chapters present scripts written by the author of this work that enabled the use of automatic taggers POS Space and Multiservice NLP - Concraft-pl (Chapter V.3.1 and V.3.3). Chapter V.3.2 presents the script used for Spacy tagging results, which automatically counted occurrences of words from grammatical categories in individual utterances of speakers in each task separately. Similarly in V.3.4, the script counts and saves the results in a spreadsheet for the annotation made with Multiservice NLP - Concraft-pl.

3.1 Part-of-Speech tagging with Spacy

```
import os
import re
import shutil
from lpmn_client import download_file, upload_file
from lpmn_client import Task

global path
path = str(os.path.abspath(os.getcwd()))

def spacy(file_name):
    task                                                    =
Task(lpmn='any2txt|spacy({„method”:”tagger”,”lang”:”pl”
})')
    task.email = „karolinapieniowska@gmail.com”
    file_id = upload_file(file_name)
    output_file_id = task.run(file_id)

download_file(output_file_id,str(path)+”\\data\\output”
)

def input_organizer(input_file,_zip=False):

    with open(input_file,”r”,encoding=„utf8”) as _input:
        content = _input.readlines()

        for line in content:
            if „_SPK” in line:
```

```
                with
open(path+"\\data\\input\\"+line.strip(„\n")+".txt","w"
,encoding=„utf8") as _output:
                    _output.write(line)

_output.write(content[content.index(line)+1].strip(„\n"
))

        if _zip == True:
            archive                                    =
shutil.make_archive(path+"\\"+input_file.strip(„.txt"),
"zip",path+"\\data\\input")
```

3.2 Part-of-Speech statistics extraction for Spacy

```
import os
import re
import xml.etree.ElementTree as ET

_all = []
_all.append([„File","Nouns","...","Verbs","...","Adject
ives","...","Adverbs","...","Part","...","Pronouns","..
.","Det","...","Adp","...","Aux","...","Cconj","...","I
ntj","...","Num","...","Propn","...","Punct","...","Sco
nj","...","Sym","...","Others (X)","..."])

def stats(folder):
    path                                               =
str(os.path.abspath(os.getcwd()))+"\\data\\output\\"+fo
lder
    path_files = os.listdir(path)

    for file in path_files:
        with open(path+"\\"+file,"r",encoding=„utf8") as
_input:
            content = _input.readlines()
            verbs = []
            nouns = []
            adjectives = []
            adverbs = []
            part = []
            pronouns = []
            det = []
            adp = []
            aux = []
            cconj = []
            intj = []
            num = []
            propn = []
            punct = []
            sconj = []
            sym = []
            x = []
```

```
            for i in content:
                if „NOUN” in i:

nouns.append(str(re.search(„<base>.+</base>„,i)).split(
„\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „ADJ” in i:

adjectives.append(str(re.search(„<base>.+</base>„,i)).s
plit(„\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „ADV” in i:

adverbs.append(str(re.search(„<base>.+</base>„,i)).spli
t(„\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „PART” in i:

part.append(str(re.search(„<base>.+</base>„,i)).split(„
\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „PRON” in i:

pronouns.append(str(re.search(„<base>.+</base>„,i)).spl
it(„\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „VERB” in i:

verbs.append(str(re.search(„<base>.+</base>„,i)).split(
„\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „DET” in i:

det.append(str(re.search(„<base>.+</base>„,i)).split(„\
'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „ADP” in i:

adp.append(str(re.search(„<base>.+</base>„,i)).split(„\
'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „AUX” in i:

aux.append(str(re.search(„<base>.+</base>„,i)).split(„\
'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „CCONJ” in i:

cconj.append(str(re.search(„<base>.+</base>„,i)).split(
„\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „INTJ” in i:

intj.append(str(re.search(„<base>.+</base>„,i)).split(„
\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „NUM” in i:

num.append(str(re.search(„<base>.+</base>„,i)).split(„\
'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „PROPN” in i:
```

```
propn.append(str(re.search(„<base>.+</base>„,i)).split(
„\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „PUNCT” in i:

punct.append(str(re.search(„<base>.+</base>„,i)).split(
„\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „SCONJ” in i:

sconj.append(str(re.search(„<base>.+</base>„,i)).split(
„\'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „SYM” in i:

sym.append(str(re.search(„<base>.+</base>„,i)).split(„\
'„)[-2].split(„</base>„)[0].split(„<base>„)[1])
                elif „X” in i:
                    try:

x.append(str(re.search(„<base>.+</base>„,i)).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1])
                    except IndexError:
                        pass



            for i in verbs:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    verbs.remove(i)
            for i in nouns:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    nouns.remove(i)
            for i in adjectives:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    adjectives.remove(i)
            for i in adverbs:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    adverbs.remove(i)
            for i in part:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    part.remove(i)
            for i in pronouns:
```

```
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    pronouns.remove(i)
            for i in det:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    det.remove(i)
            for i in adp:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    adp.remove(i)
            for i in aux:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    aux.remove(i)
            for i in cconj:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    cconj.remove(i)
            for i in intj:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    intj.remove(i)
            for i in num:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    num.remove(i)
            for i in propn:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    propn.remove(i)
            for i in punct:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    punct.remove(i)
            for i in sconj:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    sconj.remove(i)
            for i in sym:
                if                  i                   ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
```

```
                    sym.remove(i)
            for i in x:
                if                  i                  ==
str(re.search(„<base>.+</base>„,content[7])).split(„\'„
)[-2].split(„</base>„)[0].split(„<base>„)[1]:
                    x.remove(i)


_all.append([str(re.search(„<base>.+</base>„,content[7]
)).split(„\'„)[-
2].split(„</base>„)[0].split(„<base>„)[1],

len(set(nouns)),set(nouns),len(set(verbs)),set(verbs),l
en(set(adjectives)),set(adjectives),len(set(adverbs)),s
et(adverbs),len(set(part)),set(part),len(set(pronouns))
,set(pronouns),len(set(det)),set(det),len(set(adp)),set
(adp),len(set(aux)),set(aux),len(set(cconj)),set(cconj)
,len(set(intj)),set(intj),

len(set(num)),set(num),len(set(propn)),set(propn),len(s
et(punct)),set(punct),len(set(sconj)),set(sconj),len(se
t(sym)),set(sym),len(set(x)),set(x)])


    with
open(str(os.path.abspath(os.getcwd()))+"\\data\\output\
\"+folder+".txt",'w',encoding=„utf8") as _out:
        for i in _all:

_out.write(str(i[0])+"||"+str(i[1])+"||"+str(i[2])+"||"
+str(i[3])+"||"+str(i[4])+"||"+str(i[5])+"||"+str(i[6])
+"||"+str(i[7])+"||"+str(i[8])+"||"+str(i[9])+"||"+str(
i[10])+"||"+str(i[11])+"||"+str(i[12])+"||"+str(i[13])+
"||"+str(i[14])+"||"+str(i[15])+"||"+str(i[16])+"||"+st
r(i[17])+"||"+str(i[18])+"||"+str(i[19])+"||"+str(i[20]
)+"||"+str(i[21])+"||"+str(i[22])+"||"+str(i[23])+"||"+
str(i[24])+"||"+str(i[25])+"||"+str(i[26])+"||"+str(i[2
7])+"||"+str(i[28])+"||"+str(i[29])+"||"+str(i[30])+"||
"+str(i[31])+"||"+str(i[32])+"||"+str(i[33])+"||"+str(i
[34])+"\n")
```

