Modelling Differences between Near-Synonyms and its Application to Paraphrasing with Formality

Aina Garí Soler

MSc. Dissertation

Department of Intelligent Computer Systems
Faculty of Information and Communication Technology

UNIVERSITY OF MALTA

September 2017

Supervisors:
Dr. Lonneke van der Plas, University of Malta
Prof. Dr. Dietrich Klakow, Saarland University

Submitted in partial fulfilment of the requirements for the degree of
Master of Science in Human Language Science and Technology at the University of Malta
and of the
Master of Science in Language Science and Technology at Saarland University

M.Sc. (HLST)
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITY OF MALTA

Declaration

Plagiarism is defined as “the unacknowledged use, as ones own work, of work of another person, whether or not such work has been published” (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I, the undersigned, declare that the Master’s dissertation submitted is my own work, except where acknowledged and referenced.

I understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Student Name:	Aina Garí Soler
Course Code:	CSA5310 HLST Dissertation
Title of work:	Modelling Differences between Near-Synonyms and its Application to Paraphrasing with Formality

Signature of Student:

Date: 30 September 2017

EIDESTATTLICHE ERKLÄRUNG

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

DECLARATION

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Place: Arenys de Mar, Barcelona, Spain

Date: 30 September 2017

Signature:

Acknowledgements

First and foremost, I would like to thank my local supervisor at the UoM, Lonneke van der Plas, for her invaluable advice, feedback and ideas, and for her active and continuous involvement and effort in the supervision of this thesis. I also want to express my gratitude to Dietrich Klakow, my supervisor from the UdS, for always being ready to help me and answer my questions.

At the administrative level, I am especially thankful to Mike Rosner, from the University of Malta; and to Bobbye Pernice, at Saarland University, for always helping me very quickly and efficiently in urgent and important matters.

I also want to express my appreciation to Marianna Apidianaki and Alexandre Allauzen for their interest in this thesis, their suggestions and assistance.

I am very thankful for having had the opportunity to be part of the LCT programme and its wonderful community of brilliant students, and I am especially grateful to my classmates, for making of these two years the best student experience I have ever had.

I am very grateful to the five anonymous English native speakers who agreed to take part in the evaluation of the paraphraser.

I also want to thank my friends, Alex and my family, for their moral support and patience, and for understanding my long absences.

And last but not least, thanks to you, whoever is reading this, for your interest; I hope you will have a pleasant reading.

Abstract

Near-synonyms are words that have very similar meanings, but cannot generally be substituted by each other in a text. These words can differ in register (*drunk*, *inebriated*), affection (*dad*, *father*), attitude (*stingy*, *thrifty*), grammatical usage (*ajar*, *open*) or collocations (*task*, *job*), among others. Choosing the appropriate near-synonym can be trivial for a native speaker, but not necessarily for L2 learners or for automatic systems. This thesis is mainly focused on modelling the register, attitudinal and collocational dimensions of difference across near-synonyms. Formality and attitude scores for words are induced from their co-occurrence similarity with manually chosen word seeds, and collocational preferences of a word are extracted based on their statistical association. The power of this information is tested in a lexical choice task, the Fill-in-the-Blanks task. In addition, formality information is used to build an automatic paraphraser based on word and phrase substitution that intends to paraphrase sentences changing their level of formality. Results show that, whereas it is possible to obtain valuable information for formality and collocations even with a limited amount of text, this is not the case for attitude. With respect to the lexical choice task, our system does not beat a language model, but it does gain a small but significant improvement over a majority class baseline when combined with frequency information. A manual evaluation of the paraphraser's performance shows that, while it is still far from the quality of human-created paraphrases in terms of naturalness and grammaticality, its performance in terms of formality change is not significantly different.

Contents

1 Introduction

Near-synonyms are words that are very similar in meaning, but that cannot generally be used interchangeably. From what their name suggests, one would think that they are close to being synonyms, but not quite: Using one in the wrong context can produce a change in the style or attitude of the text, it could sound unnatural, or it could even slightly change the meaning of what is said. Being able to choose the appropriate word for the right context is therefore very important, and it can be a hard task both for a second language learner and for a machine.

Near-synonyms can differ in many different ways. Two near-synonyms can differ in formality (*lucky*, *fortunate*), in attitude (*persistent*, *stubborn*), in their collocational preferences (*terrible error*, *terrible mistake*), in their geographical use (*lift*, *elevator*), or in their denotation (for example, in terms of manner: *stare at* and *glimpse*). Sometimes, they differ in several of them at the same time. For example, *pass away* is not only more formal than *die*, but it can in principle only be used for humans, which makes their selectional preferences different.

In this thesis, we want to model near-synonyms based on these dimensions in an (almost) unsupervised fashion. In other words, we want to induce lexical information from text that is relevant to the specific dimensions. Concretely, we focus on formality, attitude and collocational preferences. The reason for this choice is threefold: first, they are interesting from a semantic point of view (as opposed to regional variation, for instance); second, there exist approaches to them on which our work can be based; and third, attempting to model more dimensions would have been out of our possibilities in terms of time and scope. We want to discover if automatically and unsupervisedly — using almost only raw corpora— acquired information on near-synonyms by dimension can be useful in a lexical choice task. In such case, the approach could then be easily extensible to other languages.

We understand formality as a continuum that ranges from vulgar or familiar language to more official language. The level of formality typically depends on the environment and the social characteristics of the situation and the people involved in a language act, and language adapts to the relevant formality with grammatical, but especially lexical resources.

With respect to attitude, we are interested in finding out the extent to which a word carries attitude or is neutral. In the given example, *stubborn* would be regarded as having more attitude because there is a negative intention or connotation to this word, as opposed to what happens with *persistent*.

We extract formality and attitude information in the form of lexical scores based on the distributional similarity of a target word with seed words, which are words that are known to be of extreme (in)formality or polarity. In the case of polarity, we use positive and negative seeds, instead of polar and neutral seeds, because it is not clear what "extremely neutral" words are. We evaluate these scores intrinsically, based on past work by Brooke (2014).

As a first approximation to see how dimensions interact within words, we investigate the relation between formality and attitude with the obtained scores, hypothesising that formal words might also be more neutral, and informal words more attitude-carrying.

"Collocational preferences" refers to the fact that words combine better with some words than with others, and this may differ across near-synonyms. For instance, one can *tell a story*, but not *say a story*. These preferences are extracted from corpora with statistical association measures previously used in the literature.

The information extracted on these three dimensions is then tested on the Fill-in-the-Blanks (FITB) task, a task that simulates lexical choice. Formality, attitude and collocations of a sentence are modelled and the appropriate candidate for that sentence is chosen. We compare the performance of our system on a modified version of this task with that of a language model and a simple majority class baseline. Our hypothesis is that having information on the specific dimensions that can distinguish near-synonyms can improve over the most basic approach to lexical choice, that is, word frequency.

Finally, we wanted to design an application that would make use of (part of the) obtained information. We decided to build an automatic paraphraser, based on lexical substitution, that changes the formality of a sentence. We believe this to be the first time such an automatic paraphraser is created. This would not only be yet another way of evaluating the acquired formality information, but also we think such a tool, if it works well, would be helpful for second language learners when having to adapt to environments they are not yet linguistically ready for. Paraphrases are analogous to synonyms and near-synonyms beyond the lexical level, and they can differ in formality and in other dimensions just as near-synonyms do. These differences can be externalised with resources of different nature, not just lexical (grammatical, reordering...), but we wanted to investigate the quality of paraphrases created based solely on simple lexical and phrasal substitution. We evaluate the output of this paraphraser with four native speakers who give their opinions on some of its linguistic characteristics and compare it to control paraphrases written by a different native speaker.

The main research questions this thesis aims to answer are the following:

1. Can lexical information on dimensions of difference between near-synonyms (concretely, on formality, attitude and collocations) be helpful in a lexical choice task?
2. Are lexical formality and attitude correlated?
3. Can a paraphraser produce correct paraphrases with a perceived change in formality applying only word and phrase substitution?

Outline

This thesis is organised as follows. Sections 2 and 3 summarise, respectively, the theoretical and practical background in near-synonymy, in (the extraction of) formality, attitude and collocational-related lexical information, in the relation between the first two, as well as in paraphrasing. Section 4 describes the methodology and resources used to induce lexical formality scores and the results of an intrinsic evaluation on them. Section 5 does the same for attitude, and in addition it presents our investigation on the appropriateness of seed lists by evaluating them as clusters. Section 6 looks at the relation between formality and attitude using results from Sections 4 and 5. Section 7 describes how we extracted collocations from corpora. Section 8 is devoted to the Fill-in-the-Blanks task, a global evaluation for the three kinds of information, and different approaches to it are presented and compared. In Section 9, we present the resources and structure needed for the automatic paraphraser, evaluate its output and discuss the results. Finally, Section 10 summarises the main conclusions of this work and puts forward some suggestions for future research.

2 Theoretical Background

In this section, we introduce the theoretical notions on which the work of this thesis is based. We first focus on near-synonymy and on how it differs from synonymy. After briefly introducing the ways in which near-synonyms may differ between them, we shift focus to the three that are relevant for this work and their characteristics: formality, attitude and collocations. Finally, the concept of paraphrase and its relation to near-synonymy are introduced.

2.1 Synonymy and Near-Synonymy

Different existing definitions of synonymy focus on various characteristics of synonyms: several make mention of the identity or mutual entailment between two words (R. Harris, 1973; Chierchia & McConnell-Ginet, 2000; Kempson, 1977); others, of the preservability of the truth conditions of a sentence when applying substitution (Cruse, 1986); of their full interchangeability (Jackson, 2014), partial interchangeability (Kreidler, 1998), or of the similarity of meaning between one of their senses (Katz, 1972). Apresjan (1973) understands synonyms as words that differ in fine nuances but designate the same thing. Church et al. (1994) conceive of a group of synonyms as a prototype-based category where one word is the most prototypical of all and the other words differ in some aspects. Warren (1988) establishes a distinction between synonyms and variants, where the first coincide in part of their meaning and usage, and the second differ stylistically and expressively, which makes them irreplaceable. Warren's definition of variants is actually equivalent to the subset of near-synonyms with which this thesis is concerned.

Many authors argue instead that true synonyms do not, or cannot, exist (e.g., (Bloomfield, 1933; Quine, 1961; Chafe, 1971; R. Harris, 1973; Cruse, 1986). Jackson (1988) says synonyms cannot exist because they should be exactly identical in all aspects, and such a complete identity cannot be found in language. Saussure (1959) notes that, within a language, "all words used to express related ideas limit each other reciprocally", implying that language avoids lexical items that overlap completely in meaning. Indeed, contrary to its opposite phenomenon, polysemy, synonymy is not a means for linguistic economy or expressivity, but rather an "extravagant luxury" (Murphy & Medin, 1985).

Two words are near-synonyms if they are very similar in meaning, but not similar enough to be interchangeable in every context. This is the definition of near-synonyms that will be used throughout this thesis, even though there exist many different definitions. The main question, given this definition, is what it means for two words to be similar. As has been noted in the literature, similarity is still an unclear, intuitive concept (Murphy & Medin, 1985), and describing concepts on the basis of a vague notion can lead to imprecise definitions. According to Murphy and Medin (1985), two words are more similar if, for example, they have more characteristics in common than other pairs of words. These characteristics include denotation, connotation and register, among others.

2.1.1 Near-Synonyms in Language Learning

Near-synonyms exist, presumably, in all languages, but they do not necessarily have direct correspondences across languages. Sometimes, a language makes distinctions that another language does not. This is, for example, the case of the Spanish verbs *saber* and *conocer*, which are present in other romance languages, and which —in their most typical meaning— both translate into English as *to know*, but are very well differentiated: the first one is used for facts and abilities, whereas the second one is used for people and things. For instance, in the sentence: *Conozco a Pablo, sé que es muy listo* (*I know* (conocer) *Pablo, I know* (saber) *he is very smart*) the two verbs cannot be switched. In other cases, two languages have different near-synonyms for one single concept, but these overlap only partially in meaning. An example of this are the English words *forest*, *woods* and *copse* in contrast with the German *Wald* and *Gehölz* (Edmonds & Hirst, 2002): *Wald* is similar to *forest*, but can also denote a smaller area of trees, being sometimes equivalent to *woods.* The word *Gehölz* can be equivalent to *copse* or to (small) *woods.*

With these examples, it is easy to envision that cross-linguistic differences in the way near-synonyms cover a range of meanings can pose difficulties for second language (L2) learners. Hemchua, Schmitt, et al. (2006) analysed lexical errors in the second language production of Thai learners of English. Their analysis was based on a fine-grained distinction of various kinds of formal and semantic errors, including wrong usage of near-synonyms and different subtypes of collocation errors as separate categories. The results of their analysis showed that the most common subtype of lexical error were near-synonyms (19.54%). If we exclude the collocation subtypes "preposition partners" and "arbitrary combinations" (for instance, *surrounded by* —not *with*— or *hitch-hike*, respectively), which are of a less semantic nature, the other two collocation subtypes ("semantic word selection" and "statistically weighted preferences") sum up to 7.28% of the lexical errors. In other words, up to 26.82% of their lexical errors fall in the category of what we understand as near-synonyms.

Another study by Shalaby et al. (2009), also concerned with the analysis of lexical errors, examined the second language production of students at a Saudi university. In this study, near-synonyms are also the most attested subcategory of lexical errors (24.08%), and they constitute the 39.84% of semantic lexical errors. In their analysis, collocations were also in a separate category, but no distinction was made among collocation subtypes.

To the best of our knowledge, work on the relative frequency of errors on the different dimensions of near-synonyms, other than collocations, cannot be found.

Because of the difficulties that they pose for learners, and sometimes even for native speakers, many monolingual, as well as bilingual specialised reference books exist (Hayakawa & Ehrlich, 1994; Farrell, 1977; Batchelor & Offord, 1993), among others.

2.2 Dimensions of Difference Between Near-Synonyms

This thesis focuses on three specific dimensions of difference between near-synonyms, namely formality, attitude, and collocational preferences; but these are not the only existing dimensions. There are many ways in which two similar words can differ that have been discussed in the literature (Palmer, 1981; DiMarco et al., 1993; Edmonds, 1999). Palmer (1981) distinguishes "at least 5" dimensions of difference: dialect (such as the difference between *fall* and *autumn*), style or register (*pass away* and *die*), emotive or evaluative character (*thrifty* and *stingy*), collocational restrictions (*rancid* occurs with *bacon* or *butter*, *addled* with *eggs*), and finally there are words that have similar or overlapping meanings, but not exactly the same. Palmer calls this "a loose sense of synonymy". In his PhD thesis, Edmonds (1999) thoroughly explores the dimensions of difference between near-synonyms and establishes a total of 35 fine distinctions. This classification, he reports, is based on intuition, on past work by other linguists (Cruse, 1986; DiMarco et al., 1993; Stede, 1993) —among others—, on various dictionaries that discriminate between near-synonyms, such as Hayakawa and Ehrlich (1994) or Batchelor and Offord (1993), and on raw corpora. These 35 distinctions are divided into four subgroups:

1. **Collocational** and **syntactic** variation (*task* and *assignment* with respect to *daunting*);

2. **Stylistic** variation, including **dialect** and **register** (*loo* and *toilet*);

3. **Expressive** variation, including **emotive** and **attitudinal** aspects (*pushy* and *aggressive*);
4. **Denotational** variation in a broad sense, including propositional, fuzzy, and peripheral aspects (*mist* and *fog*, which differ in degree).

In this work, we focus on the groups (1), (2) and (3) of Edmond's classification. Concretely, the dimensions we will be working with could be included in the sub-distinctions "Formality", "Euphemisms" (in 2) and "Expressed attitude" (in 3), as well as all kinds of collocations (in 1).

2.2.1 Formality

As has been said, style is one of the possible dimensions of variation across near-synonyms. In Edmonds (1999), "style" can refer to different types of dialects (geographical, temporal, etc.) or to what he calls "stylistic tone", which encompasses various word style characteristics such as force, concreteness, simplicity, formality and others. This work will be specifically focused on the dimension of formality.

Formality in language has been often defined situationally rather than linguistically. According to the Dictionary of Language Teaching and Applied Linguistics (Richards et al., 1997), what characterises formal speech is the carefulness of the speaker in choosing what and how to say it. Heylighen and Dewaele (1999) suggest that language is more formal when the "distance in space, time or background increases", as well as when "the speaker is male, introverted or academically educated". Therefore, they claim, linguistically, a speaker in a formal setting would avoid ambiguity as much as possible, following Grice's maxims (Grice, 1975).

Formality is a subjective concept and it is a matter of degree, that is, there is a gradual, continuous progression from informality to formality, and not a categorical distinction where a text would be considered to be either formal or informal. This difficulty in separating "careful" from "casual" speech objectively has been, according to Rickford and McNair-Knox (1994), a reason why it has not been studied quantitatively by researchers so much. Indeed, Lahiri et al. (2011) performed an inter-annotator agreement between two annotators who had to label sentences as formal or informal, which showed very low scores. In a later paper, Lahiri and Lu (2011) make a similar experiment but this time with a 5-point Likert scale. The agreement was higher in this case but still low, which shows that formality may be better understood in terms of degree but it remains a subjective notion. Cosine similarity between the two ratings showed that, even if different, judgments were similar in direction mostly. Certainly, it is important to note that judgments were made on sentences, and not on larger texts, where there is more linguistic evidence to assign a formality level.

Another indication of the gradual nature of formality is found in Mosquera and Moreda (2011), where texts of three different levels of informality —that is, only within the informal part of the formality spectrum— are successfully differentiated by a clustering algorithm. Specialised dictionaries such as Choose The Right Word (Hayakawa & Ehrlich, 1994) also point to this gradualness by describing near-synonyms' formality in relative terms: "*Absorb* is slightly more informal than the others and has, perhaps, the widest range of uses", when referring to the set of near-synonyms formed by *absorb*, *assimilate*, *digest*, *imbibe*, *incorporate*, and *ingest*.

2.2.2 Attitude

Another dimension of difference among near-synonyms that we will focus on is expressive variation. Edmonds (1999) distinguishes two kinds of expressive variation in near-synonyms: emotional and attitudinal. Words can express an emotion by themselves (for example, *love*), or they can indicate the speaker's opinion or attitude with respect to some entity (for instance by using the word *threat*, as opposed to *warning*).

Emotion can be classified into many subtypes, but for practical applications (see Section 3.3) typically only the polarity and strength of an emotion are distinguished. Polarity refers to the direction of the sentiment expressed, that is, whether the word inherently carries a positive or negative sentiment; and strength quantifies this direction. For example, both *happy* and *delighted* are of positive polarity, but *delighted* is stronger.

With respect to attitude, words are considered to be pejorative, neutral, or favourable with respect to an entity (Edmonds, 1999). As an example, we have the words *stingy* and *thrifty*, which could be used to describe the same personality trait but indirectly expressing the speaker's judgment on it. These words do not express an emotion by themselves, but rather show the stance of a speaker on a specific matter.

The study of polarity and other emotion-related notions, like subjectivity, is typically known in Natural Language Processing as Sentiment Analysis. Within Sentiment Analysis, Subjectivity Analysis intends to distinguish sentences that express an opinion as opposed to an objective fact. The definition and applications of Sentiment Analysis will be discussed in more detail in Section 3.3. In this work, we make use of the tightly related concept of polarity to model attitude, identifying pejorative attitude with negative polarity and favourable attitude with positive polarity.

Human annotations of polarity at the sentence level with three labels (positive, negative and neutral) have been reported to be reliable (Cohen's kappa κ of 0.91 for 100 sentences and 2 annotators, (Kim & Hovy, 2004)). Other annotations carried out for polarity (Somasundaran et al., 2007) and for subjectivity (Wiebe et al., 2