UNIVERSITÀ
DEGLI STUDI
DI PADOVA

**Università degli Studi di Padova**

Dipartimento di Studi Linguistici e Letterari

Corso di Laurea Magistrale in **Linguistica**

Classe LM-39

Tesi di Laurea

# *A computational analysis of hedging in English to Polish translations of film subtitles*

Relatrice

Prof.ssa Erica Biagetti

Correlatore

Prof. Flavio Cecchini

Laureanda

Natalia Anna Matura

No. matr.2024049 / LMLIN

Anno Accademico

2022/2023

# Abstract

The thesis presents a new pragmatic annotation scheme for the phenomenon of hedging within the Universal Dependencies framework. The proposal allows for a computational analysis of hedging occurrences in English and Polish film subtitles. First and foremost, the definition of the phenomenon in question is provided, as well as an overview of its evolution within linguistic studies and its several classifications. Discussion starts with some general notions concerning corpus and computational linguistics, in particular linguistic annotation and parallel corpora, along with some references to the relevance of a computational analysis to other fields of study, such as computer-assisted translation. The second chapter introduces the Opensubtitles and the ParTy corpora and explains the selection of texts which, having been converted into CONLL-U format, are subsequently the object of a quantitative and contrastive analysis of various occurrences of hedges in English original material and its Polish translation. Thus assembled evidence constitutes the basis for the development of a pragmatic annotation scheme specific to hedges, elaborated according to the UD guidelines. Lastly, the application of the presented scheme to the chosen texts allows for a more thorough analysis and discussion of certain cases presenting the phenomenon of hedging. The thesis ends with some considerations on the value of the annotation scheme for future study.

# Abstract

La tesi presenta un nuovo schema di annotazione pragmatica per il fenomeno dell'hedging all'interno del quadro di Universal Dependencies. La proposta consente un'analisi computazionale delle occorrenze di hedging nei sottotitoli di film inglesi e polacchi. In primo luogo viene fornita la definizione del fenomeno in questione, nonché una panoramica della sua evoluzione nell'ambito degli studi linguistici e delle sue diverse classificazioni. La discussione inizia con alcune nozioni generali sui corpora e sulla linguistica computazionale, in particolare sull'annotazione linguistica e sui corpora paralleli, insieme ad alcuni riferimenti alla rilevanza dell'analisi computazionale agli altri campi di studio, come la traduzione assistita. Il secondo capitolo introduce i corpora Opensubtitles e ParTy e spiega la selezione dei testi che, convertiti in formato CONLL-U, sono successivamente oggetto di un'analisi quantitativa e contrastiva delle varie occorrenze di hedges in sottititoli originali in inglese e nella loro traduzione polacca. Le prove così raccolte costituiscono la base per lo sviluppo di uno schema di annotazione pragmatica specifico per hedges, elaborato secondo le linee guida di UD. Infine, l'applicazione dello schema presentato ai testi scelti consente un'analisi e una discussione più approfondita di alcuni casi che presentano il fenomeno di hedging. La tesi si conclude con alcune considerazioni sul valore dello schema di annotazione per studi futuri.

# TABLE OF CONTENTS

# Introduction

"*It is an old maxim of mine that when you have excluded the impossible, whatever remains, however improbable, must be the truth*."
~ Sir Arthur Conan Doyle, *The Adventure of the Beryl Coronet*

"*The truth is rarely pure and never simple. Modern life would be very tedious if it were either, and modern literature a complete impossibility!"*
~ Oscar Wilde, *The Importance of Being Earnest*

In everyday communication, barely any linguistic exchange is strictly informative. Almost every situation requires some sort of functional language able to encompass all the nuances that the participants of the exchange need to convey. The subjective truths, uncertainty, and necessity of accommodating the results of unpleasant circumstances, impel us to seek alternative linguistic strategies to achieve our communicative goals. Given how imprecise and malleable these language functions and techniques may be, for many years they remained within *the wastebasket of the study of meaning* (Lakoff 1973:477), namely pragmatics, unappealing for the serious linguists. However, along with the development of modern linguistic theories which do not exclude a priori more ambiguous phenomena, even such 'undefinable' expressions as hedges.

A hedge in linguistics is understood as a word or a phrase, as well as a communication strategy, which result in the weakening of the referent or the illocutionary force. Research studies concerning this phenomenon have been established for many years but only in the past two decades they evolved into more modern approaches, including that of computational study of language. The thesis presents a new pragmatic annotation scheme for the phenomenon of hedging which allows for a computational analysis of its occurrences in English and Polish film subtitles. The choice of comparing the data from two distinct languages, along with that of the methods, was determined by the need to elaborate an interlinguistically applicable and generally versatile instrument, which conforms to the current trends in the computational and corpus linguistics.

The exceptional success of Alan Turing and the rest of Bletchley Park team in 'breaking the Enigma', based on previous studies and parallel work of a Polish team (Velupillai.2020), attracted and kept the attention of the public, from the moment the corresponding information were released, to this day. However, it did more than that. Namely, it inspired a multitude of scientific developments in cryptology, mathematics, and, of course, linguistics. Corpus linguistics, since its crisis in the 1950s, has developed into a systematic framework for investigating language phenomena by analysing extensive collections of texts, known as

corpora. This approach allows researchers to explore authentic language use across different contexts and genres, providing a more comprehensive understanding of linguistic patterns and structures. As a consequence of extensive criticism presented by such researchers as Chomsky and Fillmore, throughout the 20th century corpus linguistics established precise rules for the creation and manipulation of corpora. Projects such as the Brown Corpus or the British National Corpus set the standards for even more extensive modern resources. This refinement granted an opportunity to employ more sophisticated tools to enhance the method of investigation on language data. Through the computer-aided examination of large corpora, researchers can identify recurring linguistic features, assess their frequency, and establish statistical trends.

The direct descendant of Turing's studies is thought to be computational linguistics. This discipline or method, on the other hand, focuses on the development and implementation of computational models and algorithms to understand and analyse natural language. From its origins in designing the Cold War cryptology machines, it managed to enter the field of non-disputable scientific study thanks to many contributions, including that of the Italian team of Roberto Busa S.J. Currently, computational linguistics combines linguistic theory with computational methods to enable the processing and manipulation of linguistic data at scale. By leveraging computational tools, researchers can explore linguistic phenomena more efficiently and extract meaningful patterns from vast amounts of textual information. There is a multitude of methods and theories employed. On the one hand, it can create confusion when exchanging the research data, on the other, it broadens the pool of potential investigation, given that some more established techniques may be less adjusted to more particular phenomena.

Corpus linguistics and computational linguistics, working in parallel, have revolutionized the study of language, enabling researchers to delve into large-scale linguistic data and extract valuable insights that were previously unattainable. This interdisciplinary field has paved the way for exploring various linguistic phenomena and their applications, including translation and pragmatic studies. An area pertaining to such research, although still quite exploratory, is that of hedging. Hedging refers to linguistic devices used to mitigate the assertiveness of speech acts, allowing speakers to express uncertainty, vagueness, or caution. It plays a crucial role in human communication, serving various pragmatic functions such as politeness, mitigating potential criticism, and managing the speaker's epistemic stance. Hedging expressions can manifest in different forms, including modal verbs (e.g., may, might), adverbs (e.g., perhaps, possibly), or lexical items that indicate doubt or speculation. The study of hedging goes beyond the realm of syntax and semantics; it intersects with pragmatics, which examines language use in context and the speaker's intentions. Pragmatics investigates how speakers adapt their

language to convey meaning effectively, considering the social, cultural, and situational factors that influence communication. By studying hedging in translation, we gain insights into the impact of pragmatics on the rendering of hedging expressions across languages and cultures, shedding light on the complex process of conveying pragmatic nuances in film subtitles.

In our study, computational linguistics, grounded in the principles of corpus linguistics, serves as the methodology to analyse the English to Polish translation of film subtitles, providing a data-driven and systematic approach to investigating hedging. Corpus linguistics forms the foundation of our investigation, facilitating an empirical examination of hedging in film subtitles. Through the investigation of the frequency, distribution, and translation patterns of hedging expressions in a parallel corpus of subtitled films, I seek to identify the challenges and strategies employed by translators when dealing with hedging and explore potential differences between the source and target languages. The main objective of the thesis, however, is the proposal of an annotation scheme for hedging within the Universal Dependencies (UD) framework. The UD framework provides a universal and cross-linguistically applicable syntactic annotation standard for a wide range of languages. The aim was to contribute to the linguistic resources available for the study of hedging in translation by incorporating a dedicated annotation scheme, hence facilitating future research in this domain, and enhancing the understanding of hedging phenomena.

The thesis opens with an overview of the broad theoretical background for the study. First and foremost, the history and evolution of the methods of corpus linguistics are discussed, followed by a corresponding presentation of computational linguistics. The general notions concerning linguistic annotation and parallel corpora are discussed in the following sections, along with some references to the relevance of a computational analysis to other fields of study, such as computer-assisted translation. Finally, the third part of the first chapter provides a definition of hedges, as well as an overview of its evolution within linguistic studies and its several classifications. Particular attention is given to the contributions of Lakoff, Fraser, Brown and Levinson, Prince, and Caffi. The second chapter introduces the ParTy corpus which inspired the selection of research data, and the Opensubtitles corpus which provided it. Subsequently, the presented selection of texts, having been converted into CONLL-U format, are the object of a quantitative and contrastive analysis of various occurrences of hedges in English original material and its Polish translation. Thus, assembled evidence constitutes the basis for the development of a pragmatic annotation scheme specific to hedges. Lastly, an analysis of the process of application of the aforementioned scheme to the chosen texts allows for a more thorough discussion of certain problematic examples of hedging devices. Through

this interdisciplinary investigation, I strive to advance our understanding of hedging in translation and its implications for effective communication across languages and cultures. The thesis ends with some considerations on the potential value of the annotation scheme for future study.

# 1 Theoretical background

This chapter introduces the theoretical background on which the study carried out in the thesis is based. The chapter is structured as follows: Section 1.1 introduces basic notions of Corpus Linguistics with its most important influences. It is subdivided into three subsections dedicated respectively to the history and definition of the methodology, guidelines for the construction of corpora, and some of the significant modern corpus-based studies. Section 1.2 concentrates on computational linguistics, namely its foundations (1.2.1), a summary of annotation practises (1.2.2), and finally the use of computational linguistics' tools in the studies on pragmatic phenomena. The last section of this chapter, section 1.3, presents the main topic of this thesis, i.e., hedges. The three subsections discuss the history of studies on hedging, the various definitions of these expressions, as well as an overview of particularities of hedges across languages.

## 1.1 Corpus linguistics

As Rühlemann (2019:1) said, "it has become somewhat fashionable in linguistics and related disciplines to assert that one's research is based on corpus." There are in fact many different definitions and theories as to what a corpus, and corpus linguistics, actually are; however, providing one extensive enough to accommodate everything and everyone is a demanding task. What most modern researchers agree on is the fact that, despite its name (Lat. *corpus* for Eng. *body*) a corpus is not simply an ample collection, or body, of texts. It is a collection of data on natural languages and their 'real life' use (Leech 2007) that needs to uphold certain standards to perform its function (cf. 1.1.2.). Moreover, an up-to-date definition of *corpus* presupposes the use of informatic tools for its consultation, as machine-readable corpora have become a standard. In the words of Wynne ed al. (2005), "a linguistic corpus is a collection of texts which have been selected and brought together so that language can be studied on the computer".

The first section of this chapter will present an overview of the history of corpus linguistics, its criticism and consecutive evolution. A broad definition will be provided subsequently, so as to introduce the topic of creation of modern corpora that will be discussed in the 1.1.2. Lastly, some remarks *apropos* the modern-day corpus-based studies will be made (1.1.3), in order to provide a general background for more detailed discussion on the use of corpora in the following section of this chapter.

### 1.1.1 An overview of history, criticism, and definition

When it comes to discussion on different fields of study, there is always a question whether a certain topic should be considered an autonomous discipline, its branch or simply a methodology of study. Indeed, prior to any further discussion, it is crucial to establish how corpus linguistics will be considered in this thesis. Positioned somehow on the verge of all of them, corpus linguistics proves to be a versatile approach adaptable to many different types of research. McEnery and Wilson (1997:2), having examined strengths and weaknesses of applying all three of these labels, decide on referring to corpus linguistics as a broad methodology of study, precisely due to its versatility. Currently, it would be easy to consider corpus linguistics as a field of study in its own right, if only on the grounds of the sheer quantity of studies and materials. Albeit not dismissing such an approach, in this thesis corpus linguistics will be roughly referred to as a methodology. In the following paragraphs, a brief background of the methodology will be provided, alongside the most influential criticisms that have shaped its development over the years.

Whether a field of study or not, the use of corpora in linguistics is not a novel phenomenon which has evolved significantly in the last few decades. Contemporary corpus linguistics, considered as such only from the onset of the use of computers (Kennedy 2001), has a long tradition, generating mostly in the structuralist framework. In fact, linguists have been using this methodology for decades (McEnery & Wilson, 1997:2-4). For instance, word indexing of the Christian Bible in the thirteenth century (Huang & Yao 2015) can in fact be considered corpus-based works. Research using some sort of corpora before the 1950s is commonly called early corpus linguistics. It is vital to distinguish between early and modern corpus linguistics, as the criticism and comments that will be discussed further, are directed at the former as opposed to the latter.

Early corpus linguists collected large quantities of data used for determining theories about languages, their development and comprehension, thus using a bottom-up methodology for establishing an empirical basis for many fields of study (Kennedy 2001). In research fields such as lexicography and acquisitional linguistics, the use of corpora has been documented since the end of the nineteenth century: see for example the pioneering works by Bennett (2010), Preyer (1889), and Stern (1924). Apart from pioneering works in language acquisition or in foreign language teaching (Thorndike 1921), corpora were also used in comparative historical linguistics (Eaton 1940) or grammar studies (Fries 1952). However, an analysis of the empirical and contrastive works cited above makes clear that, despite all their merits, the

organisation of the inquiries lacks some of the structure and scientific approach that characterises modern corpus linguistics.

The general weaknesses of the methodology that could be observed at the time and that partly persevere to this day are that corpus linguistics is not able to provide negative evidence or explain the reasons for a certain phenomenon, nor can it be applied to all languages simultaneously. The fact that corpora cannot provide negative evidence means that it cannot offer an actual analysis of what is possible or correct in a language as a whole, but rather the analysis is limited to what is observable in the corpus. The sole data present in corpora cannot explain why a phenomenon is present either. Lastly, it would be rather impossible to create a principled collection of texts extensive enough to cover all the possible phenomena of a language, which raises the question of representativeness (see 1.1.2) (Bennett 2010). Nonetheless, considering that data cannot speak for themselves as per definition (Zins, Chaim 2007), it is the scope of human analysis to explain why a given phenomenon is attested in a given corpus.

As previously mentioned, the 1950s were an important period for the development of corpus linguistics when much criticism for corpus-based studies was raised. Some argued that a limited corpus cannot describe language in its entirety, while others claimed corpora to be the only plausible method to study it (McEnery & Wilson 1997). Certainly, the most influential for the methodology as a whole during that time was the criticism put forward by Noam Chomsky, following the publication of his most influential work *Syntactic structures* (1957) and later *Aspects of the theory of syntax* (1965). Paradoxically, despite Chomky's opposition to corpora, it was precisely his criticism, along with that of Abercrombie, that prompted their evolution and determined the modern shape of corpus linguistics.

Chomsky's primary influence was that of shifting the focus of linguistic enquiry. The Chomskyan revolution and the theory of generativism started or rather exacerbated the debate between empiricists and rationalists. The core of the distinction is on the nature of the object of study. While empiricists' research was based on the observation of real, spontaneous data, usually drawn from corpora, e.g., in order to prove the (un)grammaticality of a certain expression, rationalists studied artificial data and focused on introspection. Their objective was to develop theories that could account for the natural processes of language (McEnery & Wilson 1997:4). This difference is the source of Chomsky's first criticism which states that corpus linguistics analyses the performance of language, while linguists should strive to model the competence (for the *performance* and *competence* distinction see Chomsky 1965). Although it is well-known (McEnery & Wilson 1997:5) that externalised language offers

additional information and features regarding competence that differ depending on the situation of language use, in Chomsky's view, all the valid linguistic data can be found within the competence. Hence, since corpora present a collection of elements of performance, corpora-based research not aiming at abstract description of language was deemed as futile and corpora themselves as not an accurate enough base for scientific research.

What is important to mention here, is the first problem with Chomsky's criticism, namely the issue of verifiability. The use of introspection as supported by Chomsky is problematic in that introspective judgments are private and non-verifiable on a larger scale. Although introspection may provide legitimate and useful conclusions, such procedure could be considered unfeasible in a comparable way to that of corpus-based output. Indeed, while conclusions drawn from natural data, for example recorded utterances remain available for consultation and disputing for other researchers, inductively-derived findings are private and subjective, and usually cannot be accurately proven, as it is impossible to replicate the 'experiment'. Along those lines, if corpus-based studies are skewed, then so are rationalist inquiries. Hence, "a corpus could never be the sole explicandum of natural language" (McEnery & Wilson 1997 p.8), but neither can sole introspection.

The concept of skewness introduces the second criticism expressed by Chomsky i.e., the fact that corpora are 'skewed', susceptible to various interpretations due to their limitedness:

> 'Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description [based upon it] would be no more than a mere list.' (Chomsky 1962:159)

One may say, though, that the wildest skew is not present in the corpus but in the judgement of linguists who artificially manipulate the evidence and consequently the results of their studies. Moreover, the object of study of an introspective linguist is usually quite different from the evidence presented in a corpus. A classic response of native speakers interrogated on targeted sentences with the intention of either proving or refuting a rationalist theory is: *Yes I could say that – but I never would* (McEnery& Wilson 2001: 14).

Chomsky's final criticism referred to the general approach of corpus linguists, raising the issue of their passive approach to research, which would make even the best corpora invalid. The famous parody of corpus linguist by Fillmore (1992: 35) represents the idea of Chomskyian view. According to Chomsky, although many times, for example to obtain accurate frequency data of a certain phenomenon, it is indispensable to refer to a corpus, a conscious observation and introspective judgement of a language user is often what determines the actual facts and grammaticality of a sentence (McEnery & Wilson 1997:). This criticism may be immediately

disproved by noticing the standard answer of speakers enquired about the grammaticality of an expression, being correct but not attested. Regarding quantitative data, rationalists claim that they should not be of interest to a linguist, still, they are indispensable to provide accurate observations and generalisations of the linguistic phenomena. They are also fundamental for the development of instruments for automatic processing of language, like parsers, PoS-taggers, etc.

The last and perhaps most constructive criticism was raised by Abercrombie (1965) as he compared early corpus linguistics to a pseudo-procedure, literally: he claimed that their studies merely masqueraded putting forward a linguistic investigation, while being conscious of its impracticality. It pertains to the fact that, in their initial phase, corpus-based studies were for obvious reasons constrained by human capacity of processing data as the only way for conducting this kind of labour was to perform the tasks manually, which turned out to be extremely costly and time-consuming in terms of production and highly prone to human error. Furthermore, they tended to be rather subjective and overall, less accurate and feasible than expected (McEnery & Wilson 1997). The situation changed drastically, however, with the invention of first computers - the machines that revolutionised the world in the second part of the 20th century, allowed the researchers to drastically improve their studies with automatic processes resulting in meticulous operations, including complicated calculations achievable in seconds. The introduction of the computer as an apparatus of enquiry has indeed invalidated Abercrombie's criticism, as it transformed corpus linguistics from a pseudo- to a solid procedure (McEnery & Wilson 2001: 16-17). Nowadays corpora are almost always understood as machine-readable and, owing to their ability to search for, gather, order, and calculate data, the computer has liberated corpus linguistics from the constraints of manual processing, thus making it less susceptible to errors and more economical in terms of time and costs.

The accuracy of Chomsky's and Abercrombie's criticism resulted in poor perception of the corpus-based methodology in the scientific community, as it became seemingly discredited in a short course of time by the newly developing approach. Nevertheless, such criticism was and still is highly influential, having incited the development of the methodology, in the sense of encouraging corpus linguists to create better techniques (McEnery & Wilson 1997). Having taken into consideration all these aspects, both for and against the use of corpora, it can be admitted that the initial criticism had valid fundamentals and was highly influential in the way the studies based on corpora have developed, henceforth becoming invalid. Current practices and models for corpus-based studies *par excellence* will be discussed in the second section of this chapter.

In fact, the two approaches to linguistic research, empirical and rationalist, should not be seen as opposite, but should continue to influence each other. This is perfectly stated in the final metaphor by Fillmore:

> *"My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body." (1992:35)*

As stated, corpus is not simply a searchable collection of texts, but it is well defined by various criterions such as its size, digitalisation, use of natural language and representativeness (Bieber 1993:243-257). Currently there are numerous types of corpora: generalised, specialised, parallel, comparable, and historical are just some of them. What connects them is the fact that nowadays corpus linguistics provides us with very efficient instruments for linguistic analysis, which is what makes corpus linguistics as a method or sub-discipline so versatile. They are often extensively annotated with information about grammatical classes, functions, etc. This information permits one to observe and analyse linguistic patterns, such as grammatical structures and concordances automatically, even in the case of languages that vary greatly per their historical, sociolinguistic, and register characteristics. Such annotation allows for gathering probabilistic descriptions regarding linguistic items and processes, which in turn increases the potential of other methodologies applied and brings *quantitative dimension to the description of languages* (Kennedy in 2001).

All of these considerations lead to an up-to-date definition of the corpus-based methodology. Biber, Conrad, & Reppen (1998: 4), indicate four major characteristics of modern corpus linguistics. First, it is strictly empirical, as mentioned in previous paragraphs. Corpora include elements of real-life use of language, such as fiction, magazines, and textbooks for the written language, but also transcriptions of conversations, TV shows, etc. This allows the researchers to study the patterns of actual natural languages. The second aspect of corpus linguistics is that to ensure better quality of studies it employs large collections of texts assembled through principled sampling techniques. Another important facet is that the corpus-based approach, while relying mostly on the quantitative data, avails of the human intuition for the qualitative analysis as well. Finally, this method uses computers not only for storing the data, but also for accurate and thorough analysis of phenomena in question, by means of specialised programs.

### 1.1.2 'Good practices' for building a corpus.

What may surely be said about corpus-based analysis is the fact that it is empirical and analyses patterns found in natural texts, which is especially useful since linguists tend to focus on unusual patterns thus risking a biased study (Biber, Conrad & Reppen 1998). There is plenty of information and standards regarding the creation of corpora (Wynne 2005). One of the first fundamental considerations to be made about a corpus is what it represents. Despite the tendency to include as broad a variety of texts as possible, a more productive approach is using a collection of carefully sampled material. In most cases, the material available for a study is huge and potentially infinite. Therefore, it is vital to prepare a sample of texts that would be most beneficial for a given study. The aim of preparing such a sample is to avoid skewness of a corpus. The chosen texts should be maximally representative of a variety one is working on, meaning it should provide as accurate a picture as possible of tendencies and proportions within such variety in the entire population that one wants to consider (McEnery & Wilson 1997, p. 22). The aforementioned Chomsky's criticism regarding the skewness of a corpus has much to do with the notion of frequency. Along with any other investigation involving data samples, which needs to be maximally representative to allow accurate conclusions, they also remind that corpus data is not and should not always be used for quantitative analysis.

In this section, the different criterions, and practices for the creation of corpora will be discussed. There are many chief characteristics to be taken into consideration, such as corpus generality, modality, language, etc. which vary depending on the scopes for which such corpus is being created. As for the generality, it describes the range of texts that has been chosen for the corpus with respect to the given variety of language. General corpora include texts of different varieties and registers of language. Those are multifunctional corpora which are created having in mind the scope or creating a flexible resource for the general study of language. Generalised corpora are generally large, containing millions of words, seeking to provide as much of a whole picture of a language as possible. Examples of general corpora could be The British National Corpus (BNC) and the American National Corpus (ANC), containing both written texts such as articles, fiction and nonfiction and spoken transcripts of informal conversations and official proceedings (Bennett 2010). Specialised corpora contain texts of one specific variety or theme like corpora of professional language and the collection of texts for a limited research objective of which it needs to be representative of the language of this type. Specialised corpora can be large or small and are often created to answer extremely specific questions, such as the CHILDES Corpus (MacWhinney, 1991), which contains

language used by children. Both the general and specialistic corpora can be limited to a certain type of data, which is how their modality is determined. The famous Brown Corpus contains only written texts, Childes, on the other hand, is a corpus of spoken language. There are however many multimodal corpora that could contain many modalities in various proportions, like for example the aforementioned British National Corpus, including not only texts and transcriptions, but also audio and audio-visual data.

Two rather intuitive classifications, especially in the fields of sociolinguistics, historical and comparative linguistics, refer to the language and time of production of texts. It is important to make a distinction as for the chronological aspect of the corpus: synchronic corpora include texts that belong to the same limited period of time, with the scope to analyse a determined part of the language's history. Here one can again mention the Brown Corpus which consists of texts published in 1961. Conversely, there are diachronic corpora which aim is to monitor linguistic change over time. The last, and quite interesting type are monitor corpora, including texts from the same language variety but from different periods, usually separated by regular time intervals. They facilitate the analysis of the development in the fields such as semantics. When it comes to the source language of texts included, apart from monolingual corpora, there exist numerous studies on multilingual corpora, with enormous research potential that is evident through the growing popularity of such enquiries. Parallel corpora (cf. 1.1.3) include texts in an original language aligned with the translations in other languages, for example the PROIEL project. Lastly, there are comparable corpora, which instead of focusing on aligning texts in different languages, compile data from various texts in many languages, chosen according to specific criteria.

Last characteristics of corpora which can be used for classification are those connected to their digital parameters. First, most of the modern corpora include various metadata added onto the texts through multilevel annotation which can refer to their linguistic properties, like the most common PoS-tagging, but can also contain information about the author, date of production and any other aspect that the creators deemed important. However, there are also so-called raw corpora, which consist only in the texts in the digital format. Finally, when it comes to the size of a corpus, they are mostly measured for the number of tokens or length of registrations. Given the gradual growth in the capacity of data processing, currently the corpora can be subdivided into three generations, based on their extent: the first-generation corpora are the ones created freshly after the chomskyan revolution and contain around million words; the second-generation corpora (1980-2000) could already reach hundreds of millions of words; nowadays, the corpora can contain even billions of millions of words.

There are two important approaches to the research on corpora to be considered: corpus-based and corpus-driven. In the former, the theory precedes the use of a corpus, it only needs its support in terms of evidence and quantitative data. In this case, raw data are not useful for the researcher, they need to be annotated. According to Tognini-Bonelli (2001:66) in some cases, a corpus can indicate where one may apply corrections in previously adopted models. The latter approach, corpus-driven, entails for the data to define the course of research. The theory does not exist independently to corpus and the study generates from observation to a unified theory. All the linguistic interpretations are to be applied a posteriori.

Having discussed the modalities and approaches to corpus studies, returns the issue of representativeness and other qualities that a well-constructed corpus must present. Each corpus is created in the view of a given objective and is a result of a selection process. This can be used for applicative studies or analysis of certain phenomena. One of the important questions is how much data will be necessary for obtaining the scope of research. As was previously mentioned, one of the biggest concerns regarding the corpus linguistics as a method of study is the fact that, since language is illimited and corpora can never be, it would be difficult to create one accurately representing the object of study. Keeping this in mind, three major aspects of corpus creation are described as a part of the 'good practices' of this methodology. These aspects are the corpus's representativity, balance, and sampling.

One may conduct an analysis on a simple sample of text, but generalisation of those results would hardly prove accurate to the whole population. According to Leech, a corpus is representative *if the findings based on its contents can be generalised to a larger hypothetical corpus* (1991: 27). Hence, the ideal corpus should be a model of linguistic properties of the studied population maintaining the original proportions. There is a problem of the idea of truism between language and corpus since the introduction of electronic corpora. Still, no corpus, however vast and carefully designed, can portray the very same nature of language, the sampling is inevitable. Although there is no accurate way in which representativity can be measured, only estimated, for example through the level saturation (Belica 1996: 61-74). According to Sinclair (2004) the design of a corpus should be enriched by information on the decision-making process in the selection of texts. Balanced corpus is one that, through correct preparation, embodies these qualities.

The representativity of a corpus depends on the two other aspects, namely balance, mentioned above, and sampling. Both of the operations depend in a way on the type of text in consideration. The corpus being widely considered as a standard in terms of balance is the

British National Corpus[1], which standards were mirrored in multiple projects. For a corpus to be balanced, a correct sampling frame must be applied, which for its part must be defined with the scope of representativeness and balance in mind. According to Sinclair (2004), most (general) modern corpora are unbalanced since they lack enough spoken components. In fact, the spoken component of BNC consists of ten million words which are approximately 10% of the whole corpus. The authors explain this inconsistency as a result of lack of time and funding, since gathering of the spoken data is much more costly than that of written texts. Here one should again consider the purpose which the corpus has to serve and decide whether a certain imbalance will intercept it or not (as is the case of BNC).

One of the first steps is to make a reasonable decision on the sampling frame, so clearly defining the limits of the population studied. There is a difference, however, in the approach to written, formal language, with respect to informal register. In creating a sampling frame for the latter, one must refer to age, sex, region of provenience etc. while gathering data. It is also important to talk about the texts' integrity when it comes to the corpora. According to Sinclair (2004) the texts should possibly always be included in their entirety, so as to maintain the vicinity to the target language. A researcher may decide to input the complete texts of the language in question into the collection. Still, in many cases, only portions of disponible texts are chosen. It can also help to create a balanced composition of the corpus and include a large number of texts. In smaller samples rare elements may occur out of proportion, limiting, meanwhile, the presence of more common phenomena. (Biber, D 1993

To design a well-structured corpus, the few criteria chosen should be discrete and clear. Their most important characteristics is the ability to interact successfully to delineate a corpus representative for the language in question (Sinclair 2004). Early corpora were normative, kept close to the "standard" language. Most of the largest corpora nowadays adopt a similar policy, while other corpora, are concentrated on a more specific variation, for example historical corpora so they aim to be internally contrastive. Similarly, parallel corpora, having inherently contrasting components, could be defined as contrastive. It is likely that a corpus component can be adequate for representing its variety within a large normative corpus, but inadequate to represent its variety when freestanding."

When it comes to sampling, it is a major step in the process of constructing a corpus to which various criteria must be applied. Some of them are: the mode of the text (whether text originates from speech or writing); type, domain, language, location, and date of the text. In a

[1] http://www.natcorp.ox.ac.uk/>

couple of cases, these may be already predetermined by the corpus design. They should also be quite simple to avoid or minimise the margin of error. The criteria have to be chosen with the balance and representativeness in mind, as they both depend on them. Other information about the texts may be stored for future reference (cf. 2.2.2.) so that scholars can potentially make their decisions with respect to their consecutive research. Nowadays, where the Internet allows for easy access to any published corpora and all its documentation, there is no problem with storing not only the plain text, but rich relevant information related token by token.

To construct a corpus according to these criteria, first one must consider the object of the study and the amount of data necessary to achieve the goals. This should help to determine the basic components that a corpus must contain. Afterwards, these components need to be subdivided into singular texts and finally cells focused on a given aspect to be observed. This binary representation that Sinclair (2004) proposes, allows for a simple calculation of the number of data needed for research. The minimum size of a corpus should depend on the methodology of study anticipated. The frequency of patterns and collocations one intends to study is a principal element of determining the optimal size of a corpus. When analysing statistically such data, it is important to keep in mind the natural frequencies following Zipf's Law (Zipf, 1932).

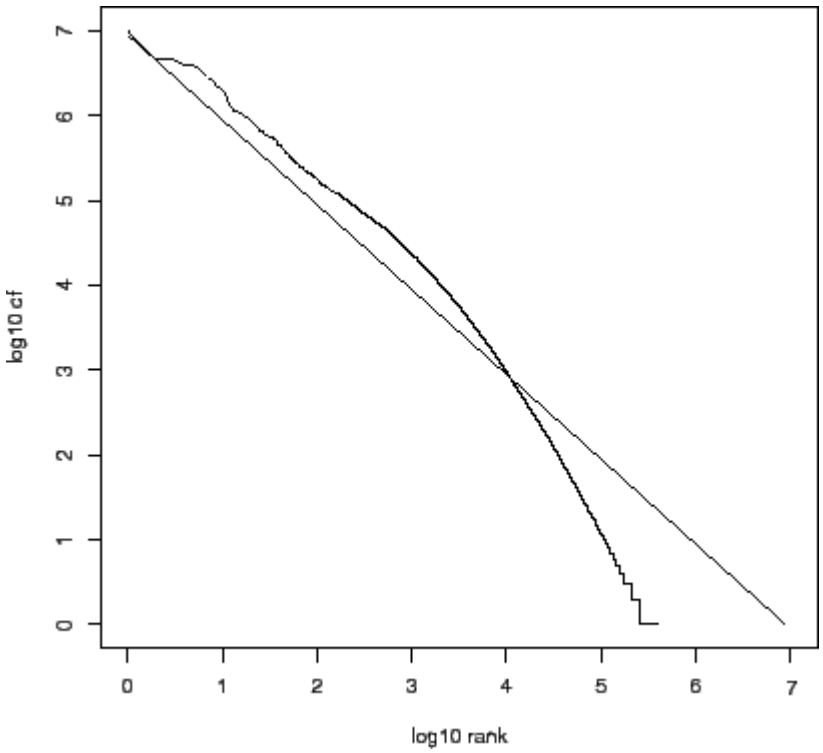



*Figure 1Representation of Zipf law applied to the frequency of occurrence of different terms as in Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.*

As the Zipf law states that the collection frequency *cfi* of the *i* most common terms is proportional to *1/i*:

$$\mathrm{cf}_i \propto \frac{1}{i}.$$

meaning that the frequency of given terms decreases rapidly with their rank, which results in the enormous amount of *hapax legomena* in any given text (Lenci x:139-141). This law can apply to the distribution of any given linguistic phenomena, thus disapproving the utility of infinite corpora (with the growth of data, the distribution reaches a point when it no longer modifies the statistical results). The more complicated the event in question, the more advanced the calculations become. Biber (1993) notices that standard statistical equations are problematic to use on a corpus, since tools such as standard deviation must be calculated for each individual feature, so he suggests basing the computation on the one most widely varying. The construction and sampling of corpus is not a trivial task, but correct application of statistical procedures should ensure its felicity.

All these good practices might not be strictly definable and depend heavily on the decisions and sensibility of the researchers. *Selection criteria that are derived from an examination of the communicative function of a text are called external criteria, and those that reflect details of the language of the text are called internal criteria. Corpora should be designed and constructed exclusively on external criteria* (Clear 1992). As discussed above, the main concept to keep in mind is representativeness, together with sampling and balance. Such practices that were already indispensable and common-sense for other scientific disciplines, are now also applied to linguistic research. There may often be a risk of bias, when a corpus is created for a certain study, since it is natural to look for data to prove what is already established, so it must also be considered not to create futile corpuses to observe qualities of texts that are already predictable. The topic of creation of corpora is much more complex and could be described in major detail, however these concepts are the most important for the considerations regarding the object of these thesis.

### 1.1.3 Modern studies and multilingual corpora

Even after the methodology has been partly discredited and before the 'machine revolution,' there were still some disciplines that were inseparable from corpora-based work, despite it being out of favour. Such disciplines as phonetics, language acquisition and historical linguistics could not possibly use introspection to infer the necessary data (McEnery & Wilson

1996). It was one of the errors of judgement that Chomsky himself admitted to in the years following his first publications. Other branches of linguistics that continued the empirical approach were quantitative sociolinguistics (Labov 1966), linguistic typology, and later functionalist and cognitive approaches based on empirical data. Thus, corpus linguistics research continued in the 1960s and 70s, although in limited circles. There is a dozen different researches worth mentioning for making notable advances in the field. One would be Quirk's Survey of English Usage (SEU), started in 1961 later digitised by Svartvik, as well as the Brown Corpus, initiated the same year, or London Lund corpus. In fact, there was rapid growth of corpus studies that were initiated already in 1965, with help of such papers as that of the Association of Computational Linguistics (1965), and paper Computer and Humanities started in 1966. As for Italy, the most important would be the studies of Father Busa in the area of Digital Humanities, and the creation of Index Thomisticus. The first computer-based corpus, so-called the Brown corpus, was created in 1961. The Brown University Standard Corpus of Present-day American English (Kučera and Francis, 1967) contains one million words from a balanced choice of texts and constitutes one of the first modern standards for corpus creation (Chu-Ren Huang, Yao Yao 2015).

The following years brought only further expansion in the field with increasing computing power and possibilities given by new generation computers. Today, corpora can reach the size of hundreds of millions of words. Contributions of corpus linguistics to various fields are immense (Bennett 2011). Due to digitalisation and annotation, which became a standard, software can be applied onto corpora to analyse and identify grammatical structure and concordances, exploring a new aspect of previous research in historical, sociolinguistic, and other kinds of research. Uniting the probabilistic and qualitative study, revolutionises modern linguistics making large steps in the development of fields such as natural language processing possible (Kennedy 2001). This proves that Fillmore's 'computer-aided armchair linguist' (1992) still applies and constitutes a base for further development of computer-based studies. Thanks to the input of such scholars as Leech, Biber, Francis, McCarthy, Sinclair and

As for the current tendencies in the corpus linguistics, what can be observed is the more significant inclusion of spoken language corpora with its transcriptions, as well as ever growing number of audio and multilingual corpora. For specific purposes there are also many specialistic corpora created, which are actually also quite easily gathered per their nature. As mentioned previously, raw corpora are rarely created, most new corpora are already enriched with some sort of metadata annotation, even minimal. The popular annotation schemes are also being adapted to more and more languages.

One of the most notable is Universal Dependencies (UD) which is an open community framework that offers annotation models for grammar in over one hundred languages. The project aims to develop a cross-linguistically consistent system of treebank annotation and provide universal inventory of categories and guidelines for research. As cited in the introduction to the project, the six elements fundamental for the design are:

> *1.… UD needs to be satisfactory on linguistic analysis grounds for individual languages.*
>
> *2.. UD needs to be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.*
>
> *3.…UD must be suitable for rapid, consistent annotation by a human annotator.*
>
> *4.. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. We refer to this as seeking a habitable design, and it leads us to favour traditional grammar notions and terminology.*
>
> *5.…UD must be suitable for computer parsing with high accuracy.*
>
> *6.. UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, …).*

The project is constantly evolving, so that the rules and features can be more universally applicable. The current guidelines can be consulted on the project's website.[2]

As it was mentioned previously, there are many different types of corpora that are being used in linguistic studies, however, they are not so easily classified since they usually share features and properties that can be attributed to different categories.[3] Talking about the number of languages that are included within a corpus, we usually subdivide them into monolingual, bi- and multilingual. A multilingual corpus, in a narrowed sense, must involve at least three languages while those involving only two languages are conventionally referred to as bilingual corpora. What mostly sparks the scientific interest now, and pertains most to the contents of this thesis, is the issue of the multilingual corpora which offer important resources for contrastive and translation studies. The main focus has fallen onto comparable and parallel corpora which offer specific uses and possibilities for these studies. In particular, they provide some insight into a comparison between more languages, including hence also different varieties than English; these comparisons allow not only to observe typological differences, but also notice the universals of language; they provide the means for observation and evaluation of translations and can be used for a variety of research goals. Before moving onto the values of these corpora, is it necessary to clarify some terminological issues. First, corpus terminology

[2] universaldependencies.org/guidelines.html
[3] www.sketchengine.eu/corpora-and-languages/corpus-types/

has taken time to settle down so that some earlier articles use the term 'parallel' for what is now called 'comparable', etc. Even now authors are prioritising different aspects in their taxonomies, one finds references to bilingual, multilingual, aligned, and comparative corpora.

Despite variety of classification systems (cf. McEnery and Xiao 2007) a parallel corpus can be generally defined as a collection of texts, each of which is translated into one or more other languages than the original. Parallel corpora can be bilingual or multilingual, what is important is that the translation direction needs to be clear within the corpus. They can be unidirectional, bi-directional, or multi-directional . Original texts are aligned to corresponding translations, usually at a sentence level. This way, the user can then search for all examples of a word or phrase in one language and the results will be displayed together with the corresponding sentences in the other language. Also, context is provided to account for equivalences between source text and target text.

In contrast, a comparable corpus is a corpus in a set of two or more monolingual corpora, typically each in a different language, built according to the same predetermined criteria. What distinguishes parallel from comparable corpora is that parallel corpora imply a common source text. This common source may be part of the corpus, or it may lie outside the corpus, as with a parallel corpus where the text pairs consist only of French and German translations of the same Dickens works. Comparable corpora may, however, bring together texts originating from different geographical areas, or drawn from diverse social varieties. The comparable texts in terms of genre/text type or topic are collected using the same sampling frame and similar balance and representativeness so the texts share various common features i.e., genre, publication date, topic. An example of comparable corpora in the CHILDES corpora. Comparable corpora are often rather small, created ad-hoc for specific tasks.

Parallel corpora are a good basis for studying how an idea in one language is conveyed in another language. They can be used for various practical purposes: contrastive analysis of features and their frequencies, systematic multilevel analysis, language-specific typological comparison, etc. all allowing for quantitative methods of analysis. *The development of corpus linguistics and the appearance of electronic parallel corpora [...] made it possible to study the usage of language specific words and expressions in translated texts* (Shmelev in Bromhead & Ye 2020). This aspect was evident even way before the introduction of digital corpora (cf. Rosetta stone). Aligning original text and translation also gives an opportunity to gain insights into the nature of translation and probabilistic machine translation systems can moreover be trained on such corpora. Many of the parallel corpora are easily accessible and offer tools for observation of concordances. Such corpora are also a rich source of materials for language

teaching (Bennett 2010). Many examples of the modern parallel corpora can be found in the CLARIN infrastructure, which contains both European and non-European language pairs, mostly sentence-aligned. One of notable parallel corpora is Europarl: European Parliament Proceedings Parallel Corpus.[4] The corpus contains the proceedings of the European Parliament in 21 European languages and its goal was to generate sentence aligned text for statistical machine translation systems.

The creators of parallel corpora must decide on several aspects of these particular corpora, for example whether to include a static or dynamic collection of texts, and entire texts or text samples. Questions of authorship, size, topic, genre, medium and style have to be considered well. In any case, a corpus needs to comply with the usual corpus requirements (cf. 1.1.2). Parallel and comparable corpora are supposed to be used for different purposes, for a comparable corpus, the sampling frame is essential. The components representing the languages involved must match with each other in terms of proportion, genre, domain, and sampling period. For a parallel corpus, the sampling frame is irrelevant, because all of the corpus components are exact translations of each other. However, this does not mean that the construction of parallel corpora is easier. For a parallel corpus to be useful, an essential step is to align the source texts and their translations which is not a trivial task. For correct alignment, one must identify the pairs or sets of elements which will be analysed, usually sentences, in the original text and their correspondences in the other languages. It is the most crucial process, since during the translation sentences might be split, deleted, or even reordered in order to create a natural translation in the target language. The degree of correspondence varies depending on the text type. The entire process is clearly very subjective, for what using solely parallel translation corpora for contrastive studies is often criticised (Johansson 2007, p. 9). Malmkjær says that simple translation only contains *one individual's introspection, albeit contextually and contextually informed* (1998). A solution could be to analyse the patterns on the basis of comparable corpora and only then study the parallel correspondences, or to include as many versions of translation as possible (Johansson 2007, p. 33).

In this first section of the thesis, an outline of the various elements of corpus linguistics was offered, so as to provide a sufficient background for the research that will be carried out further. History and criticism of corpora was considered, along with their typology, principles, and possible applications. In the following section of the thesis, an intricately connected field

[4] https://www.statmt.org/europarl/, last accessed 21.11.2022.

of computational linguistics will be presented, as the general background for the instruments involved in the preparation of the proposal in the second chapter.

## 1.2 Computational linguistics and applications of corpus-based studies

The present thesis presents a computational analysis of hedging (cf. 2.3). Such analysis consists in the application of computational techniques to a traditional analysis of data. Just as the previous section of this chapter was dedicated to outlining the origin and development of corpus linguistics, the second section is focused on presenting the field of computational linguistics, so as to understand the concepts pertaining to the study. First, the history and definition of the term will be provided, followed by some information regarding the techniques and standards of annotation of data. Finally, some insights into corpus-based computational studies on pragmatic phenomena will be presented, as a means to introduce the methodology of research that was applied for here.

### 1.2.1 Computational linguistics

According to the definition in Stanford Encyclopaedia of Philosophy[5] *computational linguistics is the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artefacts that usefully process and produce language*. The term "computational linguistics" itself was introduced by David Hays, a founding member of both the Association for Computational Linguistics (ACL) and the International Committee on Computational Linguistics (ICCL). Since the invention of the first modern computer, linguistic and informational studies have gone hand in hand. The researchers try to identify and employ what a natural language can do to increase a computer's potential and, simultaneously, how computers can help with understanding of natural language. Linguists of the field consider a computer an instrument of enquiry, offering a new perspective on language. It is considered an interdisciplinary field of inquiry, which is developing rapidly ever since its beginning.

The symbolic start of both fields is often indicated as the invention of the so-called Turing machine in 1936. Specifically devised for the computing of real numbers, there are simple abstract computational devices intended to help investigate the extent and limitations of what can be computed and are considered to be one of the foundational models of computability and (theoretical) computer science. Turing used these abstract devices to prove that there is no effective general method or procedure to solve, calculate or compute every instance of the

[5] https://plato.stanford.edu/index.html, last accessed 10.10.2022.

following problem. He solidified the concept of algorithm and applied directly to human intelligence, thus also to language capacity.

Further studies in the field had the advantage of funding from the governments for the scopes of military operations, i.e., cryptologic machines used for decryption of Enigma during World War II and ENIAC, first general electronic computer developed in Philadelphia in 1946. This and following advancements served mostly for the refinement of machine translation. Because of the capacities of computers to execute complex calculations in a brief period of time, there was hope for developing a way for complete automatic processing of language. One of those early works, by Warren Weaver and Andrew Donald Booth in 1946, were the efforts of a group of people with experience in decryption who sought to apply their knowledge base to translation seeing it merely as a complicated code (Manaris 1998). Even when technology was accessible, processing speeds on even the most advanced machines with the newest algorithms could be as long as 7 minutes for a single long sentence. Despite the tedium such limitations would inflict on researchers, they persisted, with special focus on syntax.

Around the 1950s the actual start of computational linguistics could be indicated. Following the publication of *Syntactic structures*, more formal methods were applied to linguistic studies, naturally uniting with continuously developing artificial intelligence. Sectors like the Natural Language Processing and programs they created for the analysis of syntax and interpretation of natural language led to such projects as ELIZA in 1964-66. The intense research driven by formal grammars aiming at describing and using the properties of natural language did not prove to be able to provide full insight into the linguistic complexity of natural language. While the modelling of human competence of language needed to take into consideration the system of rules and structure of symbols that a language may be defined by, it often resulted in oversimplification of linguistic theories and conception of toy models – programs that, despite being able to analyse some of the linguistic constructions, failed at computing more diverse, natural data (Lenci 2020). The initial research on models of automatic translation did not meet the expected standards. The negative 1966 ALPAC report on machine translation resulted in a turn of interest in the studies and more profound research in computational linguistics. The machine translation was not completely abandoned but focused rather on developing methods and resources to accelerate human assessment and translation.

The NLP approaches were not the only use computational linguistics had at the time. First projects also concerned the analysis of philosophical and literary texts. For example, quantitative methods were used for reconstruction of earlier forms of modern languages and other applications of socio- and historical linguistics. One of the most renowned projects, highly

respected for its influence on the development of the field in Italy is the Index Thomisticus project began by Roberto Busa S.J. It is a corpus containing the collected works of Saint Thomas Aquinas and other related authors with a total of approximately eleven million words, annotated syntactically. The whole project started in 1949, as a paper collection with manually prepared index, and continued with the goal of digitising the texts, thanks to the funding from IBM[6]. Busa's pioneering work is said to have started the tradition of computational approaches to the study of languages.

As was presented in the previous section, despite the rationalist hegemony in the middle of the last century, the empirical approaches and corpus linguistics still developed. Corpus linguistics, in fact, utilised quantitative and statistical analysis as the means to explore regularities of language emerging from natural texts. Through the merger between computational and corpus linguistics, these methods of quantitative research started to be applied to instances of everyday language. The 1980s brought growth in the NLP methods of research, adjusting the computational operations, parsing, etc. to the demands of natural language. In the next decade, more attention to improving statistical analysis was seen, as well as the development of mark-up languages and machine-learning algorithms. The later spread of the phenomenon of the Internet, introduced the biggest source of raw material in history, which enforced more extensive research on the techniques fit to confront it. Computational linguistics as a field reached its modern shape, determining its standards and principles.

Computational linguistics occupies itself with working on linguistic data to discover evidence for theories and models of language, answer their research questions and develop and test the instruments for the automatic data processing. The data can be controlled or spontaneous. The first are gathered through experiments and introspection, the latter are drawn from any type of written texts of spoken language. The data are gathered, organised, analysed, and published as linguistic resources. Those can be subdivided into actual data, resources for their processing and information or so-called "best practices" for their use.

There is no standard classification for the linguistic resources for computational linguistics yet. Nonetheless, those that are usually applied, propose a few principles necessary for the creation of accurate and effective resources. Linguistic Linked Open Data is a method that follows the standards of Linked Open Data <ref> which establishes principles of research and of reuse of data. To ensure transparency in the discipline, the FAIR principles have been proposed: findability, accessibility, interoperability, and reusability. Findability means that data

[6] International Business Machine Corporation, American technology corporation

should be accompanied by metadata which are registered in a way that allows for their identification globally. Accessibility should ensure that these data, once found, are retrievable by other users and interoperability allows for their sharing by providing a standard model of their description. Each element of the resource should have an identifiable URI, accessible via HTTP protocol. As for the methods for the actual annotation of data in the resources, these will be discussed in the following paragraphs.

### 1.2.2 Annotation

As already mentioned a few times, the work of both computational and corpus linguists follows certain regulations to best accommodate the various needs that this type of research may have and allow for a facilitated exchange of data and results. It is especially important for the bigger projects with a scope of creating linguistic models and training the automatic instruments of analysis. The need for accessibility and interoperability of these information translates to the practices for linguistic annotation.

Linguistic annotation consists in codifying the linguistic information associated with the data (Lenci 2020:211). It has a fundamental role in the field of computational linguistics because it is the component that allows for computer analysis by making the linguistic structure of the text explicit. The process of annotation is multilevel and incremental which is highly important for the interpretation and understanding of presented phenomena. The role of a linguist is not only to define the annotation scheme and define of how to apply it. The process of codifying these information is an integral part of the work.

There are four aspects that need to be taken into consideration when preparing the annotation scheme[7]. Firstly, one must think about the range and applicability of the scheme. Often a theoretical system defined a priori, turns out to be insufficient to cover all the aspects of given typology in a corpus. On the other hand, an overly specified scheme will also fail when applied to natural language data. The idea needs to be replicable as well, in the sense that the annotation can and would be consistently applied to all the phenomena in question by independent linguists. The designed scheme needs to be integrated with other levels of annotation levels, allowing to extract already existing relations between them. Finally, the proposed theory needs to be concise, so as to avoid being redundant, but expressive enough to convey the necessary information.

[7] https://users.ox.ac.uk/~martinw/dlc/chapter2.htm, last accessed 20.10.2022.

The information delivered by linguistic annotation can vary greatly. That is why we define a few different types or levels on which it can be expressed. The most common and well-formulated is the morpho-syntactic annotation. In fact, the first forms of corpora' annotation was the use of PoS taggers. (Lenci ed al. 2020:2017) Its role consists in assigning the information on grammatical category to each token of the text. This informs the readers of values such as genre and number of singular words. Moreover, the syntactic part of this annotation, conveys the relations between these different elements. The first step for executing a correct morpho-syntactic annotation is often the process of lemmatization which transforms the tokens into more general types, which simplifies further computational analysis. This multidimensional level of annotation creates the basis for improving the precision of information retrieval, since it is a *condicio sine qua non* for most other levels of encoding. Other types of annotation include that of phonetic, semantic, pragmatic, discourse, stylistic, and lexical information.

For a corpus to be useful, the annotation must be carried out according to certain standards. First of all, annotations should be separable, so that one may be able to retrieve the raw data if needed. Furthermore, just as for any other part of modern computational studies, an accurate documentation of methods and reasons for applying the annotation must be available for anyone. This should include information on whoever worked on the annotation, as well and confirmation as for the verification of the scheme applied. Each annotation scheme should provide the conventions and coding that were used in the process. It serves as an instruction for both the users of annotated corpus, and the potential collaborators. Another key point raised by Leech (2014) is that of consensuality. Given the many classifications of different, even basic linguistic phenomena, it is important to try to adhere to those more universally acknowledged, so that the annotation could be widely applicable. A group of experts gathered by the European Union, since 1990s works on the standardization of natural language processing practices as a part of EAGLES[8] initiative.

Then it comes to presentation of any annotations scheme, the encoding is extremely important. The presentation of singular tags may be more or less complicated, but the general practice is to opt for the simplest, shortest, and most transparent labels for the annotations. One of the main points if that each and every tag has to be unambiguously defined. In the past 20 years, there has been an increasingly popular tendency to standardize the annotation using standard mark-up languages which allow for encoding features of the text in a way that does

[8] http://www.ilc.cnr.it/EAGLES/home.html, last accessed 22.10.2022.

not risk losing the information during the exchange of documents. With time, this approach facilitated creation of many tools that offer support with working and interpreting texts encoded in these languages. There are some drawbacks that may be attributed to the mark up languages, such as the need for more elaborated description for each set of tags, their standard of validation that is not so readily applicable to the natural language data, however the ever-growing field of NLP studies inevitably improves them, creating more and more advanced instruments. (Leech 2014)

For one intending to provide an annotation scheme for a corpus, the aforementioned annotation manual is absolutely necessary, especially in the case of manual editing. An annotation manual should include a list of annotation devises and a specification of practices. The first includes the tagset, namely the list of symbols representing different categories of interest, along with their definition and some examples. The latter is more extensive and should include many aspects and decisions taken with reference to the proposed scheme, so any information on segmentation, parsing, and guidelines for applying the particular tags. The manual should allow the annotators to take informed arbitrary decisions, answering possibly the largest amount of potential questions in a principled manner.

To talk about the quality of linguistic annotation, Leech (*ibidem*) cites two notions. The first, being notion of realism, describes the tagsets which are well designed enough to logically connect the categories of words sharing some sort of affinity. The other, more practically measurable, is accuracy, which on its part can be subdivided into recall and precision. (van Halteren 1999: 81-86) They both refer to the results of an automatic annotation, i.e., recall describes as a measure of correct annotations within the output of a tagger, while precision defines the quantity of incorrect annotations that were rejected. Although it is possible to achieve results as high as 98% of recall for an automatic tagger, it is important that the preferred quality of annotation is always achieved with the help of human post-editing. Whether fully automatic or not, preparing an annotation is always a laborious task which is due to lack in perfection. However, the correct practices, help to improve its potentiality. Annotation can be made using any general-purpose editor manually, but for more efficient work, especially on larger corpora, it is better to use a tagger, as well as a validator – tool controlling is the annotation occurs in a way which does not undermine syntactic consistency of the text.

The level of linguistic annotation that pertains the most to the present thesis is that of pragmatic annotation. It deals with the phenomena that have to do with the communicative functions of language. (Lenci ed al. 2020:2016) This type of annotation often refers to the elements outside of a single sentence but has to do with the entire speech act. Commonly used

pragmatic annotations may use a tagset designed for attributing an illocutionary function. Pragmatic annotation may be used to improve the text with extralinguistic information that helps with the interpretation and analysis of the structure of the text.

### 1.2.3 Studies on pragmatics

The intensive development of computational and corpus linguistics ensured that currently one may find studies on an impressive number of linguistic phenomena which adhere to their methodologies. Therefore, even such elusive a field as pragmatics has a stronger and stronger representation among computational studies, although the possibility of studying pragmatic phenomena by means of computational methods has been questioned in the past. Gries (2009) contrasted corpus linguistics and pragmatics, highlighting that one is strictly quantitative and the other qualitative. As discussed previously, it is not necessarily always the case. Myers (1991)[9] observed that pragmatics and discourse analysis rely strongly on context of the situation, while corpora and especially computational studies often risk depriving the data of that contextual information. Still, there have been studies that managed to incorporate such aspects. McEnery and Wilson (1997: 98) mention the London-Lund corpus, due to its focus on conversational language which provides more opportunities to observe pragmatic elements at work. These facts may not be easily extracted by usual concordances, but still the quantitative accounts can help improve their understanding.

In fact, given the samples of natural spoken language, corpus-based studies may offer quantitative data analysis for discourse phenomena, and the expansion of these types of corpora permits to increase the number of these kinds of studies. *These quantitative approaches add to our understanding of linguistic behaviour because they can provide more specific accounts of what choices are available to the speaker in which contexts, and which of these choices are most prototypical or unusual* (McEnery & Wilson 1997, p. 99). Rühlemann (2019: 35) notices that corpus queries can help to distinguish lexico-grammatical patterns of speech acts and improve their identification and disambiguation. As for how this can be achieved, he (*ibidem*) suggests that one of the ways could be subdividing a corpus according to the speech act functions, tagging and examining how they are presented. Even though this approach is not ideal, especially because it would mostly require costly manual labour of a specialist.

[9] In McEnery & Wilson 1997: Myers, G. (1991) 'Pragmatics and corpora', talk given at Corpus Linguistics Research Group, Lancaster University.

Johansson (2007, p. 37-38) cites some corpus-based, multilingual studies that focused on discourse phenomena. Some of them are the study by Aijmer (1999) on epistemic possibility in English and Swedish or the one on discourse particle *well* in contrastive perspective by Aijmer and Simon-Vandenbergen (2003). An interesting qualitative approach to some discourse structures was presented in the paper by da Cunha, Iruskieta and Taboada (2014) through a cross-linguistic comparison. These varied inquiries prove the potential benefits of devoting the time to apply computational methods in pragmatic research. One of the topics in which it can be quite functional is hedging.

## 1.3 Hedges

The focus of this thesis is on the phenomenon of hedging which verges on the field of pragmatics. In the general understanding, it is a discourse strategy that can have various purposes but mostly modifies, or rather reduces, the truth or illocutionary force of an expression. As a pragmatic function, it concerns all levels of linguistic analysis such as morphology, syntax and semantics and is realized differently across languages (Kaltenböck ed al. 2010:3). Failures in the correct application of hedges when using a second language often results in the speaker being considered as impolite, offensive, arrogant, and, on the whole, negatively affect the felicity of communication. These consequences are usually most salient in the domains of categorisation, politeness, mitigation, and other discourse effects which will be discussed further. In this section, after a general overview of the studies on the phenomenon, some of the most influential classifications of hedging will be elaborated on. For better understanding of this unique strategy, some considerations on its presence in cross-linguistic perspective will be provided in the last section. Thus, the first chapter will be concluded with a few predictions as for the results of analysis first introduced in the second chapter of this thesis.

### 1.3.1 An Overview of the studies

The term *hedging* is relatively new in linguistics, as it has only been introduced by Lakoff in 1972 (here cited in the 1973 version). It started as a quite narrow and purely semantic concept of elements that render something more or less fuzzy. Only after the increasing interest in the topic in various linguistic backgrounds (such as speech act and politeness theory, genre-specific interrogations, and studies of vague language in general), it was attributed a wider definition. Lakoff's understanding and description of hedges is strictly connected to the prototype theory (as the topics of cognitive linguistics and categorization were also widely discussed at the time) by Rosch (1973) and is very similar to the fuzzy set theory by Zadeh (1965). In his study, Lakoff focused on logical properties of expressions such as *sort of, kind of* which served as a modifier of category boundaries. He has also paved the way for the widening of concept's understanding into pragmatics, stating that hedges operate with felicity conditions and other rules of conversation which led to further analysis of their influence on performatives.

This expansion of the concept of hedging resulted in the increased importance of applying the corpus-based approach to the studies of the phenomenon. Such step was indispensable to confront the many classification systems proposed with the actual linguistic data. Albeit initially the research on hedges concerned almost exclusively the colloquial, spoken language,

it now widened to many genres and texts, focusing also on the strong interdependence of hedging with the context and other pragmatic-based elements of communication (Kaltenböck ed al. 2010:3). Apart from the focus of study widening per material, as previously mentioned, in fact, it became apparent that the concept cannot be claimed to be purely semantic, as some of the expressions classified as hedging also modified speech acts. After Lakoff's statement, Fraser (1975) elaborated on the concept of 'hedged performative', i.e. *I **must** ask you, I **can** say*.

Shortly afterwards, Brown and Levinson (1978) showed that hedging acts not only on the propositional content but also on illocutionary force and speaker commitment in general. In their understanding, hedges indicate primarily that the speaker is not adhering to one of Grice's (1975) maxims. While illustrating these various kinds of breeching, they focus mostly on the means influencing negative politeness.

Another very important development was that of Prince ed al. (1982) who created a framework which distinguished between two types of hedges, namely those within the propositional content of the phrase and those that work on the relationship between this content and the speaker's commitment as for the truth of the proposition (*ibidem*, p. 85). They further subdivided those concerning the pragmatic aspect into 'plausibility shields' (expressing doubt) and 'attribution shields' (attributing the belief to someone other than the speaker).

These proposals will be more extensively discussed in the next section, as they were the most fundamental for the annotation scheme presented further. However, it is essential to establish that over the years there have been numerous other studies, highly influential for the general debate. One of the more recent proposals is that of Hübler (1983) who subdivides the phenomenon into understatements which modify the phrastic i.e., propositional content and hedges, modifying the neustic i.e., speaker's attitude towards his or her utterance (for other examples and an overview of criticism on the aforementioned studies see Kaltenböck et al. 2010:4-7).

One important problem that will be addressed is that hedging is achievable by semantically different mechanisms as generic expressions (placeholders), which can also play a role in approximating constructions or indefinite quantifying expressions. As said by Markkanen and Schröder (1997: 6), "*almost any linguistic item or expression can be interpreted as a hedge*". An already demanding task of providing a definition for a new phenomenon is made much more strenuous by such variety. Still, it is necessary to have some classificatory framework.

As briefly outlined, the study of hedges is a varied field. The following chart presents a comparison of the most important theories on hedging, which will be elaborated on in the following paragraphs.

| | Propositional content | Illocutionary force indication | Felicity conditions (speaker commitment, etc.) | Evidence and source |
|---|---|---|---|---|
| Lakoff (1972) | Hedge | | | |
| Fraser (1975) | | Hedged performative | | |
| Brown and Levinson (1978) | Hedge | Hedge | Hedge | |
| Prince *et al.* (1982) | Approximator (Adaptor, rounder) | | Plausibility shield | Attribution shield |
| Caffi (1999, 2001, 2007) | Bush | Hedge | Hedge | Shield |

*Figure 1Comparison of classification as per Kaltenböck et al. 2010:6*

### 1.3.2 Theories on classification

In this chapter, the theories presented by Lakoff, Fraser, Brown and Levinson, Prince et al., and Caffi will be given a broader introduction and background, so as to accurately represent the nuances in the classification. As mentioned previously, the initial idea of hedging that emerged in the 60s was closely connected to the ongoing debate on the concepts of categorization. The issue of the representation of (linguistic) knowledge raised many debates over the years. According to Lakoff (1987:7) categorisation is a fundamental human cognitive process which organises and classifies elements of the outside world, or rather their interpretations as mental categories. The phenomenon is subjected to both sensory perception and socio-cultural conditions of a person. It was during the 70s when the classic Aristotelian idea of category was dismissed. First for the Roch's (1976) prototype theory in which sufficiency and necessity conditions (Grzegorczykowa 1990, Kleiber 2003) were substituted for the idea of family resemblance (Wittgenstein 2000) that connected various elements to the prototype of a category, meaning an element considered the most central and typical. The borders of categories were considered open and allowed overlapping. This way, the affiliation

to a category was no longer a binary concept, as it became a question of degree, both on vertical and horizontal scale (cf. animal, dog, or poodle in Langacker 1993).

Lakoff's 1973 article, *Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts*, introduces and explores the concept of hedges in language and its role in expressing degrees of certainty and uncertainty. Lakoff begins by introducing the idea that language is not always precise and that speakers often need to express varying degrees of certainty or doubt. The article proves that in the natural language truth is a matter of degree. He deems it a *convenient* fiction that sentences are only to obtain values of truth, falsity, or nonsense (p. 458).He also demonstrates that fuzzy concepts have internal structure, semantics is not independent of pragmatics, and algebraic functions play a role in the semantics of hedges. Given the considerable importance of having established this particularly vague topic as an area of scientific research, Lakoff's contribution will be given a more detailed presentation.

Introducing the topic, he discusses at length some new theories on propositional logic. Grounding his thought in recent developments in the categorisation theories, he references Rosch Heider work, proving the concept of central and peripheral category membership: if people perceive categories as clearcut, they would not be able to distinguish the degrees of *birdiness*:

(4) Degree of truth (corresponding to degree of category membership)

| | |
|---|---|
| a. A robin is a bird. | (true) |
| b. A chicken is a bird. | (less true than a) |
| c. A penguin is a bird. | (less true than b) |
| d. A bat is a bird. | (false, or at least very far from true) |
| e. A cow is a bird. | (absolutely false) |

*Figure 2 Adapted from Lakoff 1973 (p. 460)*

Lakoff marks the fact that even the aforementioned 'degrees of membership' are something quite undefinable, relying on a very subjective judgement of the speakers. Having pointed out that classical set theory cannot work with more vague concepts, he analyses in depth the strengths and weaknesses of Zadeh's 1965 fuzzy set theory. Lakoff also explores the concept of "fuzzy logic" as it is a mathematical framework that allows for the representation of imprecise and uncertain concepts. Lakoff argues that hedges can be understood as linguistic manifestations of fuzzy concepts, where meanings are not binary (true/false) but exist on a continuum: an individual can be placed in a set to a certain degree ($\in$ R 0-1).

We can get a better idea of what fuzzy propositional logic is like if we look at the classical tautologies that are valid and not valid in FPL.

(5)

| *NOT VALID IN FPL* | *VALID IN FPL* |
|---|---|
| $P \vee \neg P$ | $P \rightarrow P$ |
| $P \rightarrow (Q \rightarrow P)$ | $(P \rightarrow (Q \rightarrow R)) \rightarrow ((P \rightarrow Q) \rightarrow (P \rightarrow R))$ |
| $\neg P \rightarrow (P \rightarrow Q)$ | $(\neg P \rightarrow \neg Q) \rightarrow (Q \rightarrow P)$ |
| $((P \wedge Q) \rightarrow R) \leftrightarrow (P \rightarrow (Q \rightarrow R))$ | $\neg\neg P \leftrightarrow P$ |
| $(P \rightarrow (Q \wedge \neg Q)) \rightarrow \neg P$ | $(P \wedge \neg P) \rightarrow P$ |
| $(P \wedge \neg P) \rightarrow Q$ | $((P \rightarrow Q) \wedge \neg Q) \rightarrow \neg P$ |
| $Q \rightarrow (P \vee \neg P)$ | $(P \rightarrow Q) \rightarrow ((Q \rightarrow R) \rightarrow (P \rightarrow R))$ |
| The above are true in FPL in all models in which $P$, $Q$, and $R$ are either 0 or 1. | De Morgan's Laws<br>Associative Laws<br>Distributive Law<br>Commutative Laws |

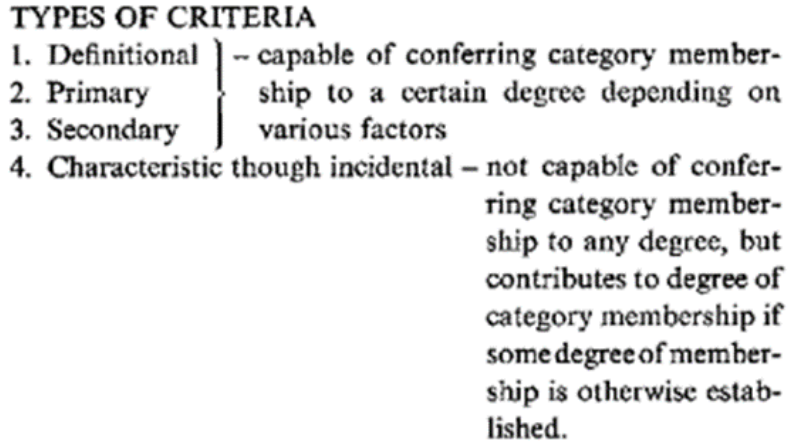
FPL reduces to ordinary propositional logic when the propositional variables are limited to the values 0 and 1.

*Figure 3 Lakoff's analysis of fuzzy set theory in view of logical equations.*

His proposal to avoid some of the problems regarding fuzzy logic, is to, instead of assigning true value to a proposition, one should assign an ordered pair of values (α, γ) for the value of true and of nonsense of the proposition respectively. This way, in simple valuations that do not adhere properly to the FPL requirements, such as $|\neg P| = 1 - |P|$, would offer more flexibility on the range between 0 and 1, where 0 would stand for the falsity of statement (since $\beta = 1 - (\alpha + \gamma)$), while in other cases the value would be positioned more freely on the truth-nonsense spectrum. Through his interpretation of fuzzy set theory shown that fuzziness can be studied seriously within formal semantics, which is quite an interesting approach that raises some questions, especially for *words whose job is to make things fuzzier or less fuzzy* which he calls hedges.

He argues that hedges serve as linguistic tools for managing these degrees of commitment and are essential for effective communication. The article discusses different types of hedges, such as adverbs ("perhaps," "maybe"), modal verbs ("might," "could"), and other lexical devices that express hesitation, speculation, or imprecision. Having investigated various modifiers in the framework of Rosh's theory, he also notes that hedges reveal more about meaning than just class distinctions. For example, *Regular*, picking up metaphorical properties of a noun, while presupposing the negation of literal meaning, makes so that 'Esther is a fish' is False, while 'Esther is a regular fish' can be True. Lakoff analyses these hedges in terms of their semantic and pragmatic implications, emphasising that hedges are not mere linguistic

fillers but carry important meaning and contribute to the overall message conveyed. This way, he places the topic, though not exclusively within the range of pragmatic research. Lakoff states, in fact, that semantics cannot be taken to be independent of pragmatics but that the two are inextricably connected. He claimed that, given the examples of hedges considered, one must distinguish at least four types of criteria for category membership:


TYPES OF CRITERIA

1. Definitional
2. Primary
3. Secondary

— capable of conferring category membership to a certain degree depending on various factors

4. Characteristic though incidental — not capable of conferring category membership to any degree, but contributes to degree of category membership if some degree of membership is otherwise established.

*Figure 4 From Lakoff 1973 (p. 477)*

Clemen (1997: 238) criticises Lakoff of basing on logical relationships in his use of words and not putting in mind the context as one of the most imperative aspects in giving hedges their outright meaning rather than viewing them as self-determining lexically. However, Lakoff does refer to a few 'inadequacies' of his proposed treatment, namely the dependence upon context, the modifiers that affect the number of criteria considered, and the problem posed by 'very' which then excluded reinforcing elements from general study on hedges. As hedges were only beginning to be studied, in the article Lakoff points out the need for a more sophisticated theory and method of research, one that would allow for the hedges to be investigated independently. He also notes that each culture categorises differently so even the elements considered primary and secondary, when talking about category membership, will be strikingly different (cf. Lakoff 1987).

In summary, Lakoff's 1973 article provides a comprehensive analysis of hedges in language, highlighting their role in expressing degrees of certainty and uncertainty. The article emphasises that hedges are not empty linguistic forms but carry significant meaning and contribute to effective communication. Lakoff's exploration of fuzzy logic and the pragmatic implications of hedges sheds light on the complex nature of language and the nuances involved in expressing meaning.

Thanks to Lakoff's contributions to establishing the hedging devices' status of elements worthy of scientific analysis, the research continued in such studies as that of Bruce Fraser from 1972 on. Moving away from the discourse around the classification theories and the questions of membership, he opened a topic which now constitutes the central aspect of hedging, namely, hedging as a rhetorical strategy which belongs to speakers' pragmatic competence. According to his definition, hedging occurs when language users rely on a linguistic device to signal *a lack of commitment to either the full semantic membership of an expression (PROPOSITIONAL HEDGING), […] or the full commitment to the force of the speech act being conveyed (SPEECH ACT HEDGING)* (Fraser in Kaltenböck 2010: 22).

Linguistic studies of Bruce Fraser have contributed to the discussion by shifting some of the focus to the expressions called hedged performatives. Hedged performatives are speech acts that contain mitigating or hedging expressions, which modify the illocutionary force of the utterance (Fraser 1975). To achieve it, the hedging expressions accompany performatives (such as English modal verbs) and act on different aspects:

1. *May I ask if you're married?*
2. *I must warn you not to discuss this in public.*
3. *I must request that you sit down.*
4. *Take the books off the table, if you can manage it.*
5. *I hope the boat has already sailed.*

In the case of (1) and (2), the illocutionary force of the expression is being modalized by verbs may and must. The second one also presents an example of mitigation with the scope of saving the speaker's face. Looking at the example (3), we even consider the felicity condition on requesting, being that the hearer is able to carry out the act. Sentence (4) could be interpreted differently in some contexts, but generally it presents another example of linguistic behaviour verging on negative politeness. Example (5) touches upon the maxim of quality where if moderated the full responsibility for the truth of proposition.

Fraser proposed a classification system for hedged performatives, categorising them into three types: hedged assertions, hortatory hedged performatives, and conditional hedged performatives. Hedged assertions involve the use of mitigating expressions to weaken the strength of the assertion. For example, phrases like "I think," "It seems," or "Perhaps" indicate a level of uncertainty or doubt. Fraser's research highlights the pragmatic functions of these hedging devices and their impact on the illocutionary force of the speech act. Hortatory hedged performatives are speech acts that express suggestions or recommendations with a degree of caution or modesty. They often employ hedging markers like "might," "may," or "should,"

which soften the imperative force of the utterance. Fraser explores the pragmatic implications of these hedged performatives and their role in persuasive communication. Conditional hedged performatives involve the use of conditional statements to attenuate the strength of the speech act. Expressions such as "If you could," "If possible," or "Would it be okay if..." introduce conditions that make the illocutionary force less direct. Fraser's work examines the interaction between conditional constructions and the mitigating effect on the speech act.

Above all, Fraser's research emphasises the pragmatic functions of hedged performatives. He argues that hedges, as linguistic devices, serve to manage interpersonal relationships, politeness, and social interaction. By employing hedging expressions, speakers can convey their intentions with greater flexibility, reduce potential face-threatening acts, and negotiate meaning in various contexts. Additionally, Fraser investigates how hedging devices can be employed strategically to manage face-saving or politeness strategies in different communicative situations. In fact, in a 2010 article (*Pragmatic competence: The case of hedging*), following a detailed overview of different linguistic means to express hedging in English, he discusses it in view of various discourse phenomena, such as vagueness and evasion. Drawing from different sources, Fraser illustrates the overall complexity of the phenomenon and highlights its relevance to issues of second-language teaching, cross-linguistic comparative politeness, equivalency of translation, and so on.

As previously mentioned, in his many works Fraser explored epistemic modality and various other pragmatic aspects of hedging devices, alluding to their influence on such discourse effects as politeness. This topic was more profoundly discussed by other prominent linguists, Brown and Levinson. Brown and Levinson's studies on hedges in linguistics are part of their broader research on politeness theory. In their seminal work, "Politeness: Some Universals in Language Usage" (1978), Brown and Levinson introduced the concept of hedges as one of the strategies used by speakers to mitigate potential face-threatening acts in communication.

Brown and Levinson (1978) were concerned with two types of politeness: positive politeness and negative politeness. Positive politeness can be understood as a compensatory action directed to the addressee’s positive face (individual's desire to be admired, ratified, and related to positively), It emphasises the speaker's connection and similarity with the addressee. Positive politeness seeks to establish rapport and build a positive relationship by stressing the similarities and highlighting the addressee's positive qualities. It can be expressed by the means of politeness markers, such as softening words and phrases or employing honorifics. Negative Politeness, on the other hand, is addressed to the addressee’s negative face (the want to have

his/her freedom unimpeded), where the speaker applies different strategies to avoid or mitigate the feeling of imposition and the potential threat to the addressee's freedom or self-image. It consists in partially satisfying their need by weakening a challenge to the negative face. Both positive and negative politeness strategies are used in different social contexts and depend on cultural norms, individual preferences, and the nature of the relationship between interlocutors. Brown and Levinson focused primarily on negative politeness strategies that include hedging the illocutionary force of an utterance; hedging any of the felicity conditions on the speech act; or hedging any of the four Gricean maxims.

Hedges are linguistic devices that allow speakers to express uncertainty, ambiguity, or lack of commitment in their utterances. That means they can serve to soften the impact of potentially offensive or threatening statements and can help to maintain harmonious social interactions while still achieving the scope of the communicative act. According to Brown and Levinson, hedges serve multiple functions in discourse. Firstly, they allow speakers to signal their own lack of knowledge or certainty, thereby avoiding making absolute or overly authoritative claims. For example, using phrases like "I think," "I believe," or "It seems to me" can indicate that the speaker is presenting their opinion rather than an indisputable fact. Secondly, hedges can function as polite strategies to lessen the potential imposition on the addressee. By softening the force of an assertion, speakers provide the listener with an opportunity to disagree or offer an alternative perspective without losing face. Furthermore, hedges can also help to minimise the impact of criticism, making it more acceptable and less threatening. For example, saying "I have a slight concern" instead of "I strongly disagree" can reduce the directness and potential offence.

Brown and Levinson's work on hedges highlights the importance of linguistic devices in managing interpersonal relationships and social dynamics. They proved that hedges provide the speakers with means to navigate the complex balance between being informative, maintaining politeness, and managing face-saving concerns. Still, no proper classification was proposed that would help distinguish between different hedging strategies. More precisely, Brown and Levinson did introduce the terms for approximators and shields, which will be discussed in the next paragraphs, but their contribution focused on the different expressions pertaining to politeness strategies.

As mentioned, Penelope Brown identified two categories of linguistic devices used to convey imprecision or vagueness: approximators (minimising the face threatening act while still conveying it) and shields (more elaborate strategies like disclaimers and apologies, used to downplay the imposition). However, it was Prince’s contribution (1982) that established those

two as distinct categories of hedging expressions, not only limited to politeness. What is worth noticing, in her work, she shifted some of the focus back onto the actual linguistic expressions conveying the role of hedges. She explored how hedges are used to express uncertainty, vagueness, or speaker reservation in different communicative contexts and shed some more light on the pragmatic aspects of hedging.

Apart from investigating different pragmatic functions of hedges, some of the main points of Prince's research covered the linguistic forms of hedges. She indicated and elaborated on a few different types of expressions that may be categorised as hedging devices. The most important were modal adverbs - words like *perhaps*, *possibly*, or *maybe* that indicate possibility or uncertainty; modal verbs – such as *could*, *might*, *would* that express tentative or conditional meaning; adjectives and adverbs of degree – words like *somewhat*, *fairly*, *rather* that indicate a moderate quality; lexical verbs – those which convey tentative meaning, like *seem*, *tend to*, or *appear*.

When it comes to the distinction between approximators and shields, as hedging devices, Prince agreed that both can serve as politeness and mitigation strategies to manage interpersonal relationships and minimise potential threats to face. She expanded those two types of hedges with more precise examples. According to Prince, approximators are linguistic devices used to express imprecision or approximation. They are employed when speakers wish to soften the impact of their utterances by intentionally being less specific or precise. Examples of approximators include words like *about*, *around*, *roughly*, *kind of*, and *sort of*. For instance, saying "It's approximately 5 o'clock" instead of "It's exactly 5 o'clock" conveys a degree of imprecision or approximation. Shields, on the other hand, are linguistic devices used to shield the speaker from potential negative reactions or implications of their statements. They provide a layer of protection by distancing the speaker from the directness or forcefulness of the statement. Shields are typically employed when speakers anticipate that their words might be perceived as impolite, offensive, or confrontational. Examples of shields include phrases like *I'm just saying*, *I'm no expert, but*, *Don't quote me on this*, or *Just a thought*. These expressions help to downplay the speaker's authority or responsibility for the statement and mitigate potential face-threatening acts.

The theory that was especially influential for the classification applied for the proposal presented later in the thesis was that of Caffi. Prince's framework on approximators and shields primarily focused on the syntactic and semantic aspects of linguistic devices used to convey imprecision or vagueness in language. It aimed to understand how speakers employ these devices to soften the impact of their statements and manage politeness. On the other hand,

Caffi's approach to hedging extends beyond the syntactic and semantic dimensions and includes an even broader analysis of pragmatic functions and communicative intentions, accompanied by another proposal of hedges' classification. The main point of her research, however, concentrates on mitigation, to which I already alluded in the previous paragraphs.

Fraser (1980) said:

> *"I will begin by saying what mitigation is not: it is not a type of speech act. To mitigate is not to perform some particular illocutionary act such as requesting, promising, or apologizing. Nor is it to perform a so-called perlocutionary act […] such as annoying, surprising or persuading."*

Mitigation is a linguistic phenomenon which aims to reduce the harshness or hostility of the force of a speech act. One of the mitigation strategies is just hedging, which can cause a statement to be less assertive or categorical. It is not the same as politeness, as it actually reduces the unwelcome effect of what is being said, while politeness is more connected to the appropriateness of linguistic behaviour in order not to achieve certain effects (by not violating the rights and obligations in effect). Thence, mitigation and politeness can coexist, ex. *I'd appreciate it if you would sit down*, but are not homonymous: *Please, sit down* is polite, but not mitigating. It means that a mitigating speaker can be perceived as impolite, and conversely, a non-mitigating, direct speaker can be perceived as exquisitely polite. Fraser (*ibidem*) actually distinguishes between self-serving and altruistic politeness. The first considers the effects of the communicative act on the recipient and aims to mitigate potential unwelcome or hostile reactions towards the speaker (me). The second type, altruistic mitigation, the objective is to soften the potential unwelcome effects that the communication may have on the hearer.

As for the Caffi's contributions (1999, 2007), she emphasises the interactional and interpersonal aspects of hedging and examines it as a pragmatic strategy used by speakers to manage face and interpersonal relationships in discourse. Caffi's analysis focuses on how hedging allows speakers to mitigate the potential threat to face, maintain a cooperative and non-confrontational stance, and promote harmonious interaction. While both Prince and Caffi explore the role of hedging in language use, their approaches differ in terms of the specific dimensions they emphasise. Prince's framework centres on the syntactic and semantic properties of approximators and shields, whereas Caffi's analysis encompasses a broader understanding of hedging as a pragmatic strategy.

In her work, the distinction is made between three types of mitigation (which can be further applied to hedging expressions):

- bushes - expressions which focalise the mitigation on the level of proposition[10]; the downgrading operates on the parameter of precision (so utilises markers of vagueness and approximation); *There were about 10 people there;*
- hedges - describing the examples where mitigation focuses on illocution; *I think you should..;*
- shields - expressions which mitigate the unwelcome effects of the speech act by focusing on its deictic origin; *He believes that…* .

Caffi's contribution is grounded in the assumption that the study of hedging operations can benefit from the analysis of language usage as it developed through the centuries in the field of rhetoric. She presents an in-depth analysis of the use of the approximation marker showing that it can be both attenuating and reinforcing. In the broader context of this thesis, it is worth pointing out that although reinforcing elements were initially included in the studies of hedging and related phenomena (see Lakoff 1973), currently it is widely considered a separate linguistic device. Fraser (in Kaltenböck 2010: 22) believes it to be stemmed from the general understanding of the concept as asymmetrical - the positive interpretation of hedging *seems counterintuitive*.

To summarise, among multiple other studies pertaining to the development of the concept of hedges, those that seemed best applicable to the present thesis, namely the contributions of Lakoff, Fraser, Brown and Levinson, Prince, and finally, Caffi, have been presented above. Lakoff introduced the concept and the name 'hedge' in the 1972 article in which the concept was investigated in the context of linguistic vagueness. She argued that hedges serve to mitigate the precision of an utterance, allowing speakers to express degrees of uncertainty or imprecision. Fraser expanded on Lakoff's work and proposed the concept of hedged performatives. The aforementioned paper examined how speakers use hedging to soften the force of speech acts, such as requests or promises, making them less assertive and more polite. Building on Lakoff and Fraser's work, Brown and Levinson developed a comprehensive theory of politeness in their book "Politeness: Some Universals in Language Usage" (1978). They incorporated the notion of hedges as one of the politeness strategies used by speakers to mitigate potential face-threatening acts. Brown and Levinson's framework identified hedges as devices to express uncertainty, imprecision, or to downplay the force of an assertion. Prince expanded on the concept of hedges, introducing the distinction between approximators and shields.

[10] In Caffi's approach, there is no space for a sharp distinction between approximators belonging to semantics and shields to pragmatics. For her, the operation of bushes on the propositional content has repercussions on the whole speech act. Hedges which would work on the propositional level by the means of accentuating vagueness or approximation, at the same time can weaken the commitment to the truth of proposition and swiftly encompass all the pragmatic aspects seen above.

Prince's focus was on the syntactic and semantic aspects of approximators and shields, examining their role in expressing imprecision and shielding the speaker from potential negative implications. Finally, Caffi further developed the understanding of hedging as a pragmatic strategy in politeness and interpersonal communication and emphasised the interactional and interpersonal functions of hedging. Caffi explored how hedging allows speakers to manage face, maintain cooperation, and navigate delicate social interactions.

Throughout this development, the concept of hedging evolved from simple expressions of linguistic vagueness and imprecision (Lakoff) to encompassing politeness and face-saving strategies (Fraser, Brown, and Levinson). Prince's contribution refined the understanding of hedges through the categorization of approximators and shields, while Caffi further explored the pragmatic dimensions of hedging in interpersonal communication. Just as in any other discipline, it's important to note that scholars, working on the topic of hedges in parallel, often built upon and contributed to each other's work. That is why, within the previous sections many notions and ideas overlapped. Nonetheless, the different researchers' theories on hedging continued to introduce new concepts and classifications, even if only slightly modified with respect to the previous one. Moreover, the phenomenon is still being interrogated. Understanding these multiple perspectives and frameworks provides a more comprehensive background for the following study.

### 1.3.3 Types of hedging expressions

Hedges can encompass various linguistic expressions that serve to soften the impact of an utterance or convey uncertainty. As can be seen from the few examples taken from the articles cited in the previous section, they can range between different parts of speech and expand from single word expressions: *like, seemingly;* through propositions, to entire speech acts. Fraser (in Kaltenböck 2010) refers to them as an open functional class. While the specific list of expressions can vary depending on context and individual usage, some common examples of linguistic devices that can constitute hedges are:

- Modal verbs - verbs such as *might, could, may*, and *would* are often used to express possibility, likelihood, or uncertainty. They can indicate that the speaker is not making a definitive or absolute statement;
- Adverbs and adverbial phrases - those which convey vagueness, imprecision, or approximation can function as hedges, like: *approximately, roughly, sort of, kind of, to some extent*, or *in a way*;
- Qualifiers - words that modify the degree or extent of another word which can indicate a level of uncertainty or imprecision: *almost, nearly, partly, quite*, or *a little*;

- Indefinite pronouns – pronouns such as *some, several, a few*, or *many* can sometimes be used to convey imprecision and avoid precise quantification;
- Phrases – for instance, *I think, I believe, it seems to me, as far as I know, if I'm not mistaken*, or *in my opinion* can signal that the speaker is presenting their viewpoint rather than an indisputable fact;
- Non-committal language - non-committal language or expressions that avoid making strong assertions can serve as a hedge. Examples include *I'm not sure, I'm not certain, I can't say for certain*, or *I'm not an expert, but*;

Fraser (in Kaltenböck 2010:23-24), compiles an even more extensive list – according to him, elements which can be classified as hedges include:

a) Adverbs/Adjectives - *He looks **sort of** sick*;
b) Impersonal pronouns - ***One** can imagine that ...*;
c) Concessive conjunctions -.***Even though** you dislike the beach, it's worth going for the view;*
d) Hedged performative - *I **must** ask you to sit down*;
e) Indirect Speech Acts - *Could you speak a little louder?*;
f) Introductory phrases - ***I believe** that he should go, if possible;*
g) Modal adverbs - *I can **possibly** do that*;
h) Modal adjectives - ***It is possible** that ...;*
i) Modal noun - ***The assumption** here is that . . .* ;
j) Modal verbs - *John **might** leave now*;
k) Epistemic verbs - *It **seems** that ...* ;
l) Negative question convey positive hedged assertion.- *Didn't Harry leave?* [I think Harry left];
m) Reversal tag - *He's coming, **isn't he**?* [I think he's coming]
n) Agentless Passive - *Many of the troops were injured*.;
o) Conditional subordinators - ***Unless** the strike has been called off, there will be no trains tomorrow.*;
p) Progressive form - *I **am hoping** you will come.*;
q) Tentative Inference - *The mountains **should be** visible from here.*;
r) Conditional clause - ***If you're going my way**, I need a lift back.*;
s) Metalinguistic comment - *strictly speaking, so to say, exactly, almost, just about*;

It's important to note that the presence of these expressions does not automatically make a statement a hedge. The interpretation and function of these expressions depend on the context, intonation, and the speaker's intention. Hedges are flexible and can have multiple interpretations and their meaning can vary depending on the context. The same linguistic form can function as a hedge in one context and as a different pragmatic device in another. For example, the phrase *I guess* can function as a hedge to express uncertainty or as a pragmatic marker to soften a request. That is why the classification of hedges may not capture the full range of their potential functions and meanings. Hedges can also exhibit overlapping features and can be multifunctional. Some linguistic devices may function as both hedges and other

pragmatic devices simultaneously a hedge can function as a mitigating device and express politeness. This overlapping nature makes it difficult to assign hedges to discrete categories.

Further, if we take into consideration that language use is dynamic, new hedges can emerge continuously and the usage of existing hedges can evolve over time. It can also vary significantly among individuals based on factors such as personality, social status, and communication style. Different speakers may employ hedges in different ways, making it challenging to create a rigid classification that applies uniformly to all speakers. Hedges can, and most certainly do, vary across different languages and cultures as well. What may be considered a hedge in one language or culture might not have an exact equivalent in another. Additionally, the usage and interpretation of hedges can vary within the same language community, making it challenging to create a comprehensive and universally applicable classification system.

Given these challenges, it is important to approach the classification of hedges with caution and recognize that it serves as a framework for understanding general patterns rather than an exhaustive and universally applicable categorization. It is crucial to consider the specific context, culture, and individual speaker when analysing the use and function of hedges in communication.

### 1.3.4 The functions and effects of hedges

Hedges can produce various pragmatic effects in communication. These effects are closely tied to the social dynamics, politeness strategies, and interpersonal relationships involved in the interaction. What is worth noticing, the effect of hedging lays in the interpretation of the speech act and not in its semantic meaning. Thence, it depends on the speaker, the context of utterance, the hedge used, and on the hearer.

Hedging, being a pragmatic/discourse strategy in itself, can sometimes invoke other discourse effects, focalised on the pragmatic scope of the utterance, some of which were discussed by Fraser (in Kaltenböck 2010). The most obvious one is vagueness, which, as a perlocutionary effect, describes a situation in which a speech act lacks expected precision. It can be correlated with politeness (not to offend), lack of knowledge, or other, more elaborated strategies. Vagueness can also stem from much simpler acts of skipping precise denomination (common occurrence in everyday language) or the fuzzy area of non-Aristotelian categorisation - within the scope of a proposition, vagueness usually acts as regular attenuation, as in *sort of* expressions. It is the speech act hedging which that is more focused on illocutionary force and can strive to convey all the other effects. However, in this case, it does not often maintain the

vagueness of the expression, ex. *It appears that we should go*. Another common discourse strategy which may involve hedging is evasion. It occurs then information received fails to meet expectations in the mind of the recipient, ex. *I think you'd better ask you mother that question*. Some vagueness can result in evasion, some can produce equivocation. The latter constitutes the use of ambiguous words in order to mislead the interlocutor. Equivocation however, rarely stems from hedging.

According to Clemen (1997:47), apart from expressing or accommodating doubt, hedges perform the function of caution, engagement, requesting, or reproving. Use of hedges serves to avoid making unnecessary conclusions and assertions thereby providing a solution for potential misunderstandings and other failures of the communicative act. Markkanen and Schroder (1989:89) attribute it mostly to hedges' function of softening and mitigating of face-threatening acts. By introducing uncertainty or imprecision, hedges can make statements less direct, forceful, or confrontational, especially when it comes to complaints or suggestions of the speaker. Through the employment of hedges, the speaker creates a mutual understanding and maintains a cooperative and non-confrontational tone in the communication. Hedges allow speakers to express disagreement, criticism, or uncertain opinions in a more tactful manner.

Furthermore, the use of hedges can make the speaker's communication approach more or less specific – speaker may employ hedging devices to appeal to the listeners and to hide his/her true feelings, or opinions. Hedges can also convey modesty and tentativeness by suggesting that the speaker does not wish to assert strong certainty or authority. They can signal that the speaker is open to alternative viewpoints, willing to consider other perspectives, or acknowledging their own limited knowledge or expertise. By using expressions that convey imprecision or approximation, speakers can leave room for negotiation or modification of their statements. Hedges can create a more tentative and open-ended conversation, facilitating cooperation and compromise.

It's important to note that the pragmatic effects of hedges can vary depending on the specific context, culture, and interpersonal relationships involved. The interpretation and effectiveness of hedges rely on the shared knowledge and expectations of the participants in the communication. Moreover, all these effects can be achieved by utilizing various hedging expressions but are not specific to hedges. They can be produced by multiple other linguistic devices.

### 1.3.5 The cross-linguistic variety and problems with translation

What many authors observed is that speakers tend to use hedges rather unconsciously. Usually, they are interpreted correctly by the interlocutors, but more often the reception remains unconscious, unless the incorrect use of these expressions broke some of the norms (like politeness standards) inherent to the language. As noted in the previous paragraphs, hedging devices vary greatly within a single language and may therefore cause significant difficulties for non-native speakers. In the same manner, they tend to complicate any interlinguistic studies and inquiries, including translation.

Meyer and Pawlack (in Kaltenböck 2010) in their study on propositional hedges (*bushes* in Caff's understanding) analysed a corpus of Brazilian Portuguese to German spoken translations of some expert talks. While analysing the source–target correspondence, they discovered that unprompted mitigating expression were added to the target text by all the interpreters (*ibidem*: 73). The initial thought was that the inconsistencies between the source and the target were caused by some common factors such as linguistic constraints, preferences of the target audience and differences in knowledge. Functional equivalence is often quite unachievable so hedging as a strategy could be a way to enhance those chances and minimize the deviation from the original, even though technically it would not be correct translation strategy since similar subjective changes modify the propositional content of the information provided. In Meyer and Pawlack study, it turned out that this tendency to vagueness corresponded mostly to unfamiliar names and uncommon, direct thoughts. Given that both the interpreter and the potential audience may not share the same cultural and epistemological basis as that of the original speaker, the instinct to adapt the message given in a foreign language to the standards of the speakers by adopting vagueness seemed validated (Meyer and Pawlack in Kaltenböck 2010:90).

Adamczyk (2015) goes into even more detail when it comes to the practical use of a certain hedge by focusing to the use of the word *jakby* (PART, *resembling*) in Polish spoken language. According to Wierzbicka (1991), who is one of the few Polish linguists investigating the topics connected to hedging, *Polish tends to overstate rather than understate*, however, the use of *jakby* in the sense of *something similar* while being only one of the canonical used of the word, corresponded to most of its presence in Adamczyk's corpus. Simultaneously, it aligns with the approximating and mitigating functions that the literature expects of hedges. She noted that in this particular spoken corpus, the more pragmatic markers were used by the speakers, the more the use of *jakby* deviated from the canonical interpretation, or at least made it more

difficult to be interpreted in that way. Nonetheless, in this exploratory study, Adamczyk notices that even non-standard uses of *jakby* served vital pragmatic roles, fitting the category of a pragmatic marker.

As seen in the aforementioned studies, as well as in plenty of others, hedging itself is a particularly ambiguous phenomenon which causes a lot of doubt even within a single language. These peculiarities are radically enhanced when other cultures and languages have to be considered. Moreover, Markkanen and Schröder (1989) point out that hedges are realised by different means in different languages, also because the speakers have individual stylistic preferences when it comes to the use of such strategies. The generally approved and the shortest definition of a good translation is that the translated text is equivalent to the original – it has to express the same semantic value but also convey the same functions. That is why cultural variation in pragmatic strategies demand a particular sensitivity from the translators. While in Markkanen and Schröder's study multilingual translators could modify their own texts, in regular translation knowing the intentions of the writer is not a foregone conclusion. Translators and second language users may also be influenced by their native language's hedging conventions and transfer those patterns to the target language which can result in pragmatic errors or misinterpretations. The pragmatic transfer can affect the appropriateness and effectiveness of communication in the target language context. Second language users may face challenges in acquiring and employing appropriate hedging strategies. Hedges are a complex aspect of pragmatics that require a deep understanding of the target language's sociocultural norms and communicative expectations. Lack of proficiency or awareness of hedging strategies can lead to either overuse or underuse of hedges, impacting the pragmatic effectiveness of communication. Hedges are highly context-dependent and can convey implicatures beyond their literal meanings. Translating hedges requires an understanding of the broader communicative context, speaker intentions, and cultural norms. To overcome these challenges, translators and second language users need to develop a strong understanding of pragmatics, including the cultural and contextual aspects of hedges. A lack of comprehension of hedging devices in the source text could result in the translator's own opinion, rather than the author's, shining through in the target text. Immersion in the target language and culture, exposure to authentic materials, and practice in using and interpreting hedges can help enhance proficiency in dealing with the complexities of hedging in translation and second language use.

According to Peterlin and Moe (2016), hedging constituted another piece of a complex puzzle of translation. Similarly, to what was just pointed out, in their work, Peterlin and Moe notice that if the translator does not render correctly what the writer wanted to convey through

their choice of words, it will undoubtedly alter the interpretation of the translated text. Through the analysis of English and Slovene translations of news discourse (*ibidem*:7-8), they managed to identify three translation strategies that can be applied to manage the hedging expressions:

- retention – when unmodified grammatical equivalent is used in the translated text; the structure of the sentence may vary but a given hedge maintains the same grammatical category in both sentences;
- omission – happens when the hedge is not transferred to the target text; it is worth mentioning that in case of omission the message of the translated text will almost certainly be somewhat altered;
- modification – when the hedging device is kept when it comes to its intent and function but it does not maintain the same grammatical category as in the original text.

Kjellström (2019), when applying Peterlin and Moe's theory to her own study, based on the observations, expanded the list by a fourth category, namely:

- addition – happens when a hedging device is added to the target text, even though it does not appear in the original.

Those strategies are useful when considering any type of research confronting the hedging strategies in different languages. Henceforth, they were used as one of the criterions for the observation on corpus data, which will be introduced in the following chapter.

# 2 The Method of study

In the first part of this thesis, I discussed the topics of corpus linguistics, with particular attention to parallel and comparable corpora, computational linguistics with its methodology and, most importantly, the phenomenon of hedging. Having established the notions to which this dissertation is devoted, it is possible to pose the foundations for the analysis. The second chapter is dedicated to describing the idea behind the analysis, the corpus and tools chosen, as well as the methodology applied during the preparation of the annotation scheme. The chapter also presents the aforementioned scheme and provides preliminary observations made during the process of annotation. Finally, the first statistical composition of the data given is laid out. Three main points are graphically presented, involving the general distribution of data, and the comparison between English and Polish data, with some focus on the dispersion amongst the annotated elements in the two corpora.

## 2.1 The objective of the thesis and choice of corpus

The purpose of the present study is to provide an efficient and versatile annotation system for the phenomenon of hedging that could be applied to a wider range of languages and registers, which, to the best of the author's knowledge, has not yet been proposed.[11] For this reason, I decided to apply such a scheme to a corpus that would allow for a comparison between two languages, one of them being English, as the research on hedging is already well established in English linguistics, and the other being Polish, on which such studies are scarce.

As previously said (cf. 2.3.1), hedges are particularly difficult to define because of their many forms which elude classification. There is no fixed category of hedging – different elements may obtain this function or not. Hedges' vague pragmatic nature, which depends upon the context and other discourse elements, render them uniquely intangible, or at the very least, problematic to define. Having that in mind, it was clear that the proposed scheme would have to be general enough to capture most of hedging manifestations, yet precise enough to distinguish their various roles so as to allow a more profound comparison between the two languages, independently of their pragmatic characteristics. I intended this system to be suitable for both literary texts and spoken language analysis, especially with the scope of semi-

[11] There have clearly been proposed some different annotation schemes for which one may find instructions that seem to be rather efficient, also in terms of automatic processing of language, however, they are mostly defined for English language only, and are generally specialised for a certain variety/register. Cf. Vincze, Veronika et al. (2008); or Sánchez and Vogel (2015)

automatic translation. For these reasons, it was necessary to choose a corpus that could fill the three requirements:

- Contain comparable material for English and Polish languages;
- Present data representative of a variety on the verge between different registers;
- Contain enough evidence for an accurate verification of the proposed scheme.

After considering a few different open-access corpora, I chose the ParTy corpus as it seemed to fit the needs of my research best. The ParTy corpus[12] contains films' and TED talks' subtitles in more than fifteen languages. It is constantly updated and was created for typological and contrastive purposes, particularly for the comparison of European languages. All files were downloaded from the online repositories *opensubtitles.org*, *subscene.com* and *ted2srt.org* and aligned automatically at the level of sentences or their smaller constituents. Each file represents a language aligned with English and the sentence IDs and line numbers correspond to the same English sentence in all files (Levshina 2015). The identification of equivalents was done automatically with the help of alignment software 'subalign' created by Jörg Tiedemann (2007). The corpus, created by Natalia Levshina as a part of the project "Mapping the causative continuum: A multivariate typological investigation of causative constructions based on a multilingual parallel corpus" (2013), was chosen for its singularity with respect to other known parallel corpora, such as Europarl or many Bible translation projects. The difference is mostly due to the informal register that these data present, while maintaining the characteristics of not being strictly the manifestations of spoken language, somehow merging the peculiarities of spoken and written varieties. In fact, through a series of quantitative analyses based on n-gram frequencies paper demonstrates that subtitles are not fundamentally different from other registers of English and that they represent a close approximation of British and American informal conversations. The comparison was made with data included in the British National Corpus and Santa Barbara Corpus of Spoken American English. A few differences that were discovered, namely the facts that the language of subtitles is more emotional and dynamic but less spontaneous and vague than that of normally occurring conversations, do not truly oppose the goals of this study.

Despite their attractive features, there are some other points in question when it comes to linguistic research on film subtitles. Levshina illustrates them rather well in her article. The first problem is specific to all parallel corpora and concerns the uncertainties of translated texts.

[12] www.natalialevshina.com/corpus.html (last access 30.10.2022)

Often the source of translation is unknown (meaning both the source language of the language pair and the creator), which only amplifies issue of finding a quality translation. Moreover, the creation of subtitles rests on strict rules (though not yet universally standardised), which determine the possible length of lines and other qualities as for the forms used. Many parts must be either omitted, or shortened, thus not providing the best example of comparable pairs of texts. Another point pertaining to the present analysis is that some narrative and discursive elements may be underrepresented in film dialogues with respect to their presence in natural language (Bednarek 2011). Still, other studies underline the similarities with the style of real-world conversations. Lastly, with regard to what was already said as about the translation authorship, subtitles found in online repositories sometimes present translation errors, as well as orthographic and punctuation mistakes. On the other hand, Levshina mentions one advantage of collaborative aspects of the repositories, namely the dedication of the users for reviewing and correcting the files. Having all of that in mind, it is also important to consider that no natural data type or corpus is bereft of mistakes and inherent difficulties. Given the characteristics of hedging, film subtitles offer a possibility of some balance between distinctively literary and conversational language.

The films contained in the corpus represent various fictional genres, according to the genre classification. from the International Movie Database IMDb[13]. For this thesis, I have chosen eight films for a total of sixteen original files (English and Polish version), seven of which were originally in English while one was originally in French, and it is not clear whether the Polish file was prepared through the English-Polish pair or through the French version. The eight films are listed below:

- Amélie (2001) orig. *Le fabuleux destin d'Amélie Poulain* by Jean-Pierre Jeunet;
- Black Swan (2010) by Darren Aronofsky;
- Bridge of Spies (2015) by Steven Spielberg;
- Frozen (2013) by Chris Buck and Jennifer Lee;
- The Grand Budapest Hotel (2014) by Wes Anderson;
- The Iron Lady (2011) by Phyllida Lloyd;
- Noah (2014) by Darren Aronofsky;
- Spectre (2015) by Sam Mendes.

[13] www.imdb.com

Although the ParTy corpus database offered an appealing insight into the cross-linguistic variation of the phenomenon, its aligned files were not suitable for the UD style annotation programmed for this project. An example of one of the ParTy files:

1 When I think of my wife ... Kiedy myślę o żonie ... zawsze myślę o jej głowie .
2 I always think of her head ... Kiedy myślę o żonie ... zawsze myślę o jej głowie .
3 I picture cracking her lovely skull ... Wyobrażam sobie , że rozbijam jej śliczną czaszkę , prostuję zwoje mózgu , usiłując uzyskać odpowiedzi .
4 Unspooling her brains ... Wyobrażam sobie , że rozbijam jej śliczną czaszkę , prostuję zwoje mózgu , usiłując uzyskać odpowiedzi .
5 Trying to get answers ... Wyobrażam sobie , że rozbijam jej śliczną czaszkę , prostuję zwoje mózgu , usiłując uzyskać odpowiedzi .
6 The primal questions of any marriage . Zasadnicze w każdym małżeństwie pytania .
7 What are you thinking ? " O czym myślisz ? " ,
8 How are you feeling ? " Jak się czujesz ? " ,
9 What have we done to each other ? " Co my sobie nawzajem zrobiliśmy ? "
10 The Irish prince graces us with his presence . Zaszczyca nas irlandzki książę .

Connecting the lines in two languages, although visually acceptable, did not allow to transform the files into a format suitable for a more detailed analysis. The alignment of subtitles was not done without errors either. In the screenshot underneath, the Polish subtitle with the ID 4 contains the sentence that in English had to be distributed between lines no.4 and 5. The problem is tackled through the repetition of the same Polish sentence in ID 5. Howbeit a superficial solution, it could have worked if only it was applied for all similar situations in the same manner.

```
0       On September 3 , 1973 ... a blue fly capable of flapping 70 beats a minute ...  3 września 1973 o godzinie 18 . 28 i 32
sekundy ... mucha plujka uderzająca skrzydłami 14670 razy/ min ...
1       landed on St .  lądowała przy ulicy St .
2       Vincent Street in Montmartre .  Vincent na Montmartrze .
3       At that moment , on a restaurant terrace nearby ... the wind magically made two glasses dance unseen ... on a tablecloth
.       W tejże sekundzie na tarasie pobliskiej restauracji ... wiatr wydymał obrus i wprawiał w taniec 2 kieliszki na stole .
4       Meanwhile , in a 5th- floor flat on Avenue Trudaine , Paris 9 ... returning from his best friend 's funeral ...
Zaś na 5 . piętrze domu przy Trudaine 28 w IX dzielnicy ... pan Colere , wróciwszy z pogrzebu przyjaciela , wymazywał go ... z
notesu telefonicznego .
5       Eugene Colere erased him from his address book .        Zaś na 5 . piętrze domu przy Trudaine 28 w IX dzielnicy ... pan
Colere , wróciwszy z pogrzebu przyjaciela , wymazywał go ... z notesu telefonicznego .
6       At the same moment , a sperm with one X chromosome ... belonging to Raphael Poulain ... made a dash for an egg in his
wife Amandine .         W tej samej chwili plemnik z chromosomem X ... Rafaela Poulain ... wyprzedził peleton do jajeczka jego
żony Amandyny z domu Fouet .
7       Nine months later ...   9 miesięcy póŸniej urodziła się Amelia Poulain .
8       Amelie Poulain was born .       9 miesięcy póŸniej urodziła się Amelia Poulain .
9       AMELIE FROM MONTMARTRE  AMELIA
10      Her father , an ex- Army doctor ... works at a spa at Enghien Les Bains .       Ojciec , były lekarz wojskowy , pracuje w
kąpielisku Enghien .
11      TIGHT LIPS , HARD HEART         ZACIŚNIĘTE USTA OZNAKŁ NIECZUŁOŚCI
```

*Figure 2 Example of an error in ParTy corpus alignment*

For these reasons, I decided to keep the film selection, but obtain the singular files directly from the Opensubtitles database. Nevertheless, the ParTy files proved valuable during the cleaning and the preparation of data, as a reference for the correct alignment of lines.

The Opensubtitles[14] website is the biggest collaborative multi-language subtitle database on which many parallel corpora have already been created and which continues to grow. It is

[14] opensubtitles.org (last access 30.10.2022)

quite functional for similar studies, since all the contents are freely available in many formats, including XML and the database offers a well-built search engine. Given that the collection is created by all users, there are usually a few files of each film that can be of different quality. However, most of the resources are well-prepared.

For this study, the abovementioned movies have been downloaded from the Opensubtitles website in the SRT format[15]. The SRT extension is a very common subtitle file format which allows the users to add or modify the subtitles of a video. As for the structure, each subtitle contains a counter of the number/position of subtitles, timestamps, the text of the subtitles, and a blank line separating the elements. One example of the SRT format for one of the files downloaded is provided below:

```
848
01:09:50,227 --> 01:09:53,646
That has nothing to do with it.
I did what I had to to get to my children.

849
01:09:53,814 --> 01:09:55,732
You led us into a war zone
with no way out?

850
01:09:55,899 --> 01:09:59,944
There is a way out. We continue on with
the job, and we do it as fast as possible...
```

In the following section, the process of preparation of the annotation scheme will be presented, with different exploratory approaches exemplified so as to portray the phenomenon of hedging in the most appropriate way. The examples given will exhibit first genuine issues that emerged during the elaboration of the scheme. Subsequently, I will explain the tools and techniques of preparation of the corpus and eventual application of the scheme. Finally, some statistical analysis of these results will be provided as a conclusion of this chapter.

[15] https://docs.fileformat.com/video/srt/

## 2.2 Preparing the annotation scheme

In the following paragraphs, I will present the process of designing the interlinguistic annotation scheme for hedges. The reason for such work is that when operating with foreign language data there is always an overwhelming amount of information to be considered, especially if one is not at all familiar with the language. While the exact meaning of the texts can also be acquired from other sources, and for most scopes of linguistic research that is enough, there are situations where a better understanding of the pragmatic role of a given element in a sentence can be extremely useful and influence the overall interpretation, as well as the researcher's goals in general. A further motivation for designing a new annotation scheme is that others, if there were any, that I was able to find in the related research (cf. Sanchez and Vogel 2015), were either English-centric or only designed for a concrete scope or type of texts. Therefore, it seemed to me that a proposal for a generally applicable scheme could be beneficial not only for my own research, but also for future studies on different languages.

As a first step, it was necessary to decide which classification of hedging phenomena to use, among the many that have been proposed in the literature (see 1.3.2). The first decision that I made, was to remain rather close to the already established models, in order to avoid unnecessary confusion. Moreover, it turned out that a proper merging of the most notable theories on hedging allows for a pretty clear and exhaustive classification. To facilitate this decision and base it on actual data, during the initial stages of the preparation of the corpus' files (cf. 2.3), I gathered the hedging expressions that I identified within the texts in a document so as to test whether they fit into the ideas for classification that I was considering.

From the initial analysis of the gathered data for both languages, it turned out that the most common occurrences of hedges mainly corresponded to Lakoff's original idea of expressions regarding categorical affiliation. Apart from that function, the remaining elements could for the most part be attributed a function of reinforcement, which I initially had not considered, as most of the literature treats it as a separate phenomenon. However, after a careful consideration, given the similarities in the pragmatic use and outcome of attenuation and reinforcement strategies, I decided that reinforcing elements may also be covered by the proposed scheme. As will become clear in the summary presented in 3.3.2, reinforcing strategies turned out to be very prominent in the results of the annotation process. As for the other data, the most present strategy for hedges found in the corpus seemed to be attenuation. Next, there were quite numerous examples of shields, mostly plausibility shields. Given the inherent pragmatic character of hedges, so their inevitable dependence on the situational

context, in many cases it was difficult to determine the exact role of a hedge (cf. 3.1), among those proposed. It became crucial to find a solution for including those vague instances in the scheme. In the following paragraphs, I will present the initial proposals that emerged during the analysis, after which the final scheme will be discussed.

### 2.2.1 Proposal number one

The first attempt to design an annotation scheme is presented in the Table (1). The tags used were thought of in view of the theories discussed in the first chapter (1.3), while not adhering to a single classification. They are to be understood as follows:

- APP stands for approximators;
- HED stands for hedges, while SAH for speech act hedges;
- RNF stands for reinforcement;
- ATT stands for attenuation;
- PRF stands for hedged performatives;
- PLAU stands for plausibility, while CMT for commitment;
- And SCH stands for shield.

The two categories introduced in the first column subdivide hedging elements into those referring to the contents of a single proposition (*approximators*) and those referring to the speech act (either HED for *hedges* or SAH for *speech act hedges*). The former category would in turn contain expressions for reinforcement (RNF) and attenuation (ATT). Attenuation could represent most elements as the propositional attenuating hedges seemed to be the most common in the initial analysis of the data. Attenuation was also expected to translate the most between languages. This tag would include the elements described as adaptors and rounders. The category of speech act hedges contains here three subcategories for performative expressions (PRF), plausibility shields (either PLAU or CMT) and attribution shields (SCH). At first, I thought that performative hedges should constitute a separate, first-level category. However, having an influence on illocutionary and perlocutionary aspects of speech, they seemed most suited for the second category. The initially gathered examples that would be denominated shields (as in Prince 1982or Caffi 1999, 2007) were to constitute two separate subcategories of plausibility/commitment and more literal shield (with the meaning of attribution shield). In the third column some general roles of those expressions were included.

| | | |
|---|---|---|
| APP | RNF | when in context of proximity to the category |
| | ATT | range, and vague category |
| HED/SAH | PRF | scope of mitigation, politeness |
| | PLAU / CMT | level of truth, so commitment to the content of the sentence |
| | SCH | attribution to someone; mitigation self-serving |

*Table 1 The first proposal for annotation scheme*

Prince's (1982) division into propositional and speech act hedging (here presented by ATT vs. HED) seemed to work fine. It was also useful to consider for the aim of confronting two distant or unknown languages: when it comes to hedges on the level of proposition, syntactic annotation would uncover the scope of hedging. However, in the case of speech act hedges, which usually have a scope broader than a single sentence, the interpretation could be confusing. Yet, this extensive pragmatic range makes it difficult for a single category to cover all the possible effects, such as vagueness, evasion, politeness, and mitigation. To be able to clarify which of the pragmatic effects is the aim of a specific expression a further level of classification is necessary.

### 2.2.2 Proposal number two

The second possibility that was considered, was to focus on the pragmatic effects and the roles of expressions, while abandoning the initial two-fold subdivision, so to treat the pragmatic effects as values of one general tag for hedging:

| | |
|---|---|
| HDG | ATT (hedge, round, h. perf) |
| | MTG |
| | RNF |
| | PLAU |
| | SHL |

*Table 2 The second proposal for annotation scheme*

Similarly to the tags used for the first proposal, the acronyms in the Table (2) stand for:

- HDG for hedge;

- ATT for attenuation;
- MTG for mitigation;
- RNF for reinforcement;
- PLAU for plausibility shields;
- and SHL for attribution shields.

As specified in the brackets following the first tag (ATT), the attenuation, being the most common scope of hedging, would have to cover many different types of expressions when it comes to their linguistic form. Namely, the ATT tag in this proposal encompasses both typical shorter forms, usually described as hedges and rounders, and hedged performatives.

A prominent issue concerned the justification of the choice of pragmatic effects to be represented in the scheme. Such a proposal can only include a limited number of tags, so a strict choice was necessary. Among the various pragmatic effects, mitigation and politeness are definitely the most discussed in literature. Moreover, they convey important information for communication that may not necessarily transfer directly into different languages. Some studies explore the different types of mitigation and politeness expressions, also in relation to hedges (Kalisz 1993, Wierzbicka 1985) and it is rather clear that these intentions can sometimes be expressed very differently. Although not every mitigating expression will be a hedge, for the instances that in fact are, I thought it would be interesting to observe how often this exact use is mirrored in translation – the same applying to all other pragmatic effects. Hence, I decided to mostly focus on mitigation, while still including tags representing reinforcement and shields as the strongest alternative strategies. Doubtlessly being a more effect-centred proposal, this scheme allows for a precise analysis of co-textual information regarding the hedging elements. However, losing the distinction between propositional and speech act hedges did not seem the correct choice, given the scope of the scheme that I was trying to present. The objective was to offer a tool that adequately classified different hedges, not simply in view of their potential pragmatic effects. That is why I continued to work on the proposal, until I reached one I found to be a due compromise.

### 2.2.3 Proposal number three

The final proposal I devised satisfied most of the requirements. First, it is necessary to address any dearth that may be pointed out. The proposed annotation scheme is inevitably limited with respect to the range of the concept of hedges that I shortly presented in the first chapter. While acknowledging the extent of the phenomenon, the scope of this thesis is to

develop a tool concise enough to be easily applicable to a vast range of texts so as to facilitate further computational analysis of hedges. That is not to say that no further extension of the scheme can be proposed; however, having in mind the potential of designing tools for semi-automated analysis of such a pragmatic, thence elusive, aspect of language, a more generic model seemed most functional.

The table beneath presents the final annotation scheme applied and successfully examined on the chosen corpus, along with some examples. To facilitate the interpretation of the tags proposed, I included a short legend, as well as corresponding classifications found in the literature.

| type | literature | values | examples | Legend | description |
|---|---|---|---|---|---|
| **PROP** | hedge, round | **ATT \| RNF \| EV** | That's **kind of** freaky. | **PROP** | propositional |
| | | | And I was hoping for **some sort of** tactical plan | **PRF** | hedged performative |
| **PRF** | hedged performatives | **MITS \| MITA \| NPL** | And you **may choose** a woman. | **CMT** | commitment |
| | | | I **might just give** you a big wet kiss. | **SCH** | shield |
| **CMT** | plausibility shield | **MITS MITA \| NPL** | **It's hard to say**, but she wasn't a redhead. | **ATT** | attenuation |
| | | | **I don't know if it's true.** | **RNF** | reinforcement |
| | | | I said you weren't interested, **right**? | **EV** | evasion |
| **SCH** | attribution shield | **MITS \| MITA \| NPL** | **She said** that he stalked her. He's in St. Louis. | **MITS** | self-serving mitigation |
| | | | And **according to your boss.** | **MITA** | altruistic mitigation |
| | | | | **NPL** | negative politeness |
| | | | | **DISC** | imprecise discourse effect |

*Table 3 The final proposal for annotation scheme*

As previously mentioned, it was deemed important to keep the distinction between the hedges referring to the propositional content and to the speech act as a whole. However, after the analysis of numerous examples, I noticed that the latter only took the form of shields and

hedged performatives (for this category also the simple modal verbs were included, as the modality information, at least in English, provide similarly valuable information and modal verbs are often included in the lists of hedges). For this reason, I decided that there will be one tag for propositional hedges which can be attributed values of attenuation and reinforcement, while tags for hedged performatives, plausibility shields (acting on commitment), and attribution shields, all referring to the speech act, can moreover obtain values of self-serving and altruistic mitigation and negative politeness. Subsequently, I added the value of evasion to the propositional hedges, as well as one value encompassing unlisted or simply not easily specified discourse effects (DISC) for speech act hedges. While the values were not restricted only to certain tags during annotation, the results proved the described subdivision to be well adjusted. This representation offers one major advantage: in case any values were to be added to the scheme, it would not negatively affect its entirety. Moreover, to accommodate an easier analysis of the multiword hedges that were expected to emerge, the positional value was also added for any instance in need of such description:

- (hedge tag)|Position=Initial – for the first element;
- (hedge tag)|N=x – for any subsequent element not being the last, where x stands for a min. 2 integer, incrementing by one for each following token;
- (hedge tag)|Position=Final – for the last element.

The presented scheme was applied during the annotation of the corpus introduced in the first section of this chapter. In the following paragraphs, the entire process of adapting and describing the files will be presented, as well as the preliminary results of distribution and hedges' frequencies that will serve as an introduction for the concluding chapter of this thesis.

## 2.3 Corpus preparation and annotation

As indicated in the first section of this chapter, the corpus chosen for this study was modelled after the one proposed by Natalia Levshina. Having downloaded the necessary files, before the data could be used for analysis and preparation of the annotation scheme, they needed to be pre-processed. While describing the database of Opensubtitles, it was mentioned that the contents are user-made. This collaborative aspect can unfortunately result in some errors in preparation or encoding. From the attached table, it is evident that some of the files changed radically through the three stages of preparation:

| **File** | **Lg.** | **Original** | | **Pre-processing** | | **CoNLL-U format** | | **Aligned** | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Tokens** | **Lines** | **Tokens** | **Lines** | **Tokens** | **Lines** | **Tokens** | **Lines** |
| *Amélie* | **eng** | 17244 | 4726 | 6962 | 1103 | 140444 | 14092 | 133549 | 13870 |
| | **pl** | 14074 | 4484 | 5164 | 922 | 145370 | 11572 | 144534 | 12296 |
| *Black Swan* | **eng** | 13056 | 3828 | 5060 | 887 | 105116 | 11097 | 87334 | 8823 |
| | **pl** | 7519 | 2286 | 3082 | 495 | 90138 | 6795 | 90416 | 6814 |
| *Bridge of Spies* | **eng** | 29924 | 8795 | 12089 | 1948 | 239469 | 24751 | 221321 | 21963 |
| | **pl** | 20548 | 6642 | 7748 | 1420 | 210512 | 16412 | 202466 | 16875 |
| *Frozen* | **eng** | 24765 | 7752 | 8528 | 1764 | 179277 | 19557 | 159094 | 17222 |
| | **pl** | 18667 | 5845 | 6541 | 1343 | 190049 | 15403 | 185681 | 16133 |
| *The Grand Budapest Hotel* | **eng** | 21720 | 6143 | 9098 | 1340 | 173308 | 17598 | 169294 | 17327 |
| | **pl** | 17294 | 5133 | 5978 | 1179 | 168548 | 13395 | 163265 | 13960 |
| *The Iron Lady* | **eng** | 21225 | 5754 | 9190 | 1288 | 175069 | 17782 | 162646 | 16791 |
| | **pl** | 15802 | 4937 | 5684 | 1119 | 157240 | 12535 | 154690 | 12149 |
| *Noah* | **eng** | 19499 | 6107 | 6263 | 1453 | 137987 | 15300 | 120287 | 13432 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **pl** | 14051 | 4520 | 3992 | 1092 | 123489 | 10397 | 121582 | 11242 |
| *Spectre* | **eng** | 19936 | 6065 | 7470 | 1363 | 152205 | 16102 | 133839 | 14159 |
| | **pl** | 13734 | 4432 | 4621 | 998 | 138841 | 11241 | 135975 | 11836 |
| **Total** | | **289058** | **87449** | **107470** | **19714** | **2527062** | **234029** | **2385973** | **224892** |

*Table 4 A summary of the corpus data throughout the different stages of the study*

In general, the changes that occurred can be easily explained. As one may see from the table, for all the files there was a dramatic reduction in the pre-processing stage which resulted from the deletion of all the structural information of srt files. This can be most obviously observed in the fourth column of the data, regarding the lines of text within the file. Subsequently, the numbers notably grew, having entered the CoNLL-U dedicated column. Given the composition of this type of files, even the mere segmentation of the sentences into lemmas, each with a dedicated line of annotation, contributes to the major increase that can be seen - from ten times (for lines of text) to even thirty times (for tokens) the number of pre-processing data.

As for the final presentation of the files, changes were not so radical, especially in the case of tokens which on average decreased by around nine thousand. Interestingly enough, in the final column for the lines of text we can see that, while English language data still consistently decreased by about 200 to 2200 points, almost all those in Polish grew, even by 730 points. For tokens, it only happened with the Polish *Black Swan* subtitles.

In the pre-processing phase of the work, it was necessary to transform the srt subtitle files into text files, containing only the necessary information, i.e., lines of text to be later properly segmented within the corresponding CoNLL-U format file. To do that, a simple Python script based on *pysubs2* was applied. *Pysubs2*[16] is a Python library for editing subtitle files. Its methods allowed for a smooth modification of the files' format, after which they could be cleaned with simple commands. Beneath, I presented the code used for the process:

[16] https://github.com/tkarabela/pysubs2, last accessed: 3.10.2022.

```
import pysubs2
subs =
    pysubs2.load("Avatar_2009_en.srt",
    encoding="Windows-1250")
f = open("output.txt", "a")
i = 0
for line in subs:

    print(subs[i], file=f)
    i = i + 1

f.close()
```

```
    f = open("output.txt", "r", encoding =
"Windows-1250")
    f2 = open("subtitles.txt", "a",
encoding = "Windows-1250")
text = f.readlines()
for line in text:

    subtitles = line[55:-2]
    print(subtitles, file=f2)

    f.close()
f2.close()
f=open("subtitles.txt")
text=f.read()
import re
new_line = text.replace("\\\\N", " ")
f2 = open("step_1.txt", "a")
print(new_line, file=f2)
    f.close()
f2.close()


f=open("step_1.txt")
text=f.read()
corsive_1 = text.replace("\\\\i1", "")
f2 = open("step_2.txt", "a")
print(corsive_1, file=f2)

    f.close()
f2.close()


f = open("step_2.txt")
text = f.read()
corsive_0 = text.replace("\\\\i0", "")
f2 = open("step_3.txt", "a")
print(corsive_0, file=f2)

    f.close()
f2.close()


f= open("step_3.txt")
text = f.read()
brac = text.replace("{}", "")
f2 = open("final.txt", "a")
print(brac, file=f2)

    f.close()
f2.close
```

One clarification is due here, namely, that the encoding type during this process varied per files, since the attempts to normalise all the files to UTF-8 caused various errors. The

transformation between UTF-8 and Windows-1250, sometimes even without specifying the encoding, worked best at this point and did not affect the following transition into CoNLL-U. Figure 3 shows the process at work.

```
In [1]: import pysubs2
   ...:
   ...: subs = pysubs2.load("Gone_Girl_2014_pl.srt", encoding="Windows-1250")

In [2]: f = open("output.txt", "a")
   ...:
   ...: i = 0
   ...:
   ...: for line in subs:
   ...:     print(subs[i], file=f)
   ...:     i = i + 1
   ...:
   ...: f.close()

In [3]: f = open("output.txt", "r", encoding = "Windows-1250")

In [4]: f2 = open("subtitles.txt", "a", encoding = "Windows-1250")

In [5]: text = f.readlines()
   ...:
   ...: for line in text:
   ...:     subtitles = line[55:-2]
   ...:     print(subtitles, file=f2)
   ...:
   ...: f.close()
   ...:
   ...: f2.close()
```

*Figure 3 The process of transforming the subtitle files using the code presented.*

For the UD CoNLL-U transformation using UDPipe online software[17], among those available for the selected languages in the UD database, I chose the UD Polish PDB[18] for the Polish versions and the UD English LinES[19] for the English files.

### 2.3.1 The Polish PDB Treebank

The Polish PDB treebank, available since UD v1.2 release, was prepared by Alina Wróblewska, Daniel Zeman, Jan Mašek, and Rudolf Rosa. It was annotated manually in non-UD style and automatically converted to UD. PDB-UD treebank is based on the Polish Dependency Bank 2.0 (PDB 2.0), created at the Institute of Computer Science, Polish Academy of Sciences in Warsaw. It contains 22,152 sentences (350K tokens) of texts from fiction, non-fiction, and news, annotated with the following tags:

- POS Tags

ADJ – ADP – ADV – AUX – CCONJ – DET – INTJ – NOUN – NUM – PART – PRON – PROPN – PUNCT – SCONJ – SYM – VERB – X

- Features

[17] https://lindat.mff.cuni.cz/services/udpipe/, last accessed: 15.10.2022.
[18] https://universaldependencies.org/treebanks/pl_pdb/index.html, last accessed: 15.10.2022.
[19] https://universaldependencies.org/treebanks/en_lines/index.html, last accessed: 15.10.2022.

Abbr – AdpType – Animacy – Aspect – Case – Clitic – ConjType – Degree – Emphatic – Foreign – Gender – Hyph – Mood – Number – Number[psor] – NumForm – NumType – PartType – Person – Polarity – Polite – Poss – PrepCase – PronType – Pun – PunctSide – PunctType – Reflex – Tense – Variant – VerbForm – VerbType – Voice

- Relations

acl – acl:relcl – advcl – advcl:cmpr – advcl:relcl – advmod – advmod:arg – advmod:emph – advmod:neg – amod – amod:flat – appos – aux – aux:clitic – aux:cnd – aux:imp – aux:pass – case – cc – cc:preconj – ccomp – ccomp:cleft – ccomp:obj – conj – cop – csubj – csubj:pass – dep – det – det:numgov – det:nummod – det:poss – discourse:emo – discourse:intj – expl:pv – fixed – flat – flat:foreign – iobj – list – mark – nmod – nmod:arg – nmod:flat – nmod:poss – nmod:pred – nsubj – nsubj:pass – nummod – nummod:flat – nummod:gov – obj – obl – obl:agent – obl:arg – obl:cmpr – obl:orphan – orphan – parataxis:insert – parataxis:obj – punct – root – vocative – xcomp – xcomp:cleft – xcomp:pred – xcomp:subj

### 2.3.2 The English LinES Treebank

The UD English LinES treebank, available since UD c1.3 release and designed by Lars Ahrenberg. The texts included are fiction, nonfiction, and spoken (Europarl), also annotated manually in non-UD style and automatically converted. The tags used are:

- POS Tags

ADJ – ADP – ADV – AUX – CCONJ – DET – INTJ – NOUN – NUM – PART – PRON – PROPN – PUNCT – SCONJ – SYM – VERB – X

- Features

Case – Definite – Degree – Gender – Mood – Number – NumType – Person – Poss – PronType – Reflex – Tense – VerbForm – Voice

- Relations

acl – acl:relcl – advcl – advmod – amod – appos – aux – aux:pass – case – cc – ccomp – compound – compound:prt – conj – cop – csubj – csubj:pass – dep – det – discourse – dislocated – expl – fixed – flat – iobj – mark – nmod – nmod:poss – nsubj – nsubj:pass – nummod – obj – obl – orphan – parataxis – punct – root – vocative – xcomp

The open-source pipeline UDPipe[20] chosen for the conversion process can be used for tokenization, tagging, lemmatization, and parsing of CoNLL-U files and offers C++, Python, Perl, Java, C# libraries, and an online software. The trained models provided included those for the treebanks above. In the figure 4, the view of the UDPipe conversion tool is given. As it shows, first the UD 2.10 version was chosen for both languages. Next, the plain text files were uploaded from the disc, and, after the processing, they could be downloaded back in the desired format.

[20] https://universaldependencies.org/tools.html#udpipe

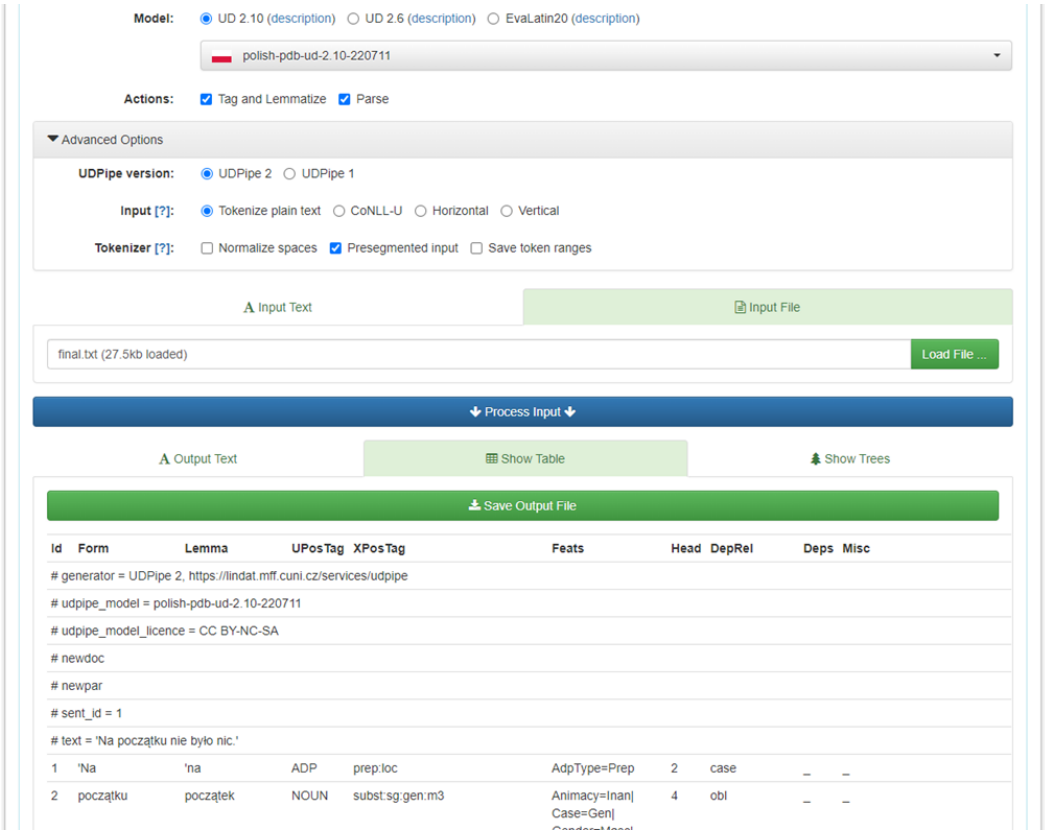



*Figure 4 The process of transforming the files into CoNLL-U using UDPipe.*

Having executed this task for all the selected subtitle files, it was possible to align and annotate them according to the scheme proposed. Both of these operations will be discussed in the next section.

## 2.4 Annotation workflow

The previous segments of this chapter presented all the various operations of preparation with relation to the annotation scheme proposal and the files formatting. This section shortly presents how the actual annotation transpired and what insights it offered before the whole process of analysis was complete.

### 2.4.1 Preparation

It was briefly mentioned in 2.2 that before confirming and applying the scheme, a few of the files were scanned for examples of hedging devices. This allowed me to revise the 'natural' data of the corpus in order to make the best possible decision for the preliminary proposal. The samples included the sentences in which the different expressions occurred, along with the POS and DEPREL tag which were attributed to the hedges.

| | A | B | C | D | E | F | G | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | Source | Id2 | Sentence | Hedge | POS | Deprel | ID | Source | Id2 | Sentence |
| 2 | 1 | blackswan_eng | 4 | It was different choreography though. It was more like the Bolshoi's. | different | adj | amod | 1 | blackswan | 3 | Ale choreografia była inna. B stylu Teatru Bolszoj. |
| 3 | | | | | very | adv cmp | advmod | 2 | | 5 | Jaki różowy. Śliczny. |
| 4 | | | | | like | adp | case | 3 | | 7 | Jesteś w dobrym nastroju. - C w tym sezonie będę więcej g |
| 5 | 2 | | 10 | He promised to feature me more this season. | promise | verb | root | 4 | | 8 | Powinien ci dać grać. Jesteś t wystarczająco długo i jesteś zdecydowanie najbardziej oc tancerką w zespole. |
| 6 | | | | | very | adv cmp | advmod | 5 | | | |
| 7 | 3 | | 11 | Well, he certainly should. You've been there long enough... | certainly | adv cmp | advmod | 6 | | 11 | Nic. |
| 8 | | | | | enough | adv cmp | advmod | 7 | | 12 | Na pewno nie chcesz, bym p |
| 9 | | | | | | | | 8 | | 14 | Widziałaś dziś Beth? Nie wie wróciła. |
| 10 | 4 | | 12 | and you're the most dedicated dancer in the company. | much | adv spl | advmod | 9 | | 15 | Oczywiście, że wróciła. - Nie Teatr jest spłukany. |
| 11 | 5 | | 16 | Nothing. | nothing | pron | root | 10 | | 16 | Nikt już nie przyjdzie, by ją o już w ogóle nie przychodzi na |
| 12 | 6 | | 17 | Sure you don't want me to come with you? | sure | adv | advmod | 11 | | | |
| 13 | 7 | | 21 | Did you see Beth today? I can't believe she's back. | can | aux pres-aux | aux | 12 | | | |
| 14 | | | | | nothing | part | advmod | 13 | | | |
| 15 | 8 | | 23 | What? She can't take a hint? The company's broke. No one comes to see her anymore. | no | pron | nsubj | 14 | | | |

*Figure 5 Example of the initial file for manual observations*

Apart from the importance for the subsequent preparation of the annotation scheme, these data showed that in both languages it can be expected that:

- The majority of hedges will be propositional;
- The pronominal hedges were mostly expressed by different adverbs;
- Speech act hedges were far less common and most often consisted of multiword expressions including verbs and sometimes nouns or elements carrying case markings.

For the following operations of annotation, UD Annotatrix[21] was chosen. It is an open source browser-based offline and online annotation tool for the UD framework. The online version displays an updated file segment by segment and allows the user to modify each and any tags and relation, in addition to generating dependency trees. Despite having already converted the file into the CoNLL-U format, the process of alignment could not start right away, as there were many problems concerning the alignment of sentences that emerged.

### 2.4.2 The process and problems:

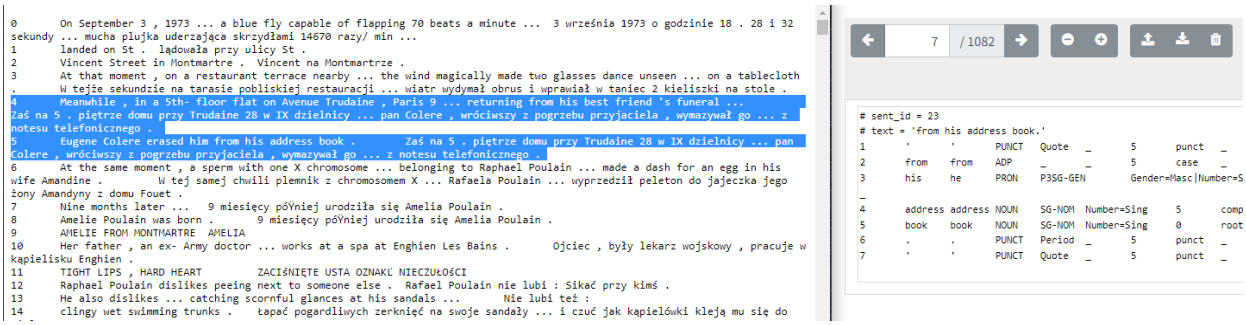



*Figure 6 Example of the differences between the ParTy corpus and the lines in the files I prepared.*

The most prominent practical problem I encountered was the issue of alignment of the subtitles. Despite a careful choice of files among those available on the Opensubtitles repository, unfortunately, all of them had bigger or smaller discrepancies between the original English and target Polish version. Some inconsistencies were due to the differences between the two languages and were naturally expected. English is an analytic language, with only simple inflections remaining causing frequent zero derivation and overall flexibility with regard to word-building, for example creation of compounds. English attaches great importance to the word order, with few word patterns that have to be followed, including the strict principles governing relative positions of various word classes, such as adjectives. Rarely is it possible to modify the word order without adding or subtracting a word. Polish is a West Slavic language, highly inflected, with characteristic presence of alternation. The syntax and word order of the whole sentence are dependent on the inflection. One of distinctive properties of Polish, and other Slavic languages, is a very salient contrast between perfective and imperfective aspect of the verb, expressed through inflection as well. Polish language has many discrete functional styles (like scientific and journalistic). While most of those differences do not cause severe problems in translation between the languages, it ensures that the contents of

[21] https://github.com/jonorthwash/ud-annotatrix, last accessed: 10.02.2023.

the corresponding texts will be quite different, if only for the languages' grammatical form. As a matter of fact, another unavoidable issue related to the standards for subtitle creation and translation, dictating the character length of lines. That, together with natural variation in sentence and word length across languages (cf. Smith 2012), would have influenced the contents of respective files in a language pair in any case.

As can be observed from Table 4, the files in English typically contained more lines of text, while having fewer tokens. That is because English in general is much more concise a language than Polish, so the original sentences were often much shorter than the translated ones, like in the case of *The Iron Lady* with 16791 lines in English and only 12149 in Polish. An example is presented in Figure 7.

| # text = 'Shall I call someone, see if anyone can come and do your hair?' | | | | | # text = 'Mam zawołać kogoś, żeby cię uczesał?' | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ' | ' | PUNCT | Quote | 1 | 'Mam | 'mać | VERB | fin:sg:pri:imperf | I-have-to |
| 2 | Shall | shall | AUX | PRES-AUX | 2 | zawołać | zawołać | VERB | inf:perf | call |
| 3 | I | I | PRON | PERS-P1SG-NOM | 3 | kogoś | ktoś | PRON | subst:sg:acc:m1 | someone |
| 4 | call | call | VERB | INF | 4 | , | , | PUNCT | interp | , |
| 5 | someone | someone | PRON | IND-SG-NOM | 5 | żeby | żeby | SCONJ | comp | in-order-to |
| 6 | , | , | PUNCT | Comma | 6 | cię | ty | PRON | ppron12:sg:acc:m1:sec:nakc | you |
| 7 | see | see | VERB | INF | 7 | uczesał | uczesać | VERB | praet:sg:m1:perf | brush-hair |
| 8 | if | if | SCONJ | _ | 8 | ? | ? | PUNCT | interp | ? |
| 9 | anyone | anyone | PRON | IND-SG-NOM | 9 | ' | ' | PUNCT | interp | |
| 10 | can | can | AUX | PRES-AUX | | | | | | |
| 11 | come | come | VERB | INF | | | | | | |
| 12 | and | and | CCONJ | _ | | | | | | |
| 13 | do | do | VERB | INF | | | | | | |
| 14 | your | you | PRON | P2-GEN | | | | | | |
| 15 | hair | hair | NOUN | SG-NOM | | | | | | |
| 16 | ? | ? | PUNCT | QuestionMark | | | | | | |
| 17 | ' | ' | PUNCT | Quote | | | | | | |

*Figure 7 Example of difference in the lengths and distribution of lines between English and Polish versions of the subtitles*

English lines were not only shorter in length, but also more subdivided, whereas lines in Polish often merged multiple sentences and lines of dialogue. That was especially evident in the case of "Black Swan," where the English file consisted of almost twice as many lines. One can observe some relevant differences even between different versions of the subtitles for a single movie:

(1) ParTy for Amélie (only the English fragment):

*ID 4* *Eugene Colere erased him from his address book.*

CoNLL-U:

*ID 6* *Eugene Colere erased him*

*ID 7* *from his address book.*

Example (1) shows how the ParTy corpus version of *Amelie* tended to keep the entire regular sentences together, while the subtitle files transformed into CoNLL-U tended to divide them in more separate lines. All of the above resulted in quite a different distribution of lines in the two files for each movie and made the process of alignment quite complicated. For the sake of the analysis, I decided to ignore the aforementioned standards for subtitles' creation, as they do not benefit the research. Instead, I simply focused on the contents of the created corpus, by aligning the corresponding sentences. This way, the contents matched in both versions (at least in context, if not for the proper translation).

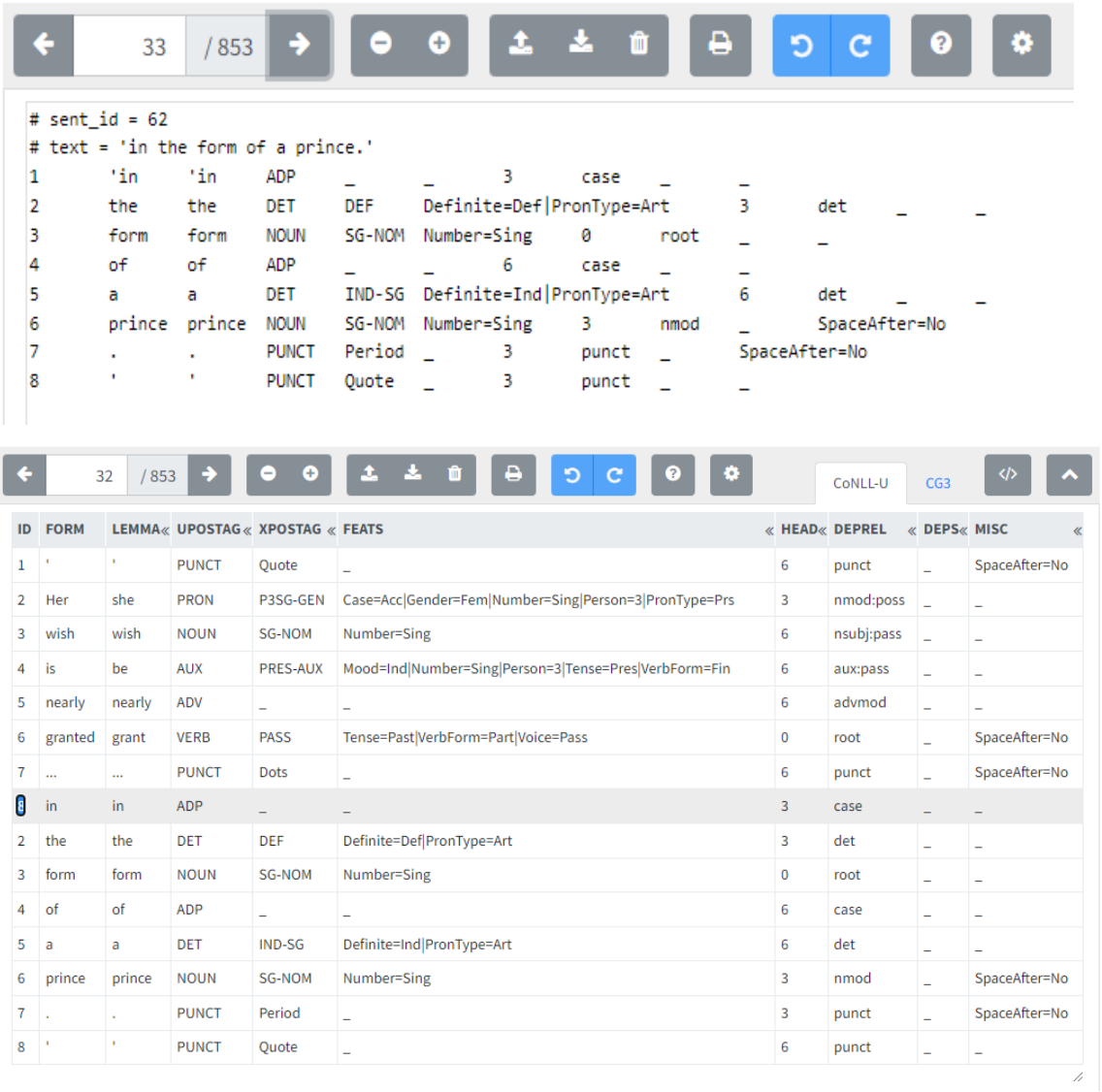



*Figure 8 Example of joining two separate sentences so as to align the English and the Polish versions of the subtitles.*

Figure 8 demonstrates how, in order to obtain comparable lines of text, segments often had to be merged and syntactic relations holding between their elements had to be fixed accordingly. As can be seen in the case of the three dots in Figure 8, punctuation marks and the original sentence segmentation were usually maintained, as they could express cues about the situational context (for example a pause here) that could potentially influence the interpretation of data.

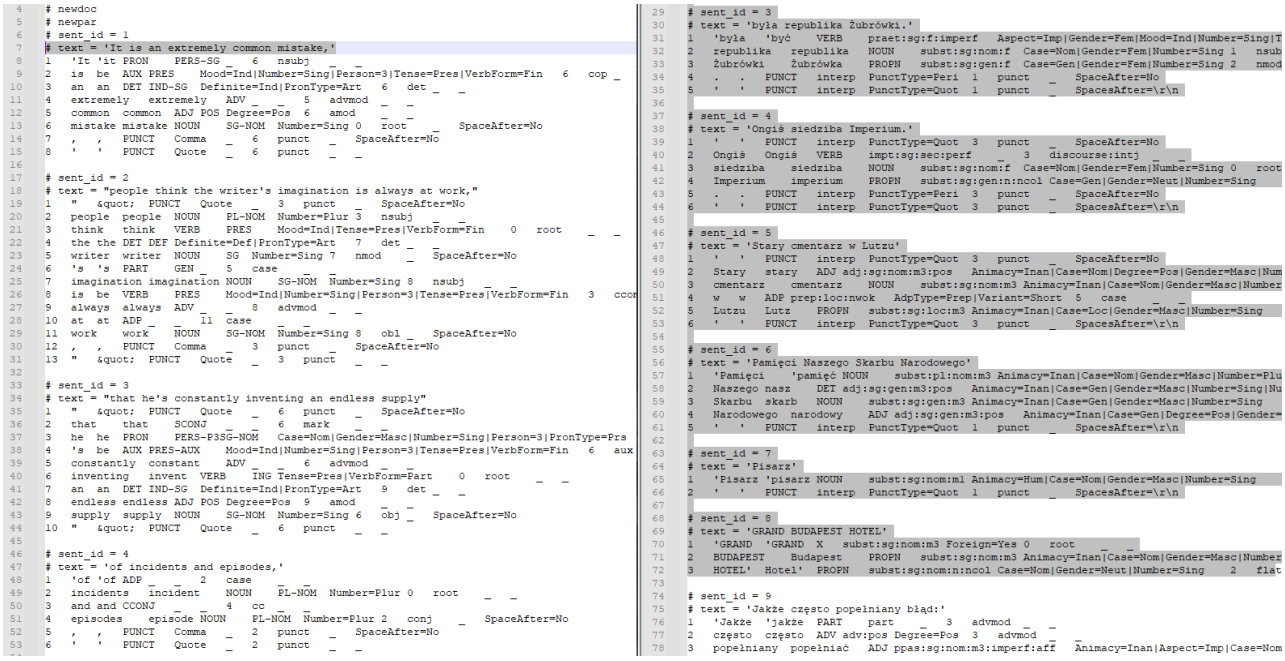


*Figure 9 Example of additional Polish lines, not present in the English subtitles.*

In other cases, some chunks of text only appeared in one of the two languages. It is difficult to determine why exactly that happened, as several explanations are plausible: in some cases, it could be due to the translators' decision; often the lines belonged to background or secondary characters; sometimes, lines spoken in a language other than English (as in the case of "*The Bridge of Spies*") were only subtitled in one version and not the other. In those cases, the best solution was to simply delete the interfering lines, as they would have made the analysis much more challenging. An example of this situation is given above: the first eight lines of Polish introduction to the "Grand Budapest Hotel" were entirely omitted in English subtitles, so they had to be deleted during the annotation. Having completed all the operations of annotation, the files were submitted to a verification using the Valideasy[22] python library.

Even though the process allowed for some captivating observations and insights into the choices and interpretation of respective translators, which inevitably influenced the results of the present research, it also slightly limited the chosen corpus, due to forced cuts, and made the entire process quite time consuming. This, however, speaks in favour to the points made previously with respect to the sustainability of early work on corpora (cf. 1.2), taking into account the relatively smaller dimensions of the chosen corpus. As for the analysis, despite the

[22] https://github.com/unipv-larl/valideasy/blob/main/README.md, last accessed: 25.02.2023.

loss of some data, cutting some sentences seemed to be the most reasonable choice. Once the process was completed, more accurate evaluation of the files could be carried out.

## 2.5 Preliminary results

Following the efforts described in the previous sections, it was possible to conduct an analysis on the data obtained. Although the more in-depth insights will be discussed in the following chapter, to properly conclude the description of methods applied in this work, I will present the overall statistics concerning the number and type of hedges observed. In the last section of this chapter, I will provide a few graphic representations concerning the quantitative distribution of tags, in general and across languages, as well as a comparison between the files in the corpus.

### 2.5.1 Quantitative distribution of hedge types and values

First of all, given the limited size of the corpus, the concern was whether there will be enough hedge-related data for a due discussion. As demonstrated in the table below, the corpus contained enough data for a comparison of hedging-related phenomena in the two languages.

| Tag | English | Polish | Singular | Multiword | Total |
|---|---|---|---|---|---|
| PROP | 1432 | 831 | 1944 | 319 | 2263 |
| PRF | 183 | 125 | 32 | 276 | 308 |
| CMT | 705 | 546 | 295 | 920 | 1251 |
| SCH | 87 | 55 | 17 | 125 | 142 |

*Table 5 Total number of hedges in the four categories*

As predicted following the observations in 2.4.1., by far the largest group of hedges belong in the PROP tag category, followed by expressions classified as CMT, with slightly over a thousand items less. PROP tag was also attributed 5,5 times more often to single tokens, rather than a longer expression. The opposite holds for all speech act tags, which were mainly applied to longer segments. The table also shows how the number of hedges is also consistently higher in English that in Polish files.

Table 5, however interesting, only offers a general view of how tags and values are distributed in the corpus. Some more detailed information can be extracted from the list of frequencies for tags and values, along with its graphic representation.

| Tags | $f_i$ | rel. $f_i$ | $fc_i$ | rel. $fc_i$ |
|---|---|---|---|---|
| PROP=ATT | 1176 | 29,939% | 1176 | 0,299389002 |
| PROP=RNF | 984 | 25,051% | 2160 | 0,549898167 |
| CMT=DISC | 464 | 11,813% | 2624 | 0,66802444 |
| CMT=ATT | 457 | 11,634% | 3081 | 0,784368635 |
| CMT=RNF | 226 | 5,754% | 3307 | 0,841904277 |
| PROP=EV | 103 | 2,622% | 3410 | 0,868126273 |
| PRF=RNF | 85 | 2,164% | 3495 | 0,889765784 |
| SCH=DISC | 81 | 2,062% | 3576 | 0,910386965 |
| PRF=ATT | 74 | 1,884% | 3650 | 0,929226069 |
| PRF=NPL | 71 | 1,808% | 3721 | 0,947301426 |
| PRF=DISC | 63 | 1,604% | 3784 | 0,963340122 |
| SCH=MITS | 48 | 1,222% | 3832 | 0,975560081 |
| CMT=MITS | 45 | 1,146% | 3875 | 0,987016293 |
| CMT=NPL | 16 | 0,407% | 3891 | 0,991089613 |
| PRF=MITS | 12 | 0,305% | 3903 | 0,994144603 |
| SCH=ATT | 8 | 0,204% | 3911 | 0,996181263 |
| CMT=MITA | 7 | 0,178% | 3918 | 0,99796334 |

| | | | | |
|---|---|---|---|---|
| PRF=MITA | 3 | 0,076% | 3921 | 0,998727088 |
| SCH=RNF | 3 | 0,076% | 3924 | 0,999490835 |
| SCH=MITA | 1 | 0,025% | 3925 | 0,999745418 |
| SCH=NPL | 1 | 0,025% | 3926 | 1 |
| TOTAL | 3926 | 29,939% | | 0,299389002 |

*Table 6 A list of frequencies for each of the tag-value combinations present*

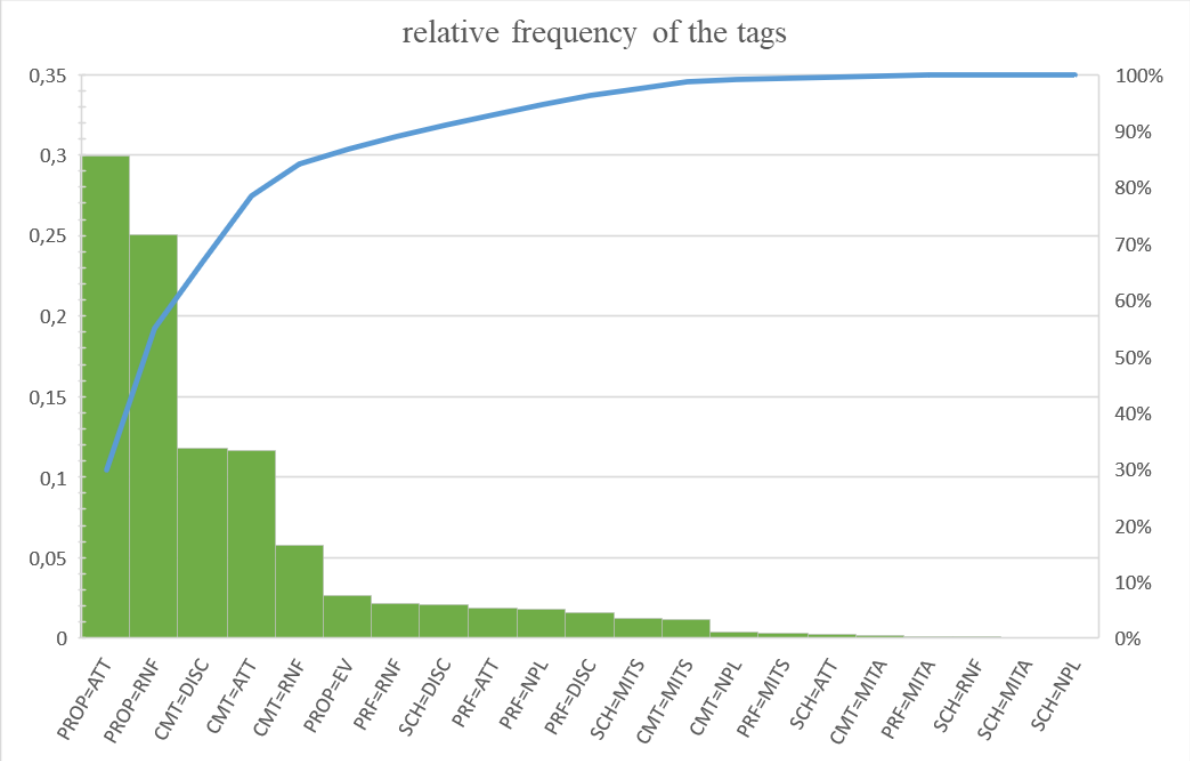



*Figure 10 Visual representation of the relative frequency of tag-value pairs*

The distribution here is clearly leaning towards unimodality, with the two most frequent tags: PROP=ATT and PROP=RNF constituting over half of the hedges observed, PROP=ATT itself being almost 30%. None of the expected hedges were inexistent, although essentially all speech act hedges (except for three) are included in the last 10% of the data. Attribution shields occupy the end of the chart, with mostly insignificant quantities of data, similarly to the MITA value. Apart from the predominant presence of propositional hedges, established previously, the leading positions are taken by general values of attenuation and reinforcement, as well as that of an unspecified discourse effects.

For more exact information on the distribution of values across the four tags, the mean for each of them has been presented in Table 7.

| Tag x̄ | Language | ATT | RNF | EV | MITS | MITA | NPL | DISC |
|---|---|---|---|---|---|---|---|---|
| **PROP** | English | 0,5006983 | 0,452514 | 0,0467877 | | | | |
| | Polish | 0,5523466 | 0,4043321 | 0,0433213 | | | | |
| **PRF** | English | 0,2677596 | 0,2896175 | | 0,054644809 | 0,016393 | 0,169399 | 0,202186 |
| | Polish | 0,2 | 0,256 | | 0,016 | 0 | 0,32 | 0,208 |
| **CMT** | English | 0,3248227 | 0,1801418 | | 0,043971631 | 0,007092 | 0,014184 | 0,429787 |
| | Polish | 0,4175824 | 0,1813187 | | 0,025641026 | 0,003663 | 0,010989 | 0,294872 |
| **SCH** | English | 0,0581395 | 0,0116279 | | 0,325581395 | 0,011628 | 0 | 0,593023 |
| | Polish | 0,0555556 | 0,037037 | | 0,333333333 | 0 | 0,018519 | 0,555556 |

*Table 7 Mean numbers for the tag-value distribution in the two languages*


*Figure 11 Percentual distribution of the four main tags in both languages*

The first observation is that the MITA value for PRF and SCH exist almost exclusively for English, while the SCH=NPL forms were observed only for Polish expressions. The PROP values distribution is highly balanced, the CMT one almost as much, with again smaller mean amount of mitigation instances for Polish. Kalisz's (1993) considerations against Wierzbicka's (1985) view of hedges in Polish comes to mind, as it seems that some speech acts expressed for example in hedges[23] in fact appear more rarely in Polish. Although, the difference is not as drastic as Smith (2012) seems to suggest (at least in Kalisz's interpretation), especially considering the non-literal translation of the subtitles, among other factors.

### 2.5.2 Interlinguistic comparison

The remarks from the previous paragraph are supported by the table and chart representation of tag values distribution English and Polish, where the further almost consistently contains less than 40% of total number of hedges, even for all the speech act values summarised (DISC total).

| Values | ATT | RNF | EV | MITS | MITA | NPL | DISC | DISC TOTAL |
|---|---|---|---|---|---|---|---|---|
| For Eng. | 1000 | 829 | 67 | 69 | 9 | 41 | 391 | 510 |
| For Pol. | 715 | 469 | 36 | 34 | 2 | 47 | 217 | 300 |
| Total | 1715 | 1298 | 103 | 103 | 11 | 88 | 608 | 810 |

*Table 8 The interlinguistic comparison of the specific values for hedges present*

[23] Both works cited discuss hedging as only one of the speech acts listed.

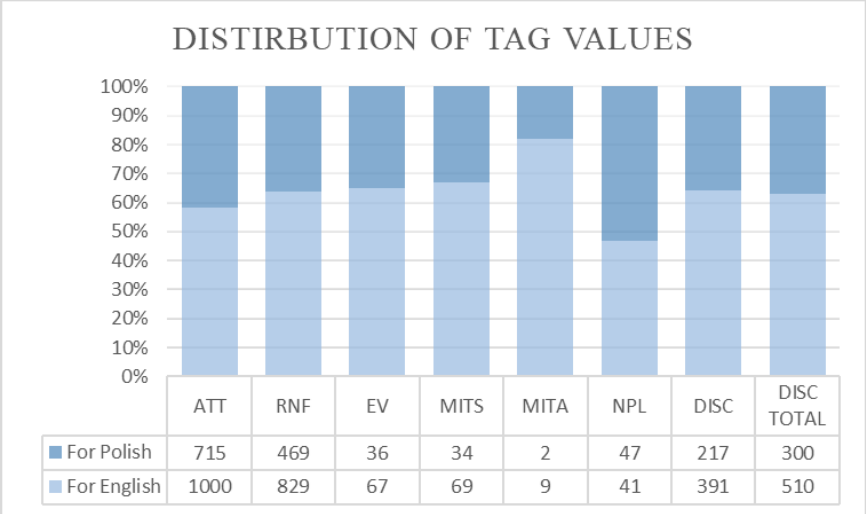



*Figure 12 Percentual distribution of the tag values for the two languages*

Here, once more, the NPL value, independently of tag, is slightly more present in Polish subtitles, while the altruistic mitigation prevails in English, although being rare in both languages. Apart from these two, the singular values of the tags seem rather imbalanced when compared to each other. More general labels such as attenuation, reinforcement, and discourse, are more represented in the data with respect to more specific tags. Still, most of the tags do not seem to be underrepresented in either language, and if so, it could speak to the given language characteristics and chosen pragmatic strategies.

### 2.5.3 Distribution within the corpus

The final topic of these introductory data representations is that of the subdivision of hedges within particular files. When introducing the idea for the choice of corpus, it was said that the vital argument for reproducing the ParTy corpus by Levshina, was that the films she proposed were all well written and each presented a very different set of characteristics when it comes to characters, topic, setting, etc, all partly identifiable in the language used. The films presented here also mirror this diversity.

| Film | English | Polish | Film/Item |
|---|---|---|---|
| Amelie | 188 | 156 | 344 |
| BS | 219 | 142 | 361 |

| | | | |
|---|---|---|---|
| BoS | 606 | 338 | 944 |
| Frozen | 297 | 275 | 572 |
| GBH | 363 | 161 | 524 |
| IL | 334 | 166 | 500 |
| Noah | 135 | 97 | 232 |
| Spectre | 265 | 186 | 451 |

*Table 9 Number of hedges per language for each film*

Overall, the smallest number of hedges were observed in *Noah*, followed by *Amélie* and *Black Swan*. The first one could be explained by the fact that dialogues in this movie aim to reproduce a solemn register as the one found in the Bible, while other are filled with imperatives and follow a pressing rhythm. The *Black Swan* contains almost 400 annotations, probably simply due to the length of the file. As for *Amélie*, the reason for the distribution is rather challenging to interpret since, apart from the regular variables, the film was originally in French (the source of translation for Polish is not specified) and contains more comments of the narrator than an actual dialogue between the characters. *The Bridge of Spies* is decisively the richest in hedges. In this case, a highly likely reason for such a result may be the many scenes of political negotiations and court trials, filled with conditional sentences and evasive language in general. It is also the one with the biggest difference between the two versions. As for the other films, hedges are still more common in the English subtitles, although both *Amélie* and *Frozen* are almost matching.

| Film | PROP | PRF | CMT | SCH |
|---|---|---|---|---|
| Amelie | 220 | 20 | 64 | 14 |
| BS | 236 | 11 | 79 | 10 |
| BoS | 524 | 104 | 102 | 59 |
| Frozen | 368 | 28 | 123 | 6 |

| | | | | |
|---|---|---|---|---|
| GBH | 302 | 32 | 107 | 15 |
| IL | 265 | 64 | 107 | 14 |
| Noah | 113 | 22 | 54 | 7 |
| Spectre | 100 | 22 | 115 | 17 |

*Table 10 A distribution of the four tags in each of the language pairs*

Table 10 focuses again on the quantitative distribution of particular tags (the tag value pairs analysis in this view would be insignificant given the small number of singular instances, especially for attribution shields). Unsurprisingly, the passages tagged as PROP prevail in all cases, leaving rather small quantities for the entirety of speech act hedges. While the *Bridge of Spies* again dominates the list when it comes to the PROP column, interestingly enough, for the second most common tag in the film, CMT, the *Bridge of Spies* loses rather visibly in numbers to both *Frozen* and *Spectre*.

All charts and tables in this section offer some insights as to the number, distribution, and possible reasons for the presence of different types of hedges. It is not a big amount of data, but still large enough for some comparisons, especially considering the relatively small corpus. Consequently, it seems that the choice of film subtitles as a sample for testing the practicality of the proposed model was correct. This preliminary overview will be enriched by some more detailed linguistic analysis in the following chapter, and later summarised for some final commentary.

# 3 The Study and analysis

The final chapter of this thesis is entirely dedicated to presenting different approaches to the analysis of the proposed annotation scheme by its implementation on the corpus. The observations will be made on how the data influenced the scheme and vice versa, what problems emerged in relation to lexical and semantic information and, finally, all the statistical insights from the following and previous sections will be summarised and commented in 3.3. Having provided these more extensive data, the thesis will be concluded with an overview and final thoughts on the entire process.

## 3.1 Observations on the implementation

First, several observations can be done relating to the expressions analysed and annotated in the corpus, namely how they were classified and if the present scheme allowed for an appropriate decision in that matter. The general thoughts on the translation and interpretation aspect will open the discussion.

The contrastive quality of translated texts may differ greatly, and section 2.1 already alluded to the potential difficulties when working with translated subtitles specifically. As expected, given the particular register, there were quite a few problems with the proper interpretation of the original text, as well as with the decisions of translators, in some texts more than in others. When it comes to actual errors within the files analysed, these generally were left unmodified. Any translation- or quality-related corrections both exceeded and contradicted the objectives of this work (being the implementation of the scheme on 'natural', not curated data). Therefore, even if some fragments happened to be unclear or simply erroneous, despite diminishing the potential material for analysis, they were left untouched.

In numerous instances, the decision on whether to apply a specific tag, and, if so, which tag should be applied, was necessarily based on the personal interpretation of linguistic and metalinguistic material, including the information provided by acting and other non-linguistic elements of the source film. Hence, it must be said that the results of this and similar analysis may not be universally consistent. One of the strategies applied to tackle problematic examples was the use of DISC value for speech act hedges, designed to incorporate unspecified or ambiguous pragmatic effects of an expression. These difficulties were expected a priori, as the study of hedges and the general field of pragmatics are difficult to define strictly. One observation following the process of annotation is that, independently of the source corpus being mono- or multilingual, it would be most beneficial and precise for the research if the

annotation of hedges were to be performed by a native speaker. It would also facilitate the consideration of a broader context of 'utterances' given that some speech act strategies referred to different lines of dialogue, cf. in *The Grand Budapest Hotel*:

(2) #line 5 *In point of fact, the opposite is true*. - referring to the content of lines #1-4.
(3) The incidents that follow were described to me … exactly as I present them here … and in a wholly unexpected way. - distributed over the lines #15-18.

In fact, both *The Grand Budapest Hotel* and *The Bridge of Spies* turned out to be a source of many thought-provoking examples, both being quite rich in hedging forms of all types, due to the discursive narrative in the former, and the topics of espionage and negotiation in the latter. The *Bridge of Spies* especially contained many conditional clauses and a variety of modal verbs, and adverbs.

It is crucial to mention some examples of usual situations and the practices that were applied during the process:

- Source: *The Bridge of Spies:*

  (4) #sent_id = 915 # text = "*I... Mr. Waters* ***had it moved*** *over to Jack Elwes's office.*"

  The causative *have* construction in this case implies that the decision for whatever happened was taken by the subject of the second sentence, which is not the person speaking but 'Mr. Waters'. Having to do with referring an unpleasant information to the speaker, it was classified as self-mitigating shield.

  (5) #sent_id = 670 # text = **'*Byłbym*** *negocjatorem reprezentującym...*'

  *by-ł-by-m negocjator-em reprezentując-ym*

  be-M-COND-1SG negotiator.1SG-

  'I would be a negotiator representing'

  *-by* in Polish is usually called a mood particle because it is a hallmark of the so-called conditional mood in many Slavic languages. Immediately after a complementizer it is assumed to introduce the subjunctive mood. In that case, and many similar ones, it was classified as a PRF=DISC.

  (6) #sent_id = 1327 # text = '*– Ale czy jest* ***możliwość****... – Że moi ludzie mnie zastrzelą?*'

  *ale czy_jest możliwość że m-oi ludzie mnie za-strzel-ą*

  but be.PRS.1SG_there possibility that my people I.GEN FUT-shoot-3PL

  'But is there a possibility.. that my people will shoot me?'

  Again, could be classified as PRF, as the (modal) noun *possibility* (following the

classification of hedging elements by Fraser in Kaltenböck, ed al. 2010:23-24) but speaks rather to the degree of potential possible truth so the general commitment CMT=DISC tag was opted for.

- Generally, if a verb was split among a few verses (as it often was for the adapted treebank for Polish), all the elements received the hedge tag.
- Similes, or more exactly approximating elements such as *like*, were mostly treated as ATT hedges, interpreted in the role of category-assigning elements in the Lakoffean vision of hedges.

Another problematic aspect of translation and source of the files were the disparities between the content of the subtitles in general. In many cases, some scenes seemed to be cut from one language version and not the other. In others, seemingly random lines were omitted in translation (as it was mostly observed in the Polish texts). For example, in *The Bridge of Spies*, lines #797-#823 were only included in English. This was a serious problem for which the only reasonable solution was to cut the 'additional' elements from the original. It was an unfortunate loss of some part of the corpus, but significantly lesser with respect to the potential impossibility of contrasting the two language versions accurately line-by-line.

As for the explanation of these situations, there could be a few theories, already touched upon in the previous chapter (cf. 2.1). The decision obviously lied in the translator's competence, however, from the observations on what types of lines were usually missing, it could be supposed that they usually belonged to background characters and were not significant to the plot (*Black Swan*), were executed in a language different than English and were only subtitled in one version (*The Bridge of Spies*), or included background noises and directorial clues (*Noah*).

The usual observations as for the translation choices are numerous, but some that could be influential to this research are, for example, the usual omission of stylistic features of language in Polish subtitles: lines of non-English speakers that included some errors in the original, were normalised; on the other hand, the biblical style of monologues in *Noah* kept the correct characteristics.

As for the practical applications of the annotation scheme, the tags chosen applied quite well to the materials of the corpus. However, there were cases where, due to various reasons, the choice of how to approach the classification was not an obvious one. À propos those more straightforward, some examples are given below.

One of the more particular instances of reinforcement was first found in *The Grand Budapest Hotel*:

(7) #sent_id = 102 # text = *'I must confess, I did **myself** inquire about you.'*

The first-person singular reflexive pronoun here serves as a reinforcing element with the scope of assuring the listener of the intentions of the speaker. In this particular case, the exact same pronoun equivalent *sam* is used in Polish version. One thing that may be worth noticing is that in the LinES treebank used for the English files there are merely 200 tokens with non-empty value of Reflex (this particular token is annotated as follows: Case=Acc|Number=Sing|Person=1|PronType=Prs|Reflex=Yes), while even within the present, much smaller corpus other reflexive pronouns with the same scope has appeared a few times. Film subtitles are not, as said in 2.1, actually 'real life' examples of language use, so it would be expected of them to deviate from the proportions contained in the natural language data. However, it was also mentioned that film subtitles proved to actually mirror those well. In my opinion, given the goal-oriented communication that takes place within the films, the quantity of normally rare expressions can be expected to increase. It only makes it a more suitable collection of data for the present analysis, considering the need to identify numerous hedging elements to draw reliable conclusions.

Another crucial decision was made as for the conditional clauses. Consider the following example from *Amélie*:

(8) #sent_id = 247.# text = *"But **if it's me that says so**, it won't count. I'm senile."*

Fraser (in Kaltenböck, ed al. 2010:24) indicates conditional clauses as one of the hedging elements of English language, stating that, in fact, it insinuates a condition under which the utterance is being made. In this view, it does in a way influence the commitment of the speaker to the truth of the utterance, however the exact pragmatic effects desired are often uncertain. Thus, unless the concrete instance was clearer as to its meaning, the conditionals were tagged as CMT=DISC. Moreover, they were tagged in their entirety, as in *The Bridge of Spies*:

(9) #sent_id = 425 # text = 'And, I am sorry **if the way I put it offends you**. We need to know. What is Abel telling to you?'

In the sentence above, the conditional clause hedges the commitment of the speaker who decides to make an excuse in advance, in order not to negatively affect his own face in this conversation. Hence, the tag chosen for this fragment was CMT=NPL.

While the propositional hedges were usually easier to interpret, the speech act hedges, though fewer, were clearly more diverse. Especially the longer expressions could cause some doubt as for their role:

- In *Spectre*:

  (10) #sent_id = 867 # text = *"And **I should tell you** I've spoken with the Home Secretary."*

  The clause in bold was classified as PERF=MITS. As mentioned previously, to determine which expressions should be treated as hedges, Fraser's (in Kaltenböck 2010) list given in 1.3.2 was broadly considered as a reference. According to him, modal verbs hedge also in their regular function and not solely as hedged performatives. In this case, the self-serving mitigation tag was applied, since the speaker's intention seems to be to warn the interlocutor:

  (11) #sent_id = 135 # text = '- ***Możesz poczuć** lekkie... - Chryste!*'

  *Mo-żesz po-czuć lekk-ie Chryst-e*
  Can-2SG FUT-feel.INF slight Christ-VOC
  'You may feel a slight… Christ!

  The modal verb here does not work as a performative, but simply expresses a possible outcome of a situation. It could be treated as a sort of mitigation, warning the other person about painful procedure, but the more appropriate interpretation seemed to be that of simply stating the possibility. That is why CMT=ATT was used.

- In *The Grand Budapest Hotel*:

  (12) #sent_id = 679 # text = *"Mendl's again? **Precisely**."*

  Instead of its usual adverbial role of modifying the intensity of another element of a sentence where it would be considered simply a propositional reinforcing element, here *precisely* refers to a wider context, namely the previous sentence. In this case, both sentences were included in a single line; however, this was not always the case. To indicate that such comments refer to the larger context, the speech act tag CMT=RNF was used.

When it comes to Polish, one quite impactful decision was made early on, which turned out to work well in all the texts, although a similar interpretation would be rather unadjusted to English. Specifically, many speech act hedges were defined as PRF=NPL, similarly to the sentence below (from *The BoS*):

(13) #sent_id = 768 # text = **'Proszę to sprawdzić** i dostosować się do sytuacji.'

*pro-szę to sprawdzić I dostosować -się do sytuacj-i*

please-1SG it.ACC check.PFV and adjust.PFV -INF.REFL to situation-GEN

'Please, check that and adjust to the situation.'

If one were to remove *Proszę* from example thirteen, it would result in an unmitigated imperative. Including the verb, however, alleviates the imperative making it seem like a more polite request. In this sense, it saves the speakers face, so the most correct tag for the clause is PRF=NPL.

In other cases, the interpretation of a situation was rather unambiguous:

(14) #sent_id = 612 # text = *'Serge?* ***I'm afraid so****.'*
Prototypical example of CMT=MITS in *The Grand Budapest Hotel*.

*(15)* #sent_id = 126 # text = ***'Słyszałem****, że coś dla ciebie przygotował.'*

*słysz-ał-em że coś dla ciebie przygoto-wa-ł*

hear-PST-M.1SG that something for you.GEN prepare-PRF-M.2SG.PST

'I heard that he prepared something for you.'

The information in the line from *Spectre* is implied to come from other people and not conveyed with an absolute certainty, providing a typical example of SCH=ATT.

Similarly, to the conditional clauses, there were many examples of ambiguous expressions and metalinguistic comments, such as this instance of a rhetorical question from *The Bridge of Spies*:

(16) #sent_id = 888.text = 'Is that the greatest weapon we have in this Cold War.'

In these situations, again the precise scope of the comment was often difficult to determine, even though the utterance itself usually had to do something with the commitment to the information conveyed. For this reason, such examples were mostly classified as CMT=DISC as well, unless there was some sort of indication to the specific scope. In this case,

given the context of the line, it was judged as self-serving mitigation – being a hedged provocative comment to the situation.

(17) #sent_id = 72 # text = *'O czym **niby**?'*

*o czym niby*
about what.GEN supposedly.PART
'About what?'

The particle serving as comparing conjunction in this position has no referent to compare to achieve the status of attenuating hedge which may be considered more inherent. It creates an emphatic relation with the root pronoun w*hat*, expressing more intensively negation of knowledge and commitment to the information given previously in the conversation. The speaker denies plausibility and distances themselves from the message insinuated by the interlocutor. This example in *Spectre* was classified as SCH=MITS.

Speech act hedges, as previously stated, were more difficult to interpret for many reasons, one of which was very practical, i.e., the length of those expressions. With them spreading over multiple tokens, it was challenging to decide where the hedge should begin and end, as well as which was its scope. In some cases, different hedges were overlapping which forced a decision if they should be treated separately, or not. Cf. *The Bridge of Spies* line:

(18) #sent_id = 116 # text = "**If** you're not **merely** being polite, and you **must** tell me **if** that's the case."

Although the different metalinguistic comments and similar expressions were often treated as CMT=DISC, there were many examples where the choice of whether to include them was arbitrary, due to their peculiarities. Some of them are:

*The Grand Budapest Hotel:*

(19) #sent_id = 103 # text = "He's perfectly capable, of course, Monsieur Jean but we can't claim he's a first, or, **in earnest**, even second-rate concierge." – as CMT=DISC;

(20) #sent_id = 667 # text = 'Right, well, **be that as it may**, find him quick and make it snappy.' – as CMT=DISC;

(21) #sent_id = 802 # text = 'Well, **in point of fact**, I'm the executor of the estate.' – as PROP=ATT.

*Spectre:*

(22) #sent_id = 856 # text = ‘Shortest meeting I can remember. South Africans on board, **I take it?**’ – as CMT=DISC;

(23) #sent_id = 197 # text = ***‘Ponoć*** *niektóre wdowy żyją dosyć krótko*.’.- as SCH=MITS.

ponoć niektó-re wdow-y ży-ją dosyć krótko
apparently.PART some-F.3PL widow-NOM.PL live-3PL quite.ADV short.ADV
‘Apparently some widows live a rather short life.’

It is impossible to discuss all the problematic cases that were found in the files, but it is important to keep in mind that the examples above provide only an overview of the issues regarding the research. The two main points to be drawn from this summary are that obtaining a more trustworthy source of translated subtitles could strongly benefit this type of study, as the quality of the corpus diminished the quantity of data available, even if not dramatically. Secondly, a pragmatic-related annotation would certainly be more robust if performed by a native speaker. Many examples, though carefully analysed, could be easily disputed for their annotation by anyone with a different interpretation of the situational context. The statistical results of this interpretative process are explained in the next section.

## 3.2 Analysis

Following the initial outline of quantitative data that was given in 3.3.2, this section will cover a more in-depth analysis of different linguistic structures and correlations present in the annotated corpus. The first part discusses the frequencies of semantic and syntactic values of tokens in the function of hedges.

### 3.2.1 Ranked frequencies and intercorrelation between semantic and syntactic roles.

An aspect certainly worth exploring is the syntactic and lexical information of hedges identified within the annotated corpus, both for specific hedge types, and for the two language versions present. The analysis begins with the former, giving the lists of frequencies which contain ten most repeated POS and Deprel per hedge tags. Next, similar ranking is prepared for the most frequent lemmas. The following summaries have been prepared by gathering the necessary data in excel files, after having extracted them from the corpus with a simple formula, for instance all PROP* elements, all SCH=ATT/n elements, etc.

#### 3.2.1.1 POS and Deprel analysis

Table 11 contains the data pertaining to the occurrences of the PROP tag.

| POS Tags | | | Polish | | | English | | |
|---|---|---|---|---|---|---|---|---|
| **RANK** | **UPOSTAG** | **$f_i$** | **RANK** | **DEPREL** | **$f_i2$** | **RANK** | **DEPREL** | **$f_i3$** |
| 1 | ADV | 1069 | 1 | advmod | 1043 | 1 | advmod | 1043 |
| 2 | DET | 292 | 2 | det | 260 | 2 | det | 260 |
| 3 | PRON | 284 | 3 | advmod:emph | 241 | 3 | case | 192 |
| 4 | PART | 277 | 4 | case | 192 | 4 | root | 178 |
| 5 | ADP | 210 | 5 | root | 178 | 5 | nsubj | 126 |
| 6 | NOUN | 195 | 6 | nsubj | 126 | 6 | mark | 112 |
| 7 | ADJ | 160 | 7 | mark | 112 | 7 | obj | 111 |
| 8 | SCONJ | 101 | 8 | obj | 111 | 8 | amod | 99 |
| 9 | VERB | 69 | 9 | amod | 99 | 9 | obl | 67 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 10 | AUX | 19 | 10 | obl | 67 | 10 | fixed | 43 |

*Table 11 Ranked POS Tags and DEPREL for PROP hedges in the two languages.*

As one can see, the top of the chart is dominated by shorter POS. Adverbs occupy the first position with an incomparable advantage of about seven hundred over determiners, pronouns, particles, and even other labels. Especially those four parts of speech function mostly as 'additional' syntactic information, merely enhancing the trees, but not providing the essential conditions for the existence of a clause or a sentence. The Deprel tag, subdivided into English and Polish column due to different classifications offered in the two treebanks, not unexpectedly contains the same types of relation within the first 10 presented in both languages, with just slight differences in the order. All the relations of adverbial modality and emphasis can be more often associated with adjuncts with respect to regular arguments of a sentence. At least for syntactic relations, this aligns well with the role of propositional hedges being modifiers of the elements of a single sentence. Those most common elements function usually as single-word tokens, which corresponds to the results presented at the end of the previous chapter as for the quantity of single and multiword expressions per tag.

The following data refer to the PRF tag, one on the borderline between single and multiword expressions, whose distribution with respect to POS and Deprel is once again almost identical in the two languages.

POS Tags Polish English

| RANK | UPOSTAG | $f_i$ | RANK | DEPREL | $f_i2$ | RANK | DEPREL | $f_i3$ |
|---|---|---|---|---|---|---|---|---|
| 1 | VERB | 361 | 1 | root | 228 | 1 | root | 228 |
| 2 | AUX | 275 | 2 | aux | 184 | 2 | aux | 184 |
| 3 | PRON | 139 | 3 | xcomp | 104 | 3 | xcomp | 104 |
| 4 | PART | 30 | 4 | nsubj | 96 | 4 | nsubj | 96 |
| 5 | ADV | 23 | 5 | cop | 40 | 5 | cop | 40 |
| 6 | NOUN | 22 | 6 | obj | 38 | 6 | obj | 38 |
| 7 | ADJ | 11 | 7 | advmod | 29 | 7 | advmod | 29 |
| 8 | ADP | 9 | 8 | aux:cnd | 24 | 8 | conj | 20 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9 | DET | 3 | 9 | conj | 20 | 9 | mark | 10 |
| 10 | SCONJ | 3 | 10 | advmod:neg | 12 | 10 | aux:pass | 9 |

*Table 12 Ranked POS Tags and DEPREL for PRF hedges in the two languages.*

The PRF hedges, being a category concerning the modal verbs and performatives, is clearly dominated by verbal and auxiliary elements, with relatively fewer pronouns, probably being a part of the modal clause. Accordingly, the relational tags of the elements include the root and auxiliary positions in the sentence, along with some complement and subject positions, all once again pointing to clauses dominated by modal verbs.

Finally, the data for commitment hedges provides a slightly more varied picture.

| POS Tags | | | Polish | | | English | | |
|---|---|---|---|---|---|---|---|---|
| **RANK** | **UPOSTAG** | **$f_i$** | **RANK** | **DEPREL** | **$f_i2$** | **RANK** | **DEPREL** | **$f_i3$** |
| 1 | VERB | 912 | 1 | root | 616 | 1 | root | 616 |
| 2 | PRON | 755 | 2 | nsubj | 613 | 2 | nsubj | 613 |
| 3 | PART | 367 | 3 | mark | 380 | 3 | mark | 380 |
| 4 | AUX | 342 | 4 | advmod | 327 | 4 | advmod | 327 |
| 5 | SCONJ | 337 | 5 | advcl | 203 | 5 | advcl | 203 |
| 6 | ADV | 296 | 6 | obj | 163 | 6 | obj | 163 |
| 7 | NOUN | 249 | 7 | aux | 151 | 7 | aux | 151 |
| 8 | ADJ | 118 | 8 | cop | 100 | 8 | cop | 100 |
| 9 | DET | 102 | 9 | det | 93 | 9 | det | 93 |
| 10 | ADP | 80 | 10 | xcomp | 90 | 10 | parataxis | 90 |

*Table 13 Ranked POS Tags and DEPREL for CMT hedges in the two languages.*

In Table 13, the main portion of the data consists of verbs and pronouns in root and subject position. Although not as clear an information as in the case of hedged performatives, it must be underlined that the instances of plausibility shields are also included in many clauses. It is enough to mention the expressions *seem to, sound like,* etc. Furthermore, in the first section of this chapter I mentioned that many longer and not precisely classifiable expressions often received a tag of CMT=DISC, automatically increasing chances of the results given above.

Moving forward, apart from particles and auxiliaries occurring once more in a higher position in the Table 13, there is also quite a significant percentage of subordination involved, with subordinating conjunctions on the fifth position and marker relation type with the third rank. This could be attributed, for example, to the consistent categorization of conditional clauses with this tag. Other elements such as adverbs, nouns, and adjectives in object, complement and modifying positions are also quite numerous, although much less common in proportion to the two major categories. This is due to the variety and scale of the CMT tag in general, being second only to PROP in the total number of annotated items, though much more diverse.

The last table in this part of the analysis concerns the smallest group of hedging devices, namely those taking the SCH tag.

| POS Tags | | | Polish | | | English | | |
|---|---|---|---|---|---|---|---|---|
| **RANK** | **UPOSTAG** | **$f_i$** | **RANK2** | **DEPREL** | **$f_i2$** | **RANK** | **DEPREL** | **$f_i3$** |
| 1 | VERB | 135 | 1 | root | 100 | 1 | root | 100 |
| 2 | PRON | 94 | 2 | nsubj | 81 | 2 | nsubj | 81 |
| 3 | NOUN | 47 | 3 | case | 26 | 3 | case | 26 |
| 4 | AUX | 42 | 4 | mark | 26 | 4 | mark | 26 |
| 5 | ADP | 27 | 5 | advmod | 22 | 5 | advmod | 22 |
| 6 | PART | 26 | 6 | obj | 22 | 6 | obj | 22 |
| 7 | ADV | 17 | 7 | aux | 19 | 7 | aux | 19 |
| 8 | DET | 16 | 8 | nmod | 14 | 8 | nmod | 14 |
| 9 | PROPN | 16 | 9 | det | 13 | 9 | det | 13 |
| 10 | SCONJ | 14 | 10 | xcomp | 11 | 10 | xcomp | 11 |

*Table 14 Ranked POS Tags and DEPREL for SCH hedges in the two languages.*

Offering less variety in the types of expressions examined the attribution shields are also dominated by tokens with verb, pronoun, noun, and auxiliary POS. The main syntactic relation to be observed is that of root and nsubj position. The following Deprel tags are rather balanced in number, with case and subordination markers with the highest position. In this case, given the size of the sample, it is difficult to advance hypotheses regarding the distribution of the tag, but it is interesting to see that, though the distribution between specific subcategories of shields

in Polish and English is quite different, the syntactic relations between them is almost identical, given the data in the fifth and eight columns.

#### 3.2.1.2 Ranked frequencies of lemmas per tag

Having observed the distribution of syntactic annotation for the four main types of hedges, it is interesting to see which lemmas exactly are the most commonly used as different hedging devices.

PROP tag

| | English | | | Polish | | |
|---|---|---|---|---|---|---|
| Rank | Lemma | f1 | rel. f1 | Lemma | f2 | rel.f2 |
| 1 | just | 165 | 9,48% | tylko - *only* | 96 | 10,01% |
| 2 | very | 84 | 4,82% | jak - *how* | 71 | 7,40% |
| 3 | like | 83 | 4,77% | tak – *yes/so* | 44 | 4,59% |
| 4 | of | 73 | 4,19% | bardzo - *very* | 43 | 4,48% |
| 5 | so | 67 | 3,85% | jakiś - *some* | 32 | 3,34% |
| 6 | a | 56 | 3,22% | nawet - *even* | 28 | 2,92% |
| 7 | only | 53 | 3,04% | troche – *a bti* | 19 | 1,98% |
| 8 | some | 46 | 2,64% | coś - *something* | 18 | 1,88% |
| 9 | really | 41 | 2,35% | może - *maybe* | 16 | 1,67% |
| 10 | any | 35 | 2,01% | zbyt - *too* | 15 | 1,56% |

*Table 15 Lemma frequencies for the PROP tag*

To start with, I will discuss the lemmas' frequency list for all the propositional hedges for both languages. Although, as seen from the relative frequency, the distribution for English most common lemmas is slightly more balanced than that of Polish, it is striking how well both lists match in terms of semantic equivalence. Ignoring any changes in meaning that might be attributed by the position in the text, there are several correspondences between the inherent meanings of hedges in the English and Polish corpus: *just* (1) equivalent to *tylko* (1), *very* (2) to *bardzo* (4), *like* (3) to *jak* (2). Other correspondences might be mentioned, but the rest of the elements could have more matches within their basic meaning, so they should not be paired out of context. It suffices to say that almost all of the lemmas in the two tenths could be matched to each other.

PROP=ATT tag

| | English | | | Polish | | |
|---|---|---|---|---|---|---|
| Rank | Lemmma | f1 | rel. f1 | Lemma | f2 | rel.f2 |
| 1 | just | 155 | 16,76% | tylko - *only* | 74 | 13,63% |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | like | 81 | 8,76% | jak - *how* | 70 | 12,89% |
| 3 | of | 53 | 5,73% | jakiś - *some* | 32 | 5,89% |
| 4 | a | 47 | 5,08% | troche – *a bit* | 19 | 3,50% |
| 5 | some | 45 | 4,86% | może - *maybe* | 16 | 2,95% |
| 6 | as | 30 | 3,24% | chyba - *maybe* | 22 | 4,05% |
| 7 | little | 28 | 3,03% | ktoś - *someone* | 11 | 2,03% |
| 8 | only | 24 | 2,59% | prosty - *simple* | 11 | 2,03% |
| 9 | bit | 23 | 2,49% | tak - *so* | 10 | 1,84% |
| 10 | any | 18 | 1,95% | niektóry - *certain* | 10 | 1,84% |

*Table 16 Lemma frequencies for PROP=ATT hedges*

The ranking for the largest subtype of propositional hedges, namely PROP=ATT, offers very similar results to the first table, at least for English. When it comes to Polish, the first two elements are again equivalent to the first and second for English respectively, but the rest of the list offers more diversity. Some of the most curious are *może* (*maybe*, modal particle), *prosty* (in this case: *simple*; probably deriving from the frequent attenuating expressions *po prostu* meaning *simply*), *chyba* (also modal particle equivalent to *maybe*), and *ktoś* (*someone/somebody*). Two new quantitative elements present in the English list are *little*, and *bit*, which could correspond mostly to *trochę* (ranked four; used both as a particle and a determiner, meaning *slightly, a bit, some, a little*).

PROP=RNF tag

| | English | | | Polish | | |
|---|---|---|---|---|---|---|
| Rank | Lemma | f1 | rel. f1 | Lemma | f2 | rel.f2 |
| 1 | very | 80 | 10,93% | bardzo - *very* | 40 | 11,02% |
| 2 | so | 64 | 8,74% | tak - *so* | 34 | 9,37% |
| 3 | really | 38 | 5,19% | nawet - *even* | 26 | 7,16% |
| 4 | too | 34 | 4,64% | tylko - *only* | 21 | 5,79% |
| 5 | even | 31 | 4,23% | zbyt - *too* | 15 | 4,13% |
| 6 | no | 30 | 4,10% | nikt - *nobody* | 14 | 3,86% |
| 7 | only | 29 | 3,96% | naprawdę - *really* | 14 | 3,86% |
| 8 | all | 21 | 2,87% | żaden – *no(one)* | 11 | 3,03% |
| 9 | of | 19 | 2,60% | całkowicie - *completely* | 10 | 2,75% |
| 10 | much | 18 | 2,46% | nic - *nothing* | 9 | 2,48% |

*Table 17 Lemma frequencies for PROP=RNF hedges*

Table 14 presents the ranking of reinforcing lemmas. Differently from what we have seen in the preceding Tables, the results for the two languages are quite different in this case. However, there are some lemmas repeating from Table 15, such as *very, so, really* (1-3) and

*only* (7) for English. The distance of four positions between *really* and *only*, and the fact that *only* is present in the first classification instead of *too*, confirms that *only* can be used both in RNF and ATT contexts. In Polish, *bardzo, tak*, *nawet* (1-3; meaning *very, so, even*), and *zbyt* (5; *too*) also appeared in the initial PROP list. Interestingly, the high-ranking lemmas for both languages are again mostly equivalent with the exception of Polish *tylko* situated on the fourth position, corresponding to the seventh *only.*in the English list.

PROP=EV tag

| | English | | | Polish | | |
|---|---|---|---|---|---|---|
| Rank | Lemma | f1 | rel. f1 | Lemma | f2 | rel.f2 |
| 1 | thing | 17 | 20,48% | coś - *something* | 11 | 20,00% |
| 2 | something | 16 | 19,28% | ktoś - *someone* | 3 | 5,45% |
| 3 | person | 10 | 12,05% | człowiek - *man* | 3 | 5,45% |
| 4 | one | 4 | 4,82% | osoba - *person* | 3 | 5,45% |
| 5 | someone | 3 | 3,61% | sprawa - *case* | 3 | 5,45% |
| 6 | it | 3 | 3,61% | się - *oneself* | 3 | 5,45% |
| 7 | the | 3 | 3,61% | ubezpieczyć – *insure/assure* | 2 | 3,64% |
| 8 | article | 3 | 3,61% | przez - *through* | 2 | 3,64% |
| 9 | wrinkle | 3 | 3,61% | pański - *yours* | 2 | 3,64% |
| 10 | insure | 2 | 2,41% | klient - *client* | 2 | 3,64% |

*Table 18 Lemma frequencies for PROP=EV hedges*

Finally, the last and least numerous of propositional subtypes EV is presented. The semantic equivalents *thing/something* and *coś* have the highest relative frequency (19-20%), not only among EV hedges, but in general. *Ktoś* (2) corresponds to *someone* (5), with the same frequency, and the third elements, *person* and *człowiek,* are also semantically equivalent (although this time the frequency is increased by *osoba* in fifth position with the same meaning). The rest of the list is rather different this time, but this difference can be attributed to the fact that, apart from *it* and *the* for English, all the elements are nominal (and thus belonging to an open class), whereas in the previous classifications adverbs and particles were more common (and more easily equivalent). It is the result of the typical objective of an evasion strategy which is evident.in the assortment of hedging elements the strategy was attributed to: in order to avoid speaking directly and conveying some type of information, different ambiguous nominal elements are used. These sorts of tokens offer more possibilities for translation, so the equivalence understandably decreased.

The striking similarities in the most frequent lemmas for all the PROP categories support the prediction that propositional hedges will most probably be much more transferable between

languages (at least in the case of languages that are closely related such as English and Polish), as their scope is limited to a single sentence and modifies its contents in a rather universal manner. The respective results for the three values for PROP also confirmed the expectations that one might have had for the distribution and type of lemmas used in this type of translated texts, with elements affecting modal aspects for ATT and RNF, and more ambiguous nominal elements with evasive function.

PRF tag

| | English | | | Polish | | |
|---|---|---|---|---|---|---|
| Rank | Lemma | f1 | rel. f1 | Lemma | f2 | rel.f2 |
| 1 | be | 55 | 10,36% | prosić - *please* | 35 | 9,97% |
| 2 | must | 53 | 9,98% | móc - *can* | 34 | 9,69% |
| 3 | may | 38 | 7,16% | musieć – *have to* | 29 | 8,26% |
| 4 | I | 37 | 6,97% | by – *in order to* | 24 | 6,84% |
| 5 | you | 29 | 5,46% | być - *be* | 24 | 6,84% |
| 6 | might | 21 | 3,95% | się - *oneself* | 11 | 3,13% |
| 7 | could | 18 | 3,39% | to - *it* | 10 | 2,85% |
| 8 | would | 16 | 3,01% | nie - *no* | 9 | 2,56% |
| 9 | should | 16 | 3,01% | usiąść - *sit* | 5 | 1,42% |
| 10 | we | 16 | 3,01% | on - *he* | 5 | 1,42% |

*Table 19 Lemma frequencies for PRF tag*

The next table (19) presents the data for the first of speech act hedges, i.e., PRF. In this case, there are no longer many similarities between English and Polish hedges. The former includes, as one could expect, mostly modal verbs, along with the copula *be*, as well as three common personal pronouns *I, you,* and *we*. These results correspond to the results of POS tags for PRF from the previous paragraphs. The Polish results are also what could be expected, although less clear, simply for the lack of exact equivalents for different modal verbs that English offers. In Polish, the most common lemma is *prosić* (*please/ask*; used in all the PRF=NPL constructions as a face-saving attenuation for an imperative). Next there are two verbs with which a weaker and a stronger modality can be expressed: *móc* (*can/ be able to*) and *musieć* (*have to*). In the third position there is *by,* the mood/subjunctive particle already mentioned in examples in 4.1, followed by impersonal reflexive pronoun *się*, copula *być* (*to be*), negative particle *nie*, as well as demonstrative pronoun *to*, third person singular masculine pronoun *on*, and finally, in the ninth position a verb *usiąść* (*to sit*), used for attenuating an order to sit.

CMT tag

| | English | | | Polish | | |
|---|---|---|---|---|---|---|
| Rank | Lemma | fi1 | rel. Fi1 | Lemma | fi2 | rel.fi2 |
| 1 | I | 314 | 13,44% | być - *be* | 82 | 6,44% |
| 2 | be | 166 | 7,10% | nie - *no* | 58 | 4,56% |
| 3 | if | 164 | 7,02% | jeśli - *if* | 101 | 7,93% |
| 4 | you | 113 | 4,84% | się - *oneself* | 41 | 3,22% |
| 5 | not | 109 | 4,66% | wiedzieć - *know* | 32 | 2,51% |
| 6 | think | 91 | 3,89% | może - *maybe* | 31 | 2,44% |
| 7 | do | 75 | 3,21% | myśleć - *think* | 28 | 2,20% |
| 8 | it | 66 | 2,82% | móc - *can* | 27 | 2,12% |
| 9 | to | 50 | 2,14% | to - it | 24 | 1,89% |
| 10 | the | 48 | 2,05% | oczywiście – *of course* | 23 | 1,81% |

*Table 20 Lemma frequencies for CMT tag*

For the plausibility shields CMT, the lists are once more less equivalent for the two languages; however, they cover, as expected, the most common expressions that were given that label, such as conditional clauses. The most frequent element for English was the personal pronoun *I*, followed by the copula *be*, the subordinating conjunction *if*, and a few more common pronouns, determiners, and verbs. From my subjective observations, the most common plausibility shields in the presented corpus were, and various metalinguistic comments. During the presentation of the statistical data in the second chapter, a few predictions were given. Expressions such as *I think* and conditional clauses are best suited for the role of.plausibility hedges. From my subjective observations, they were in fact quite common in the corpus, so I expected them to be the most numerous examples of CMT tag. Given the numbers presented in Table 20 for the conjunction *if* and copula *be*, those predictions were confirmed in English. The situation is similar in Polish. Within the ten most common lemmas there are a few verbs (again copula *być*, but also *widzieć = to see*, *myśleć = to think, móc = can, be able to*), along *jeśli* equivalent to *if*, the reinforcing adverb *oczywiście* (*of course*), and *nie, się, może,* and *to* that we have seen in Table 19. Given the results regarding the POS and Deprel tags, the lemmas' lists may not be in exactly the predictable order (one might expect verbs to dominate), but they still match well in the two languages.

SCH tag

| | English | | | Polish | | |
|---|---|---|---|---|---|---|
| Rank | Lemma | fi1 | rel. Fi1 | Lemma | fi2 | rel.fi2 |
| 1 | I | 25 | 8,17% | mówić - *speak* | 15 | 17,24% |
| 2 | say | 24 | 7,84% | być - *be* | 9 | 10,34% |
| 3 | be | 23 | 7,52% | jak - *how* | 5 | 5,75% |

| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | you | 22 | 7,19% | to - *it* | 5 | 5,75% |
| 5 | to | 16 | 5,23% | wiedzieć - *know* | 4 | 4,60% |
| 6 | the | 10 | 3,27% | on - *he* | 4 | 4,60% |
| 7 | think | 8 | 2,61% | ponoć - *apparently* | 3 | 3,45% |
| 8 | mean | 8 | 2,61% | słyszeć - *hear* | 3 | 3,45% |
| 9 | know | 7 | 2,29% | powiedzieć - *say* | 3 | 3,45% |
| 10 | as | 7 | 2,29% | pan - *you* | 3 | 3,45% |

*Table 21 Lemma frequencies for SCH tag*

At last, the SCH tag results are given, with probably the most equivalent elements out of the three speech act tags. Within the first ten lemmas are those corresponding to the verbs *say* (*mówić*), *be* (*być*), *know* (*wiedzieć*), as well ad *think* and *mean* for English and *słyszeć* (*to hear*) and *powiedzieć* (*tell*) in Polish. The following five for each language are covered by determiners, particles, and pronouns. Once again, SCH hedges constitute the least represented in the corpus; this holds for both languages and is in line with the previous results.

Compared to the PROP lists, there is a greater mismatch in the frequency lists for speech act hedges in the two languages, although they still express similar ideas in comparable ways (owing probably to the vicinity of languages). The attribution shield, while being the least numerous, appears to offer the best equivalency of structures in types and number for both languages, probably because of its attributive scope being rather visible and well transferable.

### 3.2.2 Average hedge length distributions

The second part of this section presents a comparison of another interesting aspect of the results of annotation. In the last section of the previous chapter, I compared the occurrences of one- and multi-word hedges of different types. I therefore decided to observe the average lengths of hedges of the four types, and asked myself whether the mean length of a hedge differs between the two languages. To prepare the subsequent calculations, for each positional value of a tag (Position=Initial, N=x) I calculated the number of the hedges within the respective lists. For each value, I multiplied the number of hedges by the corresponding length, giving the overall number of words for each hedge length. After calculating the number of hedges for each category (items), the number of all hedging tokens (total), and adding the number of one-word hedges per category from previous calculations, the relative frequency of various length hedges, I extracted their average length and mode.

#### 3.2.2.1 Hedge length per language

The first table presents the results for English:

| tag fi | no. hedges | length | no. words | rel. fi |
|---|---|---|---|---|
| 2 | 2 | 12 | 24 | 0,99% |
| 5 | 3 | 11 | 33 | 1,36% |
| 10 | 5 | 10 | 50 | 2,07% |
| 16 | 6 | 9 | 54 | 2,23% |
| 39 | 23 | 8 | 184 | 7,61% |
| 73 | 34 | 7 | 238 | 9,84% |
| 124 | 51 | 6 | 306 | 12,65% |
| 205 | 81 | 5 | 405 | 16,74% |
| 337 | 132 | 4 | 528 | 21,83% |
| 578 | 241 | 3 | 723 | 29,89% |
| 1101 | 523 | 2 | 1046 | 43,24% |
| 1318 | 1318 | 1 | 1318 | 54,49% |
| 2419 | 2419 | items | mean | mode |
| 4916 | 4916 | total | 2,029351 | 1 |

*Table 22 Hedge length values for English*

Although the average length turned out to be merely two tokens (with mode at only 1), it is interesting to see that the longest hedges for English could reach even 12 tokens.

Next, there are the respective results for Polish:

| tag fi | no. hedges | length | no. words | rel. fi |
|---|---|---|---|---|
| 2 | 2 | 9 | 18 | 1,18% |
| 10 | 8 | 8 | 64 | 4,20% |
| 20 | 10 | 7 | 70 | 4,59% |
| 47 | 27 | 6 | 162 | 10,63% |
| 95 | 48 | 5 | 240 | 15,75% |
| 168 | 73 | 4 | 292 | 19,16% |
| 319 | 151 | 3 | 453 | 29,72% |
| 554 | 235 | 2 | 470 | 30,84% |
| 970 | 970 | 1 | 970 | 63,65% |
| 1524 | 1524 | items | mean | mode |
| 2812 | 2812 | total | 1,797244 | 1 |

*Table 23 Hedge length values for Polish*

First of all, the hedges in general were shorter than in English (reaching the maximal length of 9 tokens), yest the average length is only slightly lower, with mode still being 1 token length.

The maximum values for each language could be expected as already in Section 3.1 I made some comments on the characteristics of English and Polish, the latter having tendentially longer words but shorter sentences. What is slightly surprising about these results it that, despite some differences, their average hedge length seems to match, furthermore, the average length of two tokens only, while there were numerous examples of multiword hedges for all the three speech act types. Still, it is important to remember the number of propositional hedges within the corpus, which certainly influenced the calculations enough to achieve these results.

#### 3.2.2.2 Hedge length per tag

The table beneath present the same estimations for the all the propositional tags.

| tag fi | no. hedges | Length | no. words | rel. fi |
|---|---|---|---|---|
| 6 | 6 | 5 | 30 | 1,32% |
| 12 | 6 | 4 | 24 | 1,06% |
| 76 | 64 | 3 | 192 | 8,45% |
| 327 | 251 | 2 | 502 | 22,10% |
| 1944 | 1944 | 1 | 1944 | 85,60% |
| | | PROP | | |
| 2271 | 2271 items | | mean | mode |
| | | total | | |
| 2704 | 2704 words | | 1,185381 | 1 |

*Table 24 Hedge length values for PROP tag*

Predictably, both the mean and the mode value are of approximately one token, with the 85% of all the PROP hedges consisting of one word. Some more interesting information may be extracted from the following calculations for PRF, CMT, and SCH types.

| tag fi | no. hedges | length | no. words | rel. fi |
|---|---|---|---|---|
| 3 | 3 | 7 | 21 | 6,75% |
| 17 | 14 | 6 | 84 | 27,01% |
| 37 | 20 | 5 | 100 | 32,15% |
| 76 | 39 | 4 | 156 | 50,16% |
| 160 | 84 | 3 | 252 | 81,03% |
| 279 | 119 | 2 | 238 | 76,53% |
| 32 | 295 | 1 | 295 | 94,86% |
| | | PRF | | |
| 311 | items | | mean | mode |
| | | total | | |
| 904 | 904 words | | 3,684887 | 1 |

*Table 25 Hedge length values for PRF tag*

The performative hedges already offer quite a different outcome. Although the mode once again is of one token, the average length for this category reaches three, almost four tokens. Given the more even distribution of hedge frequencies for PRF, this result seems plausible, taking into consideration the general expectation for speech act hedges to be multiword expressions.

| tag fi | no. hedges | | length | no. words | rel. fi |
|---|---|---|---|---|---|
| 2 | 2 | | 12 | 24 | 0,16% |
| 4 | 2 | | 11 | 22 | 0,16% |
| 9 | 5 | | 10 | 50 | 0,41% |
| 16 | 7 | | 9 | 63 | 0,58% |
| 44 | 28 | | 8 | 224 | 2,30% |
| 82 | 38 | | 7 | 266 | 3,12% |
| 143 | 61 | | 6 | 366 | 5,01% |
| 236 | 93 | | 5 | 465 | 7,64% |
| 374 | 138 | | 4 | 552 | 11,34% |
| 575 | 201 | | 3 | 603 | 16,52% |
| 922 | 347 | | 2 | 694 | 28,51% |
| 295 | 295 | | 1 | 295 | 24,24% |
| | | | CMT | | |
| 1217 | 1217 | items | | mean | mode |
| | | | total | | |
| 3658 | 3658 | words | | 2,977814 | 2 |

*Table 26 Hedge length values for CMT tag*

The CMT hedges again differ slightly from the previous results, with mode at two and average length at almost three tokens. Though the plausibility shields achieve longer hedges in general, arriving at the maximal value of 12, the distribution is surely more concentrated between values 1-4. Given the balance between many shorter expressions such as *I think*, and the long comments and clauses included in this category, the average of three appears to be understandable.

| tag | no.hedges | length | no.words | rel. fi |
|---|---|---|---|---|
| fi | | | | |
| 1 | 1 | 11 | 11 | 0,69% |
| 1 | 0 | 10 | 0 | 0,00% |
| 2 | 1 | 9 | 9 | 0,69% |
| 5 | 3 | 8 | 24 | 2,08% |
| 8 | 3 | 7 | 21 | 2,08% |
| 11 | 3 | 6 | 18 | 2,08% |
| 21 | 10 | 5 | 50 | 6,94% |
| 43 | 22 | 4 | 88 | 15,28% |
| 86 | 43 | 3 | 129 | 29,86% |

| | | | | |
|---|---|---|---|---|
| 127 | 41 | 2 | 82 | 28,47% |
| 17 | 17 | 1 | 17 | 11,81% |
| 144 | 144 | SCH items | mean | mode |
| 457 | 457 | total words | 3,118056 | 3 |

*Table 27 Hedge length values for SCH tag*

Finally, the SCH hedges, arriving at the total of eleven tokens per hedge, turned out to reach the mean and mode for the hedge length of three. This result, similar to that of other speech act hedges, is however most likely to change with a bigger sample of material.

## 3.3 Summary and commentary

In the previous section I presented a detailed analysis of the annotated corpus data. The aim was to observe comparable characteristics such as hedge length, tags, and lemmas' frequencies, to be able to observe the correlations and differences between the different types of inter' and cross-linguistically. The results mostly confirmed the expectations given in the second chapter (2.4, 2.5).as to the potential outcomes describing the general approach to the analysis and, primarily, the annotation scheme for the phenomenon.

The propositional hedges were already established as the most numerous categories of the hedges within the corpus. From the summary in the previous chapter, it was clear that they mostly consist of one-word expressions and are decisively more common in English. The more in-depth analysis in Section 3.2, showed that they in fact mostly consist of single tokens, with the mean length of 1,19. Interestingly, despite occurring less frequently in Polish, they proved to be almost equivalent in terms of most recurrent lemmas, indeed, a bigger difference in the lemmas used could only be observed for the ones with the value EV, the reason being a wider range of possible translations for nominal elements with generalizing function. The POS and Deprel tag comparison also revealed the majority of short element tags like ADV and DET in modifying positions.

The most problematic tag to analyse was the one representing attribution shields, SCH. Because it was the least frequent hedging device in the corpus, the calculations' results can be expected to slightly change if a similar study was to be applied to a bigger corpus. A curious fact discovered in the first analysis was that the most common Deprel tags for both languages mirrored exactly. That, together with the results of lemmas' frequencies, seems to suggest that this type of hedges is the most transferable between languages, at least in the case of the two examined. The average length of SCH expressions offered an unsurprising result of three tokens; however, the results of this particular analysis appear most prone to error due to potentially insufficient amount of data.

With regard to the second most numerous category, namely that of CMT hedges, they also seemed to present in the words and expressions suggested in the literature. From the corpus data, it is known that there is quite a variety of expressions included within this tag. They range from long conditional clauses, through shorter ambiguous comments often attributed value DISC, to two-word typical plausibility attenuating shields. Being more balanced when it comes to the number of occurrences in two languages, they were expected to reflect this information in the analysis. In fact, the POS and Deprel tag distribution mirrored that of typically longer

expressions, containing many clause-building elements as verbs and pronouns in root and subject positions. Similarly, in lemmas' ranked lists the different verbs and pronouns prevailed, generally offering at least partial equivalence between English and Polish. The latter offer more different verb versions and a strong position of reinforcing adverb *oczywiście*. The conditional subordinating element (eng. *if*, pl. *jeśli*) was also quite common. Finally, the hedge length approximation showed that even though CMT hedges can appear in a long multiword form, they mostly range from one to four-word expressions, averaging at almost three tokens.

As for the hedged performative tag PRF, given the rather well-defined variety of expressions that it may include, it was expected to present balanced analysis' results as well. While being only slightly more common in English, is also turned out to correspond to Polish data when it comes to the distribution of Deprel tags and frequency of lemmas. The recurring syntactic elements described as PRF consisted in verbs, pronouns, and particles which in English corresponded to various modal verbs and personal pronouns expected in hedged performatives. Unaligned lemmas' results for Polish should not be treated a significant difference to English results, given the lack of diversity in Polish range of modal verbs, represented mostly by two forms, indeed present on top of the list. Concerning hedge length for PRF tag, it showed probably the most even distribution between its one-token and maximum of seven-token realisations, averaging on 3,7, however, with the mode of one token. Given the fact that single token hedges in general are most common, as seen in 3.3.2, without a challenging unimodal distribution of lengths, these results are proving acceptable.

All the results summarised above were preceded by an outline of problematic relative to the actual application of selected tags onto the corpus. Some decisions that had to be made to normalise the method, like that of PRF=NPL tag only present in Polish attenuated imperatives, or that of labelling ambiguous metalinguistic expressions as CMT=DISC, were better grounded and seem appropriate for the scheme proposed, while other singular examples cause more doubt. The main point that I have been stressing in the introductory Section 3.1, are that given the characteristics of hedging as a pragmatic phenomenon similar study would require annotation to be performed by a native speaker. If such approach is impossible, the focus should be made on the quality of the texts chosen, having in mind the limits inherent to parallel and comparable corpora.

Regardless of all the aspects undermining the present study, it proved to achieve comparable and reliable results. All the discussed characteristics of the texts of which the corpus is composed, the different modalities and versatility of hedging devices, and finally, those natural to English and Polish, proved to match in the results of annotation. Such outcomes

of a tentative study on the proposed annotation scheme confirms its potential and seem to be applicable on a larger scale, also including other languages.

# 4 Conclusions

In this last chapter I will summarize the contents and results of the thesis in view of the utility and applications of the study outcomes. The main goal of the present thesis was to propose a versatile and adaptable annotation scheme for the phenomenon of hedging. As a tool, it was supposed to be not specialised, but applicable to any language and text type. To be able to take into consideration all the possible variables for the success of such scope, first an overview of studies on corpora (1.1), and computational linguistics (1.2) was provided, followed by a presentation of hedging (1.3), with the focus on most important studies and classification. Next, a more practical approach to the study could commence.

Among the many possible sources of data, a parallel corpus was chosen. The motivation for such decision was the possibility of immediate comparison of functionality of the scheme in two different languages. As most of the literature with regards to hedging focuses on English, it was resolved that English source material was to be contrasted with Polish, being one of the languages where similar research is scarce. Opting for a corpus of film subtitles was a result of an estimation of the source of easily comparable texts possibly richest in hedging devices, as compared to literary texts. Given the subtitles' characteristics of merging the qualities of literary and spoken language, a ParTy corpus was taken into consideration. Although it did offer promising results, the available files' format did not satisfy the expectations and would not be easily appliable to the planned analysis. Henceforth, a working corpus inspired by Levshina's project was created *ad hoc*. After a multiphase preparation of files, they could finally be annotated according to the proposed scheme and verified.

As explained in Section 2.2, there were a few different visions taken into consideration as for the annotation scheme, until they merged into the final form constituting the proposal:

| **type** | **literature** | **values** | **examples** | **legend** | **Description** |
|---|---|---|---|---|---|
| **PROP** | hedge, round | **ATT \| RNF \| EV** | That's **kind of** freaky. | **PROP** | Propositional |
| | | | And I was hoping for **some sort of** tactical plan | **PRF** | Performative |
| **PRF** | hedged performatives | **MITS \| MITA \| NPL** | And you **may choose** a woman. | **CMT** | Commitment |
| | | | I **might just give** you a big wet kiss. | **SCH** | shield |

| | | | | | |
|---|---|---|---|---|---|
| **CMT** | plausibility shield | **MITS \| MITA \| NPL** | **It's hard to say**, but she wasn't a redhead. | **ATT** | attenuation |
| | | | **I don't know if it's true.** | **RNF** | reinforcement |
| | | | I said you weren't interested, **right**? | **EV** | evasion |
| **SCH** | attribution shield | **MITS \| MITA \| NPL** | **She said** that he stalked her. He's in St. Louis. | **MITS** | self-serving mitigation |
| | | | And **according to your boss.** | **MITA** | altruistic mitigation |
| | | | | **NPL** | negative politeness |
| | | | | **DISC** | imprecise discourse effect |

*Table 28 Copy of the final annotation scheme proposal.*

Firstly, the hedges were to be classified according to their form and main function into four types with the tags: PROP for propositional hedges, working on the level of a single sentence; PRF for hedged performatives and other modal elements with the focus of illocutionary force; CMT for traditionally called plausibility shields; and, lastly, SCH for traditional attribution shields. The last three were supposed to cover the forms affecting the whole speech act, both in illocutionary and perlocutionary aspects, as the division between propositional and speech act hedges was one of the most fundamental. To further specify the role of a given expression, I attributed an additional value to each of the original tags: ATT for attenuation, RNF for reinforcement, EV for evasion, MITS for self-serving mitigation, MITA for altruistic mitigation, NPL for negative politeness, and DISC for other less common and imprecise discourse effects. The last value proved to be especially useful for expressions complicated to classify as serving strictly one scope.

As further described in the third chapter, the process of annotation presented some complications. The main technical problems originated in the previously predicted uncertain quality of subtitle files. The more or less appropriate decisions of the authors resulted in a loss of a fraction of the corpus, due to unavoidable cuts and modifications. The final numbers of tokens and lines of text was presented in Table 4. Considering the need for the corpus contents to be more natural than curated, it was not considered a serious difficulty. Nonetheless, the same uncertainties regarding the choices of translators, being secondary authors of the files,

resulted in double the issues with interpretation of the texts and *ipso facto* decision concerning the annotation.

As presented in 4.1 there were numerous instances of problematic expressions, both in English and Polish files. During the annotation process, I had to make arbitrary choices regarding the attribution of proper tags and decision on how to standardise the process. It makes the results of the present exploratory study admittedly debatable, but such approach is unavoidable in the studies concerning phenomena insomuch ambiguous as the pragmatic aspects of hedging. The biggest concerns that emerged from the analysis are those related to the SCH type classification, and the MITA value, both originating in the insufficiency of data.

The MITA value for hedges serving as an altruistic mitigation consisted in the smaller percentage of data gathered, as given in 3.3.2. Although the speech act hedges sharing this value were not inexistent, their quantity would not have allowed for a more detailed analysis. However, all of the speech act hedges' values in the sample chosen for analysis were rather finely distributed and they still managed to exhibit some general differences with the limited source material. Still, an issue could be raised of whether the division into self-serving and altruistic mitigation is truly needed. If it were simply to facilitate a more detailed analysis of data or, potentially, an automatic attribution of these labels, the answer might be to consolidate the two. Nevertheless, the main goals of designing this proposal were to create a tool that would be applicable to various source material, but also allowed for an easier interpretation of pragmatic aspects of languages that may differ extensively in their discourse standards. Having that aim in mind, the separate value for MITA seem to be worth being retained. With regard to the SCH hedges, as a general category, they were definitely the least common within the corpus. For that reason, similar considerations to MITA value label were made but, ultimately, the attribution hedges category is certainly even more important than the former.

A crucial concern that could be made for the application of the scheme is in fact the limited amount of data on which its first implementation was tested. There are however several counterarguments:

- The scope of this thesis was in fact only to propose a potential tool that seemed missing in the field. Implementation of such scheme would inevitably have to undergo a few trials on different materials, so the work presented here is certainly not concluded;
- Even though the test sample in the form of this corpus is much limited when compared to the modern standards of corpora, it still provided quite a reliable amount of data for each of the hedges expected;

- Having designed the proposal à priori, apart from some superficial assessment of hedging devices present, the scheme proved to work well on the chosen corpus, offering an appropriate tag for each example. The problems concerning classification were not due to the lack of a label, but to the difficulties of interpretation;
- The annotation was necessarily subjective, however, the results presented in the closing chapter demonstrated that they correlate well with the assumptions that could have been made about each particular tag, or rather, class of hedges.

Similarly, there are both negative and positive arguments that may be put forward given the results of the contrasts between the two languages. Most importantly, even though English and Polish belong to vastly different groups, they both belong to the Indo-European family, so it could be said that the conformity of the scheme to both of them is not difficult to achieve. On the other hand, there are still some significant differences within the way they are constructed and the linguistic behaviours of their speakers, which seem to be depicted in the results of the statistical analysis.

By way of conclusion, one may only reiterate the awareness and acknowledgement of all the disputable issues of the proposal. Taking all that into consideration, the results of the test analysis of the annotation scheme that this thesis presents, offer a promising chance for further development and automatization that could benefit many types of linguistic research, including studies on translation.

# 5 References


Abercrombie, D. (1963). “Pseudo-Procedures in Linguistics”. *STUF - Language Typology and Universals*, 16 (1-4), 9-12. From https://doi.org/10.1524/stuf.1963.16.14.9.

Abercrombie, D. (1965). *Studies in Phonetics and Linguistics*. London: Oxford University Press.

Adamczyk, M. (2015). “Do hedges always hedge? On non-canonical Multifunctionality of ‘jakby’ in Polish”. *Pragmatics*, 25 (3), 321-344.

Aijmer, K. & Simon-Vandenbergen, A. M. (2003). “The discourse particle well and its equivalents in Swedish and Dutch*”. Linguistics,* 41 (6), 1123-1161.

Bednarek, M. (2011). “Expressivity and televisual characterization”. *Language and Literature,* 20 (1), 3-21.

Belica, C. (1996). “Analysis of temporal Changes in Corpora”. *International Journal of Corpus Linguistics*, 1, 61-73.

Bennett, G. R. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers.* Ann Arbor: Michigan University Press.

Biber, D. (1993). “Representativeness in corpus design”. *Literary and Linguistic Computing,* 8 (4), 234-257.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press.

Brown, P. & Levinson, S. (1978). *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Caffi, C. (1999). “On mitigation”. *Journal of Pragmatics*, 31 (7), 881–909.

Caffi, C. (2007). *Mitigation*. Elsevier Science Ltd.

Chomsky, N. (1957). *Syntactic Structures.* Berlin: Mouton de Gruyter.

Chomsky, N. (1962). “The logical basis of linguistic theory”, *in 3rd Texas Conference on Problems of Linguistic Analysis in English*, Austin, January.

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

Clear, J. (1992). “Corpus sampling”, in G. Leitner (ed), *New directions in English language corpora*, 21-31. Berlin: Mouton de Gruyter.

Clemen, G. (1997). *The Concept of Hedging: Origins, Approaches and Definitions.* Berlin: Walter de Gruyter.

Eaton, H. S. (1940). *Semantic Frequency List for English, French, German, and Spanish.* Chicago: University of Chicago Press.

Fillmore, C. (1992). “Corpus linguistics” or “Computer-aided armchair linguistics”, in J. Svartvik (ed), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991,* 35-60. Berlin, New York: De Gruyter Mouton.

Fraser, B. (1975). “Hedged performatives*”*, in P. Cole and J. L. Morgan (eds), *Syntax and semantics 3: Speech acts* 187-210. New York: Academic Press.

Fraser, B. (1980). “Conversational mitigation”. *Journal of Pragmatics*, 4, 341-350.

Fries, C. (1952). *The Structure of English: An Introduction to the Construction of Sentences.* New York: Harcourt-Brace.

Grice, H. P. (1975). “Logic and conversation”, in P. Cole et al. (eds), *Syntax and semantics 3: Speech act*s, 41-58. New York: Academic Press.

Gries, S. Th. (2009). *Quantitative corpus linguistics with R: A practical introduction.* New York and London: Routledge.

Grzegorczykowa, R. (1990). *Wprowadzenie do semantyki językoznawczej*. Państwowe Wydawn. Naukowe.

Huang, C. R. & Yao, Y., (2015). Corpus Linguistics, in *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., 949-953). Amsterdam: Elsevier.

Hübler (1983). *Understatements and hedges in English*. Amsterdam/Philadelphia: Benjamins.

Iruskieta, M., da Cunha, I. & Taobada, M. (2014). "A Qualitative Comparison Method for Rhetorical Structures: Identifyng different discourse structures in multilingual corpora". *Language resources and evaluation*, 2 (49), 1-47.

Johansson, S. (2007). *Seeing through multilingual corpora. On the use of corpora in contrastive studies.* Amsterdam/Philadelphia: John Benjamins.

Kalisz, R. (1993). "Different cultures, different languages, and different speech acts revisited". *Papers and Studies in Contrastive Linguistics*, 27, 107-118.

Kaltenböck, G. et al. (2010). *New Approaches to Hedging.* Emerald Group Publishing Limited.

Kennedy, G. (2001). Corpus Linguistics, in *International Encyclopedia of the Social & Behavioral Sciences* (2816-2820). Oxford: Pergamon. from https://doi.org/10.1016/B0-08-043076-7/03056-4.

Kjellström, A. (2019). *What may or may not be certain: A Study of the Translation of Hedging Devices from English to Swedish in a Non-Fiction Text*. [Independent thesis Advanced level, Linnaeus University]. Digitala Vetenskapliga Arkivet. http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-86270

Kleiber, G. (2003). *Semantyka prototypu: kategorie i znaczenie leksykalne*. Kraków: Towarzystwo Autorów i Wydawców Prac Naukowych "Universitas".

Koehn, P. (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation", In *MT Summit*, Phuket, Thailand, September, (79-86) from https://www.statmt.org/europarl/.

Kučera, H. & Francis, W. N. (1967). *Computation Analysis of Present-Day American English.* Providence: Brown University Press.

Labov W. (1966). "Hypercorrection by the lower middle class as a factor in linguistic change", in Sociolinguistics – Proceedings of the UCLA Sociolinguistics Conference, New York City, 1964, (84-113).

Lakoff, G. (1973), "Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts". *Journal of philosophical logic,* 2 (4), 458-508.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Langacker, R. W. (1993). "Universals of construal". *Annual Meeting of the Berkeley Linguistics Society,* 19 (1), 447-463.

Leech, G. (2007). "New resources, or just better ones? The holy grail of representativeness", in M. Hundt, N. Nesselhauf and C. Biewer (eds). *Corpus linguistics and the web*, 133-150, Amsterdam and New York: Rodopi.

Leech, G. (2014). *The Pragmatics of Politeness.* Jericho: Oxford University Press.

Lenci, A., Montemagni, S. & Pirrelli, V. (2020). *Testo e computer. Elementi di linguistica computazionale.* Carocci Editore.

Levshina, N. (2015). "Online film subtitles as a corpus: An n-gram approach". *Corpora*, 12 (3), 311-338.

MacWhinney, B. (1991). *The CHILDES Project: Tools for analysing talk.* Hillsdale, NJ: Erlbaum.

Manaris, B. (1998). “Natural Language Processing: A Human-Computer Interaction Perspective*”. Advances in Computers*. 47, 1-66.

Markkanen, R., & Schröder, H. (1989). “Hedging as a translation problem in scientific texts”. *Special languages: From human thinking to thinking machines*. 171-179.

Markkanen, R., & Schröder, H. (1997). “Hedging: A challenge for pragmatics and discourse analysis”. *Hedging and discourse. Approaches to the analysis of a pragmatic phenomenon in academic texts*, 24, 3-18. Berlin-New York: Walter de Gruyter.

McEnery, T. & Wilson, A. (1997). *Corpus Linguistics.* Edinburg: Edinburg University Press.

McEnery, T. & Xiao, R. (2007). “Chapter 2. Parallel and Comparable Corpora: What is Happening?”, in G. Anderman & M. Rogers (eds), *Incorporating Corpora: The Linguist and the Translator*, 18-31. Bristol: Blue Ridge Summit: Multilingual Matters. from https://doi.org/10.21832/9781853599873-005

Peterlin, A. P. & Moe, M. Z. (2016). “Translating hedging devices in news discourse”. *Journal of pragmatics*, 102, 1-12.

Preyer, W. (1889). *The Mind of a Child.* New York: Appleton.

Rühlemann, C. (2019). *Corpus Linguistics for Pragmatics: A Guide for Research.* London and New York: Routledge.

Prince, E. F., Frader, J. & Bosk, C. (1982). “On hedging in physician-physician discourse”, *Linguistics and the professions*, 8 (1), 83–97.

Rosch, E. H. (1973). “Natural categories”. *Cognitive psychology*, 4 (3), 328-350.

Rosch, E. H. (1976). “Basic objects in natural categories”. *Cognitive Psychology*, 8 (3), 382-439.

Sánchez, L. M. & Vogel, C. (2015). “A hedging annotation scheme focused on epistemic phrases for informal language”, *In Proceedings of the Workshop on Models for Modality Annotation*, London, April, Association for Computational Linguistics. from http://hdl.handle.net/2262/74004

Shmelev, A. (2020). “Russian language-specific words in the light of parallel corpora”, in H. Bromhead & Z. Ye (eds), in *Meaning, Life and Culture. In conversation with Anna Wierzbicka*, 403-419. Acton (Australia): Australian National University Press.

Sinclair, J. (2004). *Corpus and Text — Basic Principles.* London: Tuscan Word Centre.

Smith, R. (2012). “Distinct word length frequencies: distributions and symbol entropies”. *Glottometrics*, 23, 7-22.

Stern, W. (1924). *Psychology of Early Childhood up to Six Years of Age.* New York: Holt.

Taylor, L., Leech, G. & Fligelstone, S. (1991). A survey of English machine-readable corpora. In S. Johansson & A. Stenström (Ed.), *English Computer Corpora*: *Selected Papers and Research Guide* (319-354). Berlin, Boston: De Gruyter Mouton.

Thorndike, E. (1921). *A Teacher's Word Book of 30,000 Words.* New York: Columbia University Press.

Tiedemann, J. (2007). “Building a Multilingual Parallel Subtitle Corpus”, *Proceedings of the 17th Conference on CLIN,* Leuven, January, (147-162).

Tognini-Bonelli, E. (2001). *Corpus Linguistics at work: Studies in corpus linguistics.* Amsterdam: John Benjamins.

Van Halteren, H. (1999). “Syntactic Wordclass Tagging”. *Computation Linguistics,* 26 (3), 456-459.

Velupillai, K. (2020). “Enigma Variations: A Review Article”. *New Mathematics and Natural Computation*, 16 (2), 377-396, from https://doi.org/10.1142/S1793005720500234

Vincze, V. et al. (2008). “The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes”. *BMC bioinformatics*, 9 (11), 38-45.

Wierzbicka, A. (1985). “Different cultures, different languages, and different speech acts: Polish VS English”. *Journal of Pragmatics,* 9 (2-3), 145-178.

Wierzbicka, A. (1991). *Cross-cultural pragmatics: The semantics of human interaction.* Berlin & New York, N.Y.: Mouton de Gruyter.

Wittgenstein, L. (2000). *Dociekania filozoficzne*. Warszawa: Wydawnictwo Naukowe PWN.

Wynne, M. (2005). *Developing Linguistic Corpora - A Guide to Good Practice*. Oxford: Oxford Books for the Arts and Humanities Data Service.

Zadeh, L. (1965). “Fuzzy sets”. *Inform Control*, 8, 338-353.

Zins, C. (2007). “Conceptual approaches for defining data, information, and knowledge”. *Journal of the American society for information science and technology*, 58 (4), 479-493.

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language.* Cambridge: Harvard University Press.

# 6 Sitography


WWW.NATALIALEVSHINA.COM/CORPUS.HTML

OPENSUBTITLES.ORG

LINDAT.MFF.CUNI.CZ/SERVICES/UDPIPE/

UNIVERSALDEPENDENCIES.ORG/GUIDELINES.HTML

UNIVERSALDEPENDENCIES.ORG/TREEBANKS/PL_PDB/INDEX.HTML

UNIVERSALDEPENDENCIES.ORG/TREEBANKS/EN_LINES/INDEX.HTML

UNIVERSALDEPENDENCIES.ORG/TOOLS.HTML#UDPIPE

GITHUB.COM/TKARABELA/PYSUBS2

GITHUB.COM/JONORTHWASH/UD-ANNOTATRIX

GITHUB.COM/UNIPV-LARL/VALIDEASY/BLOB/MAIN/README.MD

USERS.OX.AC.UK/~MARTINW/DLC/INTEX.HTM

WWW.SKETCHENGINE.EU/CORPORA-AND-LANGUAGES/CORPUS-TYPES/

PLATO.STANFORD.EDU/INDEX.HTML

CHILDES.TALKBANK.ORG

WWW.CORPUSTHOMISTICUM.ORG/IT/INDEX.AGE

WWW.LANCASTER.AC.UK/FASS/PROJECTS/CORPUS/ZJU/XCBLS/CHAPTERS/A02.PDF

# 7 List of tables

# 8 List of figures

# 9 Ringraziamenti

In primis, desidero esprimere un sentito ringraziamento alla mia relatrice, Prof.ssa Erica Biagetti, che mi ha guidato, con disponibilità e gentilezza, nella stesura dell'elaborato. La ringrazio per un'infinita pazienza dimostratami in questi mesi e Le auguro gli studenti più tempestivi per le prossime esperienze da relatrice. Rivolgo le grazie per la prontezza e il supporto anche al mio correlatore, Prof. Flavio Cecchini.

Vorrei ringraziare i miei genitori per avermi sostenuto fin troppo in questo sogno dell'università italiana, nonostante i continui prolungamenti del percorso. Grazie per tutte le corse per i vari documenti, i numerosi 'pacchi da sù' e per aver sopportato qualche volta il caldo italiano, solo per passare le vacanze con me.

Ringrazio la mia carissima sorella Paulina che in qualche modo mi ha sempre fatto sentire capace. Le auguro di poter sperimentare lo stesso sentimento, assolutamente meritato, più spesso. Grazie per tutti i fangirl moments e per tutte le serate passate lamentandoci a vicenda.

Tengo a ringraziare anche mio padrino Leszek e mia zia Maryla per la disponibilità immensa e il supporto insostituibile. Sono veramente grata per tutti i regali inaspettati e la gentilezza sempre dimostrata. Ringrazio anche alle mie nonne Ania e Bronia per un'infinita affezione.

Aggiungo un abbraccio a Wojtek, mio cugino, per avermi tifato durante tutto questo percorso, per avermi riempito la testa con mille curiosità musicali cosicché non pensavo solo allo studio e per avermi sempre ricordato delle cose che mi rendevano felice da bambina.

A Monika, che mi è rimasta accanto sin dalle superiori, potrei ringraziare per tante cose. Più che altro, vorrei dire grazie per esser sempre stata per me un'ispirazione e per continuare a trasmettermi una costante motivazione, voluta o meno. Grazie per aver risvegliato in me l'amore per il ballo e per le lingue, apprezzo tutti i viaggi vicini e lontani, sia quelli fatti insieme che quelli fatti per vederci. In questa occasione, ringrazio in particolar modo per avermi consigliato così tanto di andare a Padova, tutti questi anni fa. Senza di te, non sarei qui adesso.

Non è possibile esprimere con parole, o almeno io non ne sono capace, di quanto sono grata a Chiara, Ale e Lisa per avermi tollerato, supportato e sopportato in questi quasi due anni che abbiamo passato insieme. Grazie per tutte le mattine, le serate, e anche i pomeriggi. Grazie per i giri e i concerti. Grazie per il cibo pronto a tavola, le tisane e i pasticcini da Manzato. A 'Giovanna' vorrei ringraziare per la pazienza e la flessibilità necessaria per condividere con me i propri spazi. Grazie per la bontà, per le serie viste e discusse, le sedute di terapia e uscite spontanee. Sono tremendamente grata a 'Gigliola' per le canzoni del giorno, sfilate di moda,

discussioni sensate o meno e tutta la tenerezza dimostrata. Infine, ringrazio ‘Petunia’ per aver sopportato le nostre crisi, per avere sempre tenuto la testa attaccata al collo, per la generosità e per avermi sempre distrutto nelle carte, giusto per sopprimere il mio ego. Conserverò sempre nel cuore la nostra convivenza.

A Deborah, ringrazio per tutti i momenti di gioia e crisi condivisi, sia a casa che al lavoro. Grazie per il comfort post-esami, per i momenti di stacco essenziali, passati sul balcone. Per la sua perenne presenza, anche in orari notturni, l’organizzazione, le zone rosse sopravissute e, peraltro, per l’aiuto nella stesura di questa tesi, sarò per sempre grata.

La mia gratitudine è dovuta anche ad un'altra persona, conosciuta a caso, che ha reso la mia magistrale molto più vivibile. Ringrazio Natalia per avermi scritto per prima, essendo che io non ci avrei mai pensato. La ringrazio per i colori che ha portato nelle giornate di studio grigie e per tutti i cappucini al latte di soia bevuti chiacchierando sui corsi. Inanzitutto, sarò sempre in debito a lei e Lazar per avermi ospitato nella loro casa, quando non avevo la propria. Grazie per aver vissuto con me tutta la burocrazia italiana e le traduzioni legali trovate all'ultimo.

Ai miei amici, Asia e Adam, vorrei ringraziare per l’entusiasmo continuo nei confronti dei miei studi a Padova che mi hanno fatto apprezzare ancora di più questa università. Sono grata per la loro voglia di venire a vedermi e di voler fornirmi supporto anche durante la discussione. Mi rende molto felice il fatto che dopo anni siamo rimasti in contatto tale che mi sono sempre sentita benvenuta a casa loro, anche per lamentarmi sui corsi di cui non sapevano nulla.

Un grande ringraziamento è dovuto anche alla mia cara amica d'infanzia, Kasia e un amico più recente ma altrettanto caro, Kamil. Per i loro piccoli gesti che mi hanno aiutato ad andare avanti e a sentirmi meno sola. Per la dedicazione di Kasia che ha passato ore ad aiutarmi a sistemare i dati necessari per questo lavoro. Per la possibilità di parlare del tutto e di niente - grazie.

Vorrei ringraziare a tutti i miei amici, vecchi e nuovi, per aver reso questi tre anni più intensi, più creativi, pieni di meraviglia, risate e degli attimi troppo fuggenti anche per la fotocamera. In poche parole, grazie per gli anni veramente felici, nonostante tutto. Ringrazio Patrycja per esser sempre stata disposta a vederci e per voler continuare a includermi nella sua vita futura con Błażej. A Francesca per avermi infettato di "che dire", per le giornate all’aula studio, ma anche quelle fuori. A Noemi, Ludovica, e tutta la gente "di via del Santo", ringrazio per numerose cene, pranzi e bellissime feste. A Iulia, grazi per avermi ricordato le gioie della mia prima permanenza a Padova. Anche se non ci siamo viste spesso, la coscienza di averla

nuovamente vicina, è stata un sollievo. A Klaudia e Iza ringrazio per ogni incontro, pur breve, che mi ricordava che le amicizie possono sopravvivere anche a distanza. A tutte le altre persone che non riesco a elencare singolarmente, indipendentemente da quando ci siamo conosciuti, ringrazio per aver partecipato in questa parte della mia vita.

Grazie di cuore all'Università di Padova e tutti i docenti meravigliosi con cui ho avuto a che fare. La loro passione e dedicazione mi ispirerà per sempre, ovunque io vada.

Infine, vorrei ringraziare me stessa per non essermi arresa, nonostante tutte le tentazioni. Sono fiera e felice di aver concluso e accettato questo percorso e questa tesi per quanto imperfetti. Vorrei ricordare questo momento come una prova della mia capacità di realizzare i miei sogni, malgrado gli ostacoli.

# 10 Podziękowania

In primis, chciałabym wyrazić serdeczne podziękowania dla mojej promotorki, prof. Erici Biagetti, za uprzejme wsparcie z którym poprowadziła mnie przez trudny proces redagowania pracy magisterskiej. Chciałabym podziękować jej za nieskończoną cierpliwość, którą okazała mi w ciągu ostatnich kilku miesięcy i życzyć samych sumiennych studentów w dalszej karierze. Dziękuję również za dyspozycyjność mojemu drugiemu promotorowi, profesorowi Flavio Cecchiniemu.

Następnie, chciałabym podziękować moim rodzicom za to, że tak bardzo wspierali mnie w tym marzeniu o włoskim uniwersytecie, pomimo ciągłych opóźnień. Dziękuję za bieganie po różne dokumenty, liczne "paczki przetrwania" i za znoszenie włoskich upałów, żeby spędzić ze mną wakacje.

Dziękuję mojej najukochańszej siostrze Paulinie, która w jakiś sposób zawsze sprawiała, że czułam się zdolna. Życzę jej, żeby częściej doświadczała tego samego uczucia, bardzo zasłużonego. Dziękuję za wszystkie fangirl moments i wszystkie wieczory spędzone na wspólnym narzekaniu.

Dziękuję również mojemu ojcu chrzestnemu Leszkowi i cioci Maryli za ogromną dyspozycyjność i niezastąpione wsparcie. Jestem ogromnie wdzięczna za wszystkie niespodziewane prezenty i zawsze okazywaną życzliwość. Dziękuję również moim babciom, Ani i Broni, za nieskończoną czułość.

Uściski dla Wojtka, mojego kuzyna, za dopingowanie mnie na tej drodze, za wypełnianie mojej głowy tysiącem muzycznych ciekawostek, żebym nie myślała tylko o studiach i za ciągłe przypominanie mi o rzeczach, które uszczęśliwiały mnie w dzieciństwie.

Monice, która trwa przy mnie od liceum, mogłabym podziękować za wiele rzeczy. Przede wszystkim chciałbym podziękować za to, że zawsze była dla mnie inspiracją i dawała mi ciągłą motywację, czy tego chciałam, czy nie. Dziękuję za rozbudzenie we mnie na nowo miłości do tańca i języków, za wszystkie podróże bliskie i dalekie, zarówno te odbyte razem, jak i te odbyte, aby się zobaczyć. Przy tej okazji, szczególnie dziękuję za to, że tak bardzo doradzała mi wyjazd do Padwy lata temu. Bez ciebie nie byłoby mnie tutaj.

Nie da się wyrazić słowami, a przynajmniej ja nie jestem w stanie tego zrobić, jak bardzo jestem wdzięczna Chiarze, Ale i Lisie za tolerowanie, wspieranie i znoszenie mnie przez te prawie dwa lata, które spędziłyśmy razem. Dziękuję za wszystkie poranki, wieczory, oraz popołudnia. Dziękuję za wycieczki i koncerty. Dziękuję za jedzenie na stole, herbaty i desery u Manzato. “Giovannie” chciałabym podziękować za jej cierpliwość i elastyczność w dzieleniu

się ze mną swoją przestrzenią. Dziękuję za dobroć, za obejrzane i przedyskutowane seriale, sesje terapeutyczne i spontaniczne wyjścia. Jestem ogromnie wdzięczna "Giglioli" za piosenki dnia, pokazy mody, rozsądne lub nie rozsądne dyskusje i całą okazaną czułość. Wreszcie, dziękuję "Petunii" za znoszenie naszych dram, za to, że zawsze trzymała głowę na karku. za hojność i za to, że zawsze niszczy mnie w kartach, żeby trochę stłumić moje ego. Zawsze będę pielęgnować nasz wspólnych czas w moim sercu.

Deborze dziękuję za wszystkie wspólne chwile radości i kryzysu, zarówno w domu, jak i w pracy. Dziękuję za pocieszanie po egzaminach, za niezbędne chwile relaksu spędzone na balkonie. Zawsze będę wdzięczna za jej nieustanną obecność, nawet nocami, organizację, wspólne przetrwanie czerwonych stref, oraz za pomoc w pisaniu tej pracy.

Wyrazy wdzięczności należą się również Natalii, przypadkowo poznanej, dzięki której łatwiej było przeżyć tę magisterkę. Dziękuję za to, że napisała do mnie pierwsza, bo sama nigdy bym na to nie wpadła. Dziękuję jej za koloryt, który wprowadziła w szare dni studiów i za wszystkie cappucino z mlekiem sojowym, które wypiłyśmy rozmawiając o zajęciach. Przede wszystkim zawsze będę wdzięczna jej i Lazarowi za ugoszczenie mnie we własnym domu, kiedy nie miałam swojego. Dziękuję za przebrnięcie ze mną przez włoską biurokrację i tłumaczenia prawne zlecane w ostatniej chwili.

Moim przyjaciołom, Asi i Adamowi, chciałabym podziękować za ich nieustający entuzjazm dla moich studiów w Padwie, który sprawił, że jeszcze bardziej doceniłam ten uniwersytet. Jestem wdzięczna za to, że chętnie do mnie przyjeżdżali i byli chętni wspierać mnie nawet podczas obrony. Cieszy mnie i wzrusza to, że po latach utrzymaliśmy kontakt do tego stopnia, że zawsze mogłam czuć się mile widziany w ich domu, nawet narzekając na zajęcia, o których nic nie wiedzieli.

Ogromne podziękowania należą się również mojej drogiej przyjaciółce z dzieciństwa, Kasi, oraz nowszemu, ale równie drogiemu przyjacielowi, Kamilowi. Za ich drobne gesty, które pomogły mi iść naprzód i poczuć się mniej samotnie. Za poświęcenie Kasi, która spędziła wiele godzin pomagając mi uporządkować dane potrzebne do tej pracy. Za możliwość rozmowy o wszystkim i o niczym - dziękuję.

Chciałbym podziękować wszystkim moim przyjaciołom, starym i nowym, za uczynienie tych trzech lat bardziej intensywnymi, bardziej kreatywnymi, pełnymi zachwytu, śmiechu i chwil zbyt ulotnych nawet dla aparatu. Jednym słowem, dziękuję za naprawdę szczęśliwe lata, pomimo wszystkiego. Dziękuję Patrycji za to, że zawsze chętnie się ze mną spotykała i chce dalej uwzględniać mnie w swoich przyszłych życiowych planach z Błażejem. Francesce za zarażenie mnie powiedzeniem "che dire", za dni w bibliotece, oraz te poza nią. Noemi,

Ludovice i wszystkim osobom "z via del Santo", dziękuję za wszystkie kolacje, obiady i wspaniałe imprezy. Iulii dziękuję za odnowienie we mnie radości z czasu mojego pierwszego pobytu w Padwie. Chociaż nie widywałyśmy się często, świadomość, że znów jest blisko, była dla mnie ulgą. Klaudii i Izie dziękuję za każde, nawet krótkie spotkanie, które przypominało mi o tym, że przyjaźń może przetrwać nawet na odległość. Wszystkim innym osobom, których nie jestem w stanie wymienić z osobna, niezależnie od tego, kiedy się poznaliśmy, dziękuję za to, że zechcieli uczestniczyć w tej części mojego życia.

Bardzo dziękuję Uniwersytetowi w Padwie i wszystkim wspaniałym wykładowcom, z którymi miałam do czynienia. Ich pasja i poświęcenie zawsze będą mnie inspirować, gdziekolwiek się udam.

Na koniec chciałbym podziękować samej sobie za to, że się nie poddałam, pomimo silnych pokus i chwil słabości. Jestem dumna i szczęśliwa, że ukończyłam tę ścieżkę i tę pracę dyplomową oraz że zaakceptowałam je, jakkolwiek niedoskonałe. Chciałbym zapamiętać tę chwilę na dowód tego, że mogę zrealizować wszelkie marzenia, nawet jeśli mają zająć więcej czasu niż bym chciała.