3.3 Part-of-Speech tagging with Multiservice NLP - Concraft-pl

```
import requests
import re
import csv
import os
from selenium import webdriver
import time
from selenium.webdriver.common.by import By
import selenium.common.exceptions
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
```

```
def open_chrome():
        time.sleep(5)
        URL                                                 =
'http://multiservice.nlp.ipipan.waw.pl/pl/'
        driver                                              =
webdriver.Chrome(ChromeDriverManager().install())
        driver.get(URL)


def files_handler():
        path_input                                          =
str(os.path.abspath(os.getcwd()))+"\\data\\input\\"
        path_input_files = os.listdir(path_input)
        path_output                                         =
str(os.path.abspath(os.getcwd()))+"\\data\\output\\"


        for file in path_input_files:
                tagger(path_input+file)
                time.sleep(10)
                path_output_files                           =
os.listdir(path_output)
                print(path_output_files)

os.rename(path_output+path_output_files[-
1],path_output+file.strip(„.")+"_IPIPAN.txt")


def tagger(input_file):
        with open(input_file,"r",encoding = „utf-8") as
_input:
                content = _input.readlines()

                textarea                                    =
driver.find_element(By.XPATH,"//textarea")
                textarea.clear()
                textarea.send_keys([content[1]])
                button                                      =
driver.find_element(By.XPATH,"//button[@id='doitButton'
]")
                button.click()
```

3.4 Part-of-Speech statistics extraction for Multiservice NLP - Concraft-pl

```
import os
import re
import xml.etree.ElementTree as ET

global _all
_all = []
```

```
_all.append(['File', 'subts', '...', 'depr', '...',
'num', '...', 'numcol', '...', 'adj', '...', 'adja',
'...', 'adjp', '...', 'adjc', '...', 'adv', '...',
'ppron12', '...', 'ppron3', '...', 'siebie', '...',
'fin', '...', 'bedzie', '...', 'aglt', '...', 'praet',
'...', 'impt', '...', 'imps', '...', 'inf', '...',
'pcon', '...', 'pant', '...', 'ger', '...', 'pact',
'...', 'ppas', '...', 'winien', '...', 'pred', '...',
'prep', '...', 'cong', '...', 'comp', '...', 'qub',
'...', 'brev', '...', 'burk', '...', 'interj', '...',
'interp', '...', 'xxx', '...', 'ign', '...'])
path                                                =
str(os.path.abspath(os.getcwd()))+"\\data\\output\\CoNL
L"
path_files = os.listdir(path)

def stats_CoNLL(w_duplicates = False):
    path                                            =
str(os.path.abspath(os.getcwd()))+"\\data\\output\\CoNL
L"
    path_files = os.listdir(path)

    for file in path_files:
        with open(path+"\\"+file,"r",encoding="utf8") as
_input:
            global content
            content = _input.readlines()


            subst=    []
            depr=     []
            num= []
            numcol=   []
            adj= []
            adja=     []
            adjp=     []
            adjc=     []
            adv= []
            ppron12=  []
            ppron3=   []
            siebie=   []
            fin= []
            bedzie=   []
            aglt=     []
            praet=    []
            impt=     []
            imps=     []
            inf= []
            pcon=     []
            pant=     []
            ger= []
            pact=     []
            ppas=     []
```

```
            winien=   []
            pred=     []
            prep=     []
            cong=     []
            comp=     []
            qub= []
            brev=     []
            burk=     []
            interj=   []
            interp=   []
            xxx= []
            ign= []

            for i in content:
                if i != „\n”:

                    if str(i).split(„\t”)[3] == „subst”:
                        subst.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „depr”:
                        depr.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „num”:
                        num.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „numcol”:
                        numcol.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „adj”:
                        adj.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „adja”:
                        adja.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „adjp”:
                        adjp.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „adjc”:
                        adjc.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „adv”:
                        adv.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „ppron12”:

ppron12.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „ppron3”:
                        ppron3.append(i.split(„\t”)[2])
                    elif  i[0].isnumeric()  ==  True  and
i.split(„\t”)[3] == „siebie”:
                        siebie.append(i.split(„\t”)[2])
```

```
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „fin”:
                        fin.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „bedzie”:
                        bedzie.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „aglt”:
                        aglt.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „praet”:
                        praet.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „impt”:
                        impt.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „imps”:
                        imps.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „inf”:
                        inf.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „pcon”:
                        pcon.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „pant”:
                        pant.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „ger”:
                        ger.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „pact”:
                        pact.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „ppas”:
                        ppas.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „winien”:
                        winien.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „pred”:
                        pred.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „prep”:
                        prep.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „cong”:
                        cong.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „comp”:
                        comp.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „qub”:
```

```
                        qub.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „brev”:
                        brev.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „burk”:
                        burk.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „interj”:
                        interj.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „interp”:
                        interp.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „xxx”:
                        xxx.append(i.split(„\t”)[2])
                    elif i[0].isnumeric() == True and
i.split(„\t”)[3] == „ign”:
                        ign.append(i.split(„\t”)[2])



        _all.append([file,             len(subst),subst,
len(depr),depr,    len(num),num,    len(numcol),numcol,
len(adj),adj,      len(adja),adja,      len(adjp),adjp,
len(adjc),adjc,   len(adv),adv,   len(ppron12),ppron12,
len(ppron3),ppron3,  len(siebie),siebie,  len(fin),fin,
len(bedzie),bedzie,  len(aglt),aglt,  len(praet),praet,
len(impt),impt,      len(imps),imps,      len(inf),inf,
len(pcon),pcon,      len(pant),pant,      len(ger),ger,
len(pact),pact,   len(ppas),ppas,   len(winien),winien,
len(pred),pred,     len(prep),prep,     len(cong),cong,
len(comp),comp,      len(qub),qub,      len(brev),brev,
len(burk),burk, len(interj),interj, len(interp),interp,
len(xxx),xxx, len(ign),ign])


    with
open(str(os.path.abspath(os.getcwd()))+”\\data\\output\
\CoNLL_stats.txt”,'w',encoding=„utf8”) as _out:
        for i in _all:
            _out.write(str(i[0]) + „\t” + str(i[1]) +
„\t” + str(i[2]) + „\t” + str(i[3]) + „\t” + str(i[4]) +
„\t” + str(i[5]) + „\t” + str(i[6]) + „\t” + str(i[7]) +
„\t” + str(i[8]) + „\t” + str(i[9]) + „\t” + str(i[10])
+ „\t” + str(i[11]) + „\t” + str(i[12]) + „\t” +
str(i[13]) + „\t” + str(i[14]) + „\t” + str(i[15]) + „\t”
+ str(i[16]) + „\t” + str(i[17]) + „\t” + str(i[18]) +
„\t” + str(i[19]) + „\t” + str(i[20]) + „\t” + str(i[21])
+ „\t” + str(i[22]) + „\t” + str(i[23]) + „\t” +
str(i[24]) + „\t” + str(i[25]) + „\t” + str(i[26]) + „\t”
+ str(i[27]) + „\t” + str(i[28]) + „\t” + str(i[29]) +
„\t” + str(i[30]) + „\t” + str(i[31]) + „\t” + str(i[32])
```

```
+ „\t” + str(i[33]) + „\t” + str(i[34]) + „\t” +
str(i[35]) + „\t” + str(i[36]) + „\t” + str(i[37]) + „\t”
+ str(i[38]) + „\t” + str(i[39]) + „\t” + str(i[40]) +
„\t” + str(i[41]) + „\t” + str(i[42]) + „\t” + str(i[43])
+ „\t” + str(i[44]) + „\t” + str(i[45]) + „\t” +
str(i[46]) + „\t” + str(i[47]) + „\t” + str(i[48]) + „\t”
+ str(i[49]) + „\t” + str(i[50]) + „\t” + str(i[51]) +
„\t” + str(i[52]) + „\t” + str(i[53]) + „\t” + str(i[54])
+ „\t” + str(i[55]) + „\t” + str(i[56]) + „\t” +
str(i[57]) + „\t” + str(i[58]) + „\t” + str(i[59]) + „\t”
+ str(i[60]) + „\t” + str(i[61]) + „\t” + str(i[62]) +
„\t” + str(i[63]) + „\t” + str(i[64]) + „\t” + str(i[65])
+ „\t” + str(i[66]) + „\t” + str(i[67]) + „\t” +
str(i[68]) + „\t” + str(i[69]) + „\t” + str(i[70]) + „\t”
+ str(i[71]) + „\t” + str(i[72])+”\n”)
```

4 Manually annotated corpus

In the attachment in the form of an .xls file, manually tagged list of words from transcripts of Tasks 5-16 from the Harmonia corpus have been added. The file contains a frequency list of unaltered word forms with and tags. The first column contains information about the recording and the speaker. In the second sheet there is a table with the applied tags and their explanation.
The corpus is available on the drive:
https://docs.google.com/spreadsheets/d/1vZo2453SwCgpiIDv4-eta2_hoexCFoiu/

5 LSM results

Table 37 presents the exact results of the LSM factor for all grammatical categories and for each dialogue in the Harmonia corpus separately. The colours in the fields of the names of the recordings define the groups of tasks (yellow - cooperation, common goal, green - expression, persuasiveness, blue - provocative, arousing emotions (students), purple - provocative, arousing emotions (students with Teacher)). The results are colour-coded on a green-red scale, where green is used for the lowest values and red for the highest. The first column contains the abbreviations of the grammatical categories for which the LSM factor was calculated. These are the same abbreviations (tags) that were used to tag the Harmonia corpus.

Tag	Meaning	Tag	Meaning
adj	adjectives	**negpro**	negative pronouns
adv	adverbs	**noun**	nouns
advtime	adverbs time	**num**	numerals
aux	auxilary verbs	**other**	other
cconj	conjunctions	**part**	particles
comadv	common adverbs	**perspro**	personal pronouns
dempro	demonstrative pronouns	**propname**	proper names
genpro	generalizing pronouns	**posspro**	possessive pronouns
indepro	indefinite pronouns	**prepos**	prepositions
indnum	indefinite numerals	**refpro**	reflexive pronouns
intj	interjections	**verb**	verbs
intpro	interogative pronouns	**welfar**	welcome/farewell

Table 36. List of tags with their meanings used for manual corpus annotation.

N04	**n04_z05**	**n04_z06**	**n04_z07**	**n04_z08**	**n04_z09**	**n04_z10**	**n04_z11**	**n04_z12**	**n04_z13**	**n04_z14_p1**	**n04_z15_p1**	**n04_z16_p1**	**n04_z14_p2**	**n05_z15_p2**	**n04_z16_p2**
aux	0,5	0,95	0,95	0,96	0,6	0,81	0,49	0,75	0,64	0,78	0,94	0,88	0,73	0,91	0,7
cconj	0,99	0,82	0,84	0,84	0,97	0,73	0,98	0,99	0,91	0,67	0,89	0,82	0,78	0,86	0,76
comadv	0,92	0,3	0,48	0,84	0,38	0,66	0,74	0,67	0,08	0,94	0,65	0,86	0,52	0,55	0,78
dempro	0,59	0,9	0,03	0,84	0,48	0,9	0,9	0,94	0,96	0,55	0,89	0,82	0,92	0,91	0,83
genpro	0,57	1	0,14	1	0,07	0,72	0,53	0,1	0,15	1	1	0,92	0,76	0,62	0,42
indepro	0,12	0,12	1	0,55	0,22	0,57	0,9	0,1	0,48	0,43	0,55	0,43	0,98	0,14	0,95
intj	0,59	0,02	0,32	0,11	0,07	0,33	0,08	1	1	1	1	0,36	0,14	0,08	0,67
intpro	0,87	0,96	1	0,84	0,44	0,73	0,64	1	0,82	1	0,84	0,68	0,21	0,81	0,36
negpro	0,14	0,12	1	1	0,04	0,24	0,12	1	0,15	0,26	1	0,46	0,43	0,22	1
part	0,73	0,75	0,82	0,84	0,78	0,72	0,79	0,95	0,74	0,92	0,97	0,89	0,91	0,96	0,93
perspro	0,7	0,71	0,55	0,55	0,1	0,74	0,39	0,04	0,77	0,86	0,98	0,86	0,68	0,96	0,59
posspro	0,06	0,16	1	1	1	0,14	1	0,09	0,14	1	0,66	0,32	0,34	0,22	0,32
prepos	0,83	0,63	0,73	0,8	0,96	0,79	0,8	0,72	0,98	0,96	0,96	0,91	0,86	0,87	0,69
refpro	0,36	0,12	0,82	0,46	0,99	0,79	0,84	1	0,77	0,18	0,79	0,97	0,78	0,72	0,76
AVR	0,57	0,54	0,69	0,76	0,51	0,63	0,66	0,67	0,61	0,75	0,87	0,73	0,65	0,63	0,70
N06	**n06_z05**	**n06_z06**	**n06_z07**	**n06_z08**	**n06_z09**	**n06_z10**	**n06_z11**	**n06_z12**	**n06_z13**	**n06_z14_p1**	**n06_z15_p1**	**n06_z16_p1**	**n06_z14_p2**	**n06_z15_p2**	**n06_z16_p2**
aux	0,81	0,85	0,52	0,86	0,93	0,9	0,68	0,92	0,75	0,78	0,74	0,75	0,97	0,85	0,82
cconj	0,8	0,86	0,93	0,94	0,89	0,84	0,76	0,94	0,94	0,94	0,99	0,94	0,89	0,94	0,96
comadv	0,06	1	0,92	0,91	0,54	0,46	0,64	0,95	0,78	0,94	0,05	0,9	0,91	0,75	0,87
dempro	0,88	0,55	0,98	0,98	0,26	0,46	0,61	0,76	0,88	0,2	0,44	0,64	0,91	0,65	0,9
genpro	0,11	1	0,18	0,22	1	1	0,97	0,1	0,79	0,18	0,66	0,49	0,18	0,19	0,89

indepro	0,57	0,22	0,22	0,37	0,81	0,69	0,77	0,91	0,63	0,05	0,84	0,93	0,37	0,73	0,89
intj	1	0,43	1	1	0,17	0,11	1	0,6	1	0,88	0,1	0,68	1	1	0,27
intpro	0,43	0,57	0,93	0,86	0,62	0,83	0,83	0,93	0,6	0,97	0,97	0,96	0,1	0,92	0,67
negpro	0,88	0,57	1	1	0,09	1	1	0,37	1	1	0,85	0,14	1	1	0,22
part	0,86	0,97	0,93	0,92	0,74	0,83	0,91	0,81	0,95	0,96	0,93	0,92	0,87	0,9	0,9
perspro	0,97	0,42	0,59	0,64	1	0,06	0,24	0,79	0,92	0,57	0,67	0,86	0,98	0,95	0,51
posspro	0,09	1	1	1	1	0,58	1	0,23	0,58	0,13	0,84	0,9	0,91	0,17	0,57
prepos	0,89	0,66	0,93	0,88	0,89	0,94	0,93	0,89	0,81	0,96	0,89	0,89	0,78	0,98	0,79
refpro	0,74	0,97	0,72	0,83	0,88	0,98	0,85	0,81	0,69	0,71	0,83	0,98	0,91	0,7	0,69
AVR	0,65	0,72	0,78	0,82	0,7	0,69	0,8	0,72	0,81	0,66	0,7	0,78	0,77	0,77	0,71
N07	**n07_z05**	**n07_z06**	**n07_z07**	**n07_z08**	**n07_z09**	**n07_z10**	**n07_z11**	**n07_z12**	**n07_z13**	**n07_z14_p1**	**n07_z15_p1**	**n07_z16_p1**	**n07_z14_p2**	**n07_z15_p2**	**n07_z16_p2**
aux	0,85	0,88	0,72	0,54	0,22	0,93	0,68	0,62	0,66	0,98	0,62	0,96	0,75	0,02	0,74
cconj	0,5	0,98	0,56	0,92	0,68	0,81	0,93	0,67	0,55	0,94	0,86	0,88	0,88	0,49	0,96
comadv	0,1	0,57	0,86	0,6	0,26	0,9	0,66	0,72	0,87	0,8	0,33	0,68	0,88	0,57	0,71
dempro	0,99	0,5	0,03	0,78	0,89	0,62	0,87	0,91	0,93	0,92	0,7	0,95	0,94	0,65	0,73
genpro	1	1	1	0,13	1	0,59	1	0,06	1	1	1	0,13	1	0,2	0,72
indepro	0,1	1	1	0,09	0,04	0,28	0,41	1	0,28	0,88	0,65	0,61	0,25	0,02	0,63
intj	0,99	0,7	1	1	0,92	1	0,07	1	0,87	1	0,72	1	1	1	0,33
intpro	1	0,43	0,16	0,69	0,06	0,06	0,13	0,04	0,56	0,08	0,7	0,99	0,99	0,75	0,93
negpro	1	1	1	0,22	1	1	0,13	0,12	1	1	1	0,42	1	1	0,2
part	0,7	0,91	0,87	0,67	0,75	0,83	0,82	0,93	1	0,86	0,76	0,97	0,87	0,8	0,94
perspro	0,99	0,02	0,33	0,54	0,58	0,1	0,03	1	0,57	0,69	0,71	0,82	0,99	0,69	0,01
posspro	0,4	1	0,16	0,55	0,1	0,62	1	1	0,15	0,8	0,22	0,87	1	0,11	0,57
prepos	0,81	0,73	0,68	0,98	0,59	0,84	0,63	0,55	0,95	0,96	0,83	0,9	0,84	0,94	0,92
refpro	0,23	0,71	0,79	0,73	0,75	0,71	0,82	0,12	0,34	0,21	0,58	0,67	0,03	0,06	0,49
AVR	0,69	0,75	0,65	0,6	0,56	0,66	0,58	0,62	0,7	0,79	0,69	0,78	0,82	0,52	0,63
N08	**n08_z05**	**n08_z06**	**n08_z07**	**n08_z08**	**n08_z09**	**n08_z10**	**n08_z11**	**n08_z12**	**n08_z13**	**n08_z14_p1**	**n08_z15_p1**	**n08_z16_p1**	**n08_z14_p2**	**n08_z15_p2**	**n08_z16_p2**
aux	0,8	0,89	0,77	0,03	0,51	0,68	0,93	0,86	0,69	0,73	0,79	0,93	0,75	0,8	0,87
cconj	0,74	0,81	0,88	0,4	0,66	0,72	0,94	0,94	0,84	0,98	0,92	0,81	0,92	0,84	0,93
comadv	0,86	0,77	0,97	0,39	0,5	0,89	0,97	0,02	1	0,62	0,84	0,64	0,68	0,63	0,66
dempro	0,94	0,77	0,03	0,36	0,66	0,14	0,96	0,16	0,69	0,8	0,97	0,95	0,93	0,68	0,8
genpro	0,08	1	1	0,03	1	1	0,1	0,81	0,03	0,43	0,12	0,92	1	0,86	0,74
indepro	0,86	0,73	1	0,03	0,04	0,66	0,54	0,47	1	0,19	0,78	0,76	0,47	0,33	0,48
intj	0,14	1	1	1	1	1	1	1	0,04	0,19	1	0,92	0,17	0,35	0,72
intpro	0,04	1	1	0,02	0,02	0,66	0,03	0,89	0,07	0,89	0,82	0,85	0,94	0,63	0,55
negpro	1	0,14	1	1	1	1	1	0,1	0,07	1	0,21	0,62	1	1	0,2
part	0,98	0,56	0,49	0,32	0,9	0,9	0,89	0,51	0,68	0,7	0,63	0,66	0,94	0,95	0,89
perspro	0,56	0,23	1	0,92	0,66	0,08	0,03	0,89	0,43	0,51	0,69	0,93	0,97	0,45	0,76

posspro	1	0,18	1	1	0,08	0,08	0,18	0,86	1	0,69	0,86	0,08	1	0,13	0,28
prepos	0,91	0,86	0,83	0,68	0,68	0,78	0,94	0,72	0,93	0,85	0,9	0,88	1	0,84	0,96
refpro	0,14	0,56	0,49	0,74	0,29	0,89	0,54	0,86	0,92	0,68	0,46	0,56	0,76	0,83	0,95
AVR	0,65	0,68	0,82	0,49	0,57	0,68	0,65	0,65	0,6	0,66	0,71	0,75	0,82	0,67	0,7
N09	**n09_z05**	**n09_z06**	**n09_z07**	**n09_z08**	**n09_z09**	**n09_z10**	**n09_z11**	**n09_z12**	**n09_z13**	**n09_z14_p1**	**n09_z15_p1**	**n09_z16_p1**	**n09_z14_p2**	**n09_z15_p2**	**n09_z16_p2**
aux	0,94	0,24	0,77	0,88	0,95	0,92	0,91	0,89	0,84	0,84	0,89	0,94	0,95	0,81	0,88
cconj	0,88	0,35	0,98	0,82	0,6	0,97	0,95	0,94	0,88	0,88	0,9	0,96	0,89	0,64	0,78
comadv	0,56	0,68	0,58	0,91	0,66	0,75	0,02	0,65	0,91	0,9	0,96	0,85	0,75	0,8	0,82
dempro	0,55	0,11	0,53	0,52	0,49	0,81	0,78	0,94	0,54	0,82	0,94	0,78	0,94	0,85	0,67
genpro	0,2	1	1	0,16	0,16	1	0,15	0,04	0,12	0,53	0,13	0,64	1	0,12	0,65
indepro	0,47	1	0,06	0,26	0,88	0,73	0,5	0,91	0,6	0,57	0,29	0,16	1	0,36	0,26
intj	0,11	0,11	1	0,11	1	0,11	1	1	1	0,22	1	1	1	0,12	1
intpro	0,56	0,11	0,13	0,44	0,31	0,99	0,29	0,56	0,37	0,53	0,77	0,85	0,04	0,51	0,05
negpro	0,39	1	1	1	1	0,19	1	1	0,6	0,17	0,8	0,78	1	0,06	1
part	0,86	0,09	0,87	0,93	1	0,97	0,57	0,99	0,84	0,99	0,99	0,94	0,83	0,96	0,99
perspro	0,94	0,06	0,03	0,82	0,05	0,43	0,15	0,95	0,83	0,55	0,46	0,93	0,53	0,71	0,61
posspro	1	1	0,13	0,08	0,06	1	0,08	0,1	0,62	0,53	0,32	0,53	0,1	0,06	0,64
prepos	0,89	0,28	0,84	0,96	0,83	0,72	0,98	0,85	0,74	0,62	0,87	0,96	0,91	0,88	0,76
refpro	0,11	0,69	0,13	0,44	0,78	0,99	0,91	0,98	0,82	0,69	0,94	0,83	0,65	0,72	0,64
AVR	0,6	0,48	0,58	0,6	0,63	0,76	0,59	0,77	0,69	0,63	0,73	0,8	0,76	0,54	0,7
N10	**n10_z05**	**n10_z06**	**n10_z07**	**n10_z08**	**n10_z09**	**n10_z10**	**n10_z11**	**n10_z12**	**n10_z13**	**n10_z14_p1**	**n10_z15_p1**	**n10_z16_p1**	**n10_z14_p2**	**n10_z15_p2**	**n10_z16_p2**
aux	0,51	0,02	0,56	0,02	0,62	0,41	0,69	0,87	0,87	0,53	0,93	0,93	0,81	0,91	0,72
cconj	0,86	0,94	0,94	0,94	0,83	0,94	0,93	0,9	0,88	0,75	0,86	0,96	0,6	0,62	0,95
comadv	0,87	0,91	0,41	0,82	0,84	0,99	0,59	0,43	0,79	0,55	0,37	0,64	0,28	0,97	0,46
dempro	0,7	0,04	0,13	0,72	0,63	0,94	0,98	0,76	0,75	0,92	0,83	0,82	0,67	0,71	1
genpro	0,13	1	1	0,19	0,12	0,07	1	0,05	0,15	0,07	0,26	0,84	0,12	0,24	0,33
indepro	0,33	1	1	0,47	0,45	0,21	0,98	0,98	0,08	0,87	0,52	0,52	0,07	0,83	0,91
intj	0,13	1	0,57	1	1	1	1	1	1	1	1	1	1	1	0,33
intpro	0,81	0,1	1	0,63	0,02	0,14	0,02	0,78	0,69	0,56	0,57	0,88	0,62	0,65	0,03
negpro	0,13	1	1	1	1	1	1	1	1	1	1	0,98	1	0,24	1
part	0,92	0,79	0,55	0,79	0,62	0,95	0,46	0,85	0,85	0,75	0,7	0,86	0,78	1	0,91
perspro	0,77	0,47	0,06	0,95	0,07	0,03	0,04	0,05	0,88	0,51	0,59	0,97	0,95	0,65	0,94
posspro	0,63	0,1	1	1	1	0,14	1	0,05	0,08	0,22	0,93	0,98	0,24	0,84	0,83
prepos	0,78	0,96	0,81	0,87	0,82	0,83	0,95	0,92	0,91	0,93	0,92	0,91	0,87	0,95	0,99
refpro	0,63	0,47	0,06	0,95	0,86	0,73	0,08	0,03	0,92	0,38	0,39	0,81	0,64	0,54	0,94
AVR	0,59	0,63	0,65	0,74	0,63	0,6	0,69	0,62	0,7	0,65	0,71	0,86	0,62	0,73	0,74

N11	**n11_z05**	**n11_z06**	**n11_z07**	**n11_z08**	**n11_z09**	**n11_z10**	**n11_z11**	**n11_z12**	**n11_z13**	**n11_z14_p1**	**n11_z15_p1**	**n11_z16_p1**	**n11_z14_p2**	**n11_z15_p2**	**n11_z16_p2**
aux	0,93	0,25	0,76	0,4	0,98	0,62	0,72	0,03	0,45	0,73	0,7	0,96	0,74	1	0,89
cconj	0,87	0,34	0,71	0,79	0,72	0,91	0,81	0,71	0,75	0,99	0,79	0,79	0,57	0,66	0,97
comadv	0,1	0,43	0,39	0,99	0,3	0,91	0,53	0,69	0,95	0,71	0,33	0,41	0,58	0,78	0,47
dempro	0,68	0,02	0,03	0,58	0,63	0,91	0,94	0,63	0,95	0,68	0,76	0,89	0,82	0,8	0,7
genpro	1	1	1	0,21	1	0,86	0,67	0,13	0,08	0,53	0,22	0,47	1	1	1
indepro	0,02	1	1	0,17	0,03	0,95	0,52	0,04	0,55	0,07	0,87	0,37	0,04	0,48	1
intj	0,1	0,06	0,06	1	0,16	1	0,06	1	1	1	0,22	0,26	1	0,29	0,24
intpro	0,74	0,14	0,9	0,75	0,08	0,86	0,55	0,99	0,95	0,39	0,94	0,83	0,43	0,63	0,43
negpro	1	1	1	1	1	0,64	0,08	1	1	1	0,12	0,97	1	0,14	0,82
part	0,97	0,1	0,5	0,71	0,91	0,94	0,76	0,99	0,78	0,98	0,73	0,67	0,95	0,97	0,77
perspro	0,56	0,03	0,72	1	0,19	0,07	0,81	0,02	0,91	0,55	0,94	0,83	0,73	0,86	0,88
posspro	1	1	0,2	0,05	0,16	1	0,11	0,13	0,09	0,05	1	0,26	0,07	0,09	0,94
prepos	0,65	0,13	0,94	0,62	0,96	0,88	0,98	0,5	0,85	0,78	0,92	0,94	0,57	0,84	0,86
refpro	0,93	1	0,98	0,51	0,78	0,94	0,81	0,68	0,94	0,55	0,93	0,91	0,87	0,58	0,97
AVR	0,68	0,46	0,66	0,63	0,56	0,82	0,6	0,54	0,73	0,64	0,68	0,68	0,67	0,65	0,78
N12	**n12_z05**	**n12_z06**	**n12_z07**	**n12_z08**	**n12_z09**	**n12_z10**	**n12_z11**	**n12_z12**	**n12_z13**	**n12_z14_p1**	**n12_z15_p1**	**n12_z16_p1**	**n12_z14_p2**	**n12_z15_p2**	**n12_z16_p2**
aux	0,57	0,03	0,75	0,88	0,97	0,77	0,51	0,02	0,92	0,61	0,94	0,66	0,95	0,7	0,75
cconj	0,78	0,83	0,69	0,76	0,72	0,77	0,85	0,78	0,88	0,83	0,94	0,93	0,94	0,72	0,87
comadv	0,06	0,69	0,9	0,92	0,98	0,49	0,67	0,99	0,98	0,87	0,83	0,72	0,98	0,56	0,61
dempro	0,82	0,79	0,97	0,63	0,76	0,99	0,77	0,22	0,83	0,84	0,83	0,83	0,68	0,98	0,48
genpro	1	1	1	0,1	1	0,71	0,13	0,09	0,08	0,58	0,65	0,57	1	1	0,07
indepro	0,51	0,1	1	0,08	0,17	0,34	0,96	0,67	0,91	0,81	0,84	0,76	1	0,85	0,58
intj	1	0,12	1	0,08	1	0,04	0,07	1	1	1	1	0,76	0,11	0,46	0,43
intpro	0,82	1	0,68	0,4	0,65	0,79	0,93	0,99	0,52	0,04	0,98	0,74	0,98	0,46	0,63
negpro	0,06	1	0,12	1	1	0,97	0,23	1	0,91	0,21	0,97	0,76	1	1	1
part	0,66	0,58	0,88	0,87	0,78	0,87	0,92	0,75	0,74	0,99	0,87	0,97	0,93	0,86	0,77
perspro	0,79	0,05	0,21	0,5	0,23	0,66	0,72	0,9	0,84	0,94	0,65	0,95	0,06	0,74	0,94
posspro	0,12	0,74	0,99	0,84	0,78	0,1	0,13	1	1	0,06	1	0,6	1	0,26	0,86
prepos	0,71	0,83	0,89	0,88	0,95	0,98	0,92	0,87	0,93	0,75	0,91	0,84	0,92	0,67	0,63
refpro	0,11	0,61	0,53	0,97	0,77	0,89	0,68	0,66	0,82	0,62	0,94	0,78	0,98	0,55	0,8
AVR	0,57	0,6	0,76	0,64	0,77	0,67	0,61	0,71	0,81	0,65	0,88	0,78	0,82	0,70	0,67
N13	**n13_z05**	**n13_z06**	**n13_z07**	**n13_z08**	**n13_z09**	**n13_z10**	**n13_z11**	**n13_z12**	**n13_z13**	**n13_z14_p1**	**n13_z15_p1**	**n13_z16_p1**	**n13_z14_p2**	**n13_z15_p2**	**n13_z16_p2**
aux	0,43	0,95	0,04	0,8	0,31	0,91	0,03	0,06	0,94	0,5	0,89	0,82	0,93	0,88	0,85
cconj	0,38	0,72	0,85	0,67	0,89	0,66	0,7	0,88	0,74	0,98	0,96	0,96	0,78	0,97	0,98
comadv	0,78	0,32	0,66	0,41	0,91	0,7	0,95	0,67	0,75	0,71	0,72	0,73	0,13	0,63	0,29

dempro	0,97	0,85	0,67	0,8	0,99	0,8	0,88	0,77	0,78	0,69	0,82	0,79	0,73	0,89	0,76
genpro	1	1	1	0,11	1	1	0,16	0,12	0,02	0,04	0,14	0,76	1	0,13	0,83
indepro	0,78	0,03	1	0,68	0,31	0,6	0,15	0,03	0,11	1	0,82	0,34	0,14	0,31	0,65
intj	0,78	0,03	1	0,02	0,82	0,05	1	1	1	1	0,14	0,27	0,14	0,19	1
intpro	0,23	0,95	0,04	1	0,59	0,91	0,96	0,91	0,62	0,84	0,67	0,99	0,05	0,66	0,85
negpro	0,06	1	1	1	0,1	0,12	0,16	1	0,11	1	0,14	0,76	1	0,19	0,11
part	0,95	0,88	0,92	0,78	0,78	0,75	0,94	0,84	0,84	0,99	0,89	0,98	0,93	0,89	0,9
perspro	0,84	0,95	0,04	0,42	0,82	0,6	0,16	0,35	0,06	0,4	0,72	0,42	0,08	0,9	0,72
posspro	1	1	0,08	0,06	1	0,12	1	0,59	0,04	1	0,9	0,76	1	0,91	0,96
prepos	0,74	0,85	0,99	0,64	0,73	0,85	0,97	0,95	0,61	0,93	0,76	0,97	0,84	0,87	0,73
refpro	0,78	0,38	0,5	0,04	0,55	0,35	0,96	0,99	0,94	0,07	0,58	0,94	0,04	0,88	0,85
AVR	0,69	0,71	0,63	0,53	0,7	0,6	0,64	0,65	0,54	0,73	0,65	0,75	0,56	0,66	0,75
N14	**n14_z05**	**n14_z06**	**n14_z07**	**n14_z08**	**n14_z09**	**n14_z10**	**n14_z11**	**n14_z12**	**n14_z13**	**n14_z14_p1**	**n14_z15_p1**	**n14_z16_p1**	**n14_z14_p2**	**n14_z15_p2**	**n14_z16_p2**
aux	0,88	0,97	0,9	0,01	0,86	0,48	0,85	0,96	0,92	0,52	0,71	0,49	0,88	0,88	0,89
cconj	0,99	0,86	0,81	0,97	0,94	0,57	0,97	0,87	0,86	0,76	0,9	0,99	0,74	0,89	0,78
comadv	0,82	0,51	0,48	0,98	0,82	0,91	0,95	0,69	0,56	0,64	0,67	0,51	0,43	0,85	0,15
dempro	0,77	0,97	0,9	0,98	0,9	0,58	0,29	0,94	0,94	0,66	0,73	0,73	0,92	0,82	0,87
genpro	0,83	1	1	0,23	0,11	1	1	1	1	0,48	0,86	0,35	0,62	1	1
indepro	0,9	0,04	1	0,07	0,4	0,59	0,21	0,08	0,26	0,96	0,99	0,63	0,94	0,84	0,07
intj	0,19	0,12	1	1	0,13	0,24	1	1	0,22	1	0,25	0,28	1	0,27	1
intpro	0,82	0,62	0,23	0,44	0,65	0,03	0,95	0,61	0,48	0,05	0,87	0,88	0,46	0,92	0,53
negpro	1	1	0,13	1	1	0,72	0,09	0,22	0,38	1	0,65	0,16	0,16	0,27	0,59
part	0,92	0,85	0,63	0,33	0,59	0,71	0,65	0,88	0,82	0,88	0,97	0,88	0,81	0,9	0,92
perspro	0,85	0,61	0,53	0,01	0,72	0,44	0,91	0,75	0,73	0,95	0,54	0,75	0,97	0,78	0,87
posspro	1	0,21	0,13	0,13	0,11	0,24	1	0,22	0,46	0,05	1	0,41	1	0,27	0,44
prepos	0,94	0,57	0,95	0,8	0,65	0,82	0,84	0,83	0,86	0,92	0,8	0,73	0,97	0,96	0,74
refpro	0,27	0,53	0,02	0,52	0,59	0,7	0,85	0,73	0,72	0,96	0,96	0,79	0,78	0,53	1
AVR	0,8	0,63	0,62	0,53	0,61	0,57	0,75	0,7	0,66	0,7	0,78	0,61	0,76	0,73	0,7
N15	**n15_z05**	**n15_z06**	**n15_z07**	**n15_z08**	**n15_z09**	**n15_z10**	**n15_z11**	**n15_z12**	**n15_z13**	**n15_z14_p1**	**n15_z15_p1**	**n15_z16_p1**	**n15_z14_p2**	**n15_z15_p2**	**n15_z16_p2**
aux	0,65	1	0,07	0,71	0,83	0,94	0,71	0,89	0,66	0,9	0,84	1	0,31	0,9	0,66
cconj	0,99	0,51	0,26	0,99	0,94	0,88	0,95	0,9	0,92	0,98	0,73	0,96	0,98	0,95	0,82
comadv	0,88	0,54	0,25	0,24	0,96	0,97	0,95	0,94	0,63	0,79	0,76	0,42	0,75	0,82	0,38
dempro	0,44	0,05	0,01	0,46	0,92	0,26	0,8	0,9	0,85	0,89	0,8	0,21	0,94	0,85	0,59
genpro	0,77	1	0,07	0,12	0,73	0,1	0,06	0,25	1	0,06	0,17	1	0,13	0,08	0,17
indepro	0,28	1	1	0,62	0,46	0,3	0,56	0,68	0,59	1	1	0,52	0,08	0,47	0,45
intj	0,21	0,05	0,09	0,22	0,03	0,25	1	1	1	1	0,17	1	0,21	0,21	0,28
intpro	0,65	1	0,07	0,57	0,89	0,92	0,91	0,84	0,67	0,66	0,68	0,69	0,4	0,37	0,39
negpro	0,21	1	1	1	0,14	0,25	1	1	1	1	0,17	1	0,21	0,18	0,52

part	0,99	0,51	0,45	0,9	0,9	0,66	0,66	0,94	0,94	0,79	0,78	0,47	0,92	0,91	0,96
perspro	0,93	0,95	0,41	0,13	0,28	0,97	0,6	0,93	0,67	0,79	0,86	1	0,6	0,65	0,87
posspro	0,77	1	1	0,22	0,04	0,04	0,06	0,25	0,05	0,11	0,07	0,09	0,92	1	0,34
prepos	0,98	0,95	0,9	0,88	0,99	0,74	0,83	0,89	0,8	0,84	0,91	0,8	0,88	0,87	0,59
refpro	0,76	0,85	0,99	0,7	0,96	0,57	0,81	0,58	0,48	0,88	0,46	0,68	0,84	0,82	0,98
AVR	0,68	0,74	0,47	0,55	0,65	0,56	0,71	0,79	0,73	0,76	0,6	0,7	0,58	0,65	0,57
N16	**n16_z05**	**n16_z06**	**n16_z07**	**n16_z08**	**n16_z09**	**n16_z10**	**n16_z11**	**n16_z12**	**n16_z13**	**n16_z14_p1**	**n16_z15_p1**	**n16_z16_p1**	**n16_z14_p2**	**n16_z15_p2**	**n16_z16_p2**
aux	0,99	0,67	0,78	0,84	0,98	0,86	0,53	0,8	0,94	0,99	0,56	0,93	0,88	0,99	0,78
cconj	0,89	0,94	0,8	0,98	0,93	0,75	0,96	0,95	0,84	0,91	0,81	0,97	0,89	0,71	0,71
comadv	0,46	0,41	0,52	0,72	0,47	0,98	0,97	0,48	0,89	0,96	0,53	0,74	0,67	0,72	0,47
dempro	0,65	0,03	0,76	0,93	0,95	0,85	0,63	0,81	0,97	0,75	0,55	0,66	0,67	0,75	0,84
genpro	0,96	1	1	0,1	0,27	0,22	0,24	0,27	0,71	0,97	0,19	0,16	0,2	0,18	0,08
indepro	0,88	0,07	0,11	0,67	0,17	0,47	0,97	0,76	0,54	0,39	1	0,76	0,79	0,64	0,93
intj	0,95	0,41	0,51	0,82	0,47	0,95	0,18	0,9	1	0,26	0,19	0,12	1	1	1
intpro	0,78	0,1	0,19	0,54	0,73	0,36	0,82	0,95	0,91	0,03	0,8	0,7	0,56	0,21	0,58
negpro	0,22	1	0,89	0,15	1	0,08	0,97	0,22	1	1	1	0,36	0,12	0,91	0,24
part	0,91	0,93	0,72	0,81	0,67	0,87	0,86	0,92	0,95	0,92	0,96	0,98	0,88	0,71	0,78
perspro	0,54	0,68	0,88	0,45	0,85	0,5	0,53	0,58	0,81	0,79	0,89	0,98	0,79	0,81	0,87
posspro	0,64	0,18	0,11	0,1	0,06	0,12	1	0,46	0,55	0,73	0,11	0,39	0,29	0,51	0,75
prepos	0,77	0,78	0,99	0,85	0,74	0,5	0,77	0,79	0,98	0,78	0,75	0,97	0,82	0,81	0,8
refpro	0,64	1	0,85	0,96	0,62	0,75	0,78	0,89	1	0,65	1	0,9	0,82	0,85	0,74
AVR	0,73	0,59	0,65	0,64	0,64	0,59	0,73	0,7	0,86	0,72	0,67	0,69	0,67	0,7	0,68

Table 37. Detailed results of LSM factor for all grammatical categories and for each dialogue in the Harmonia corpus separately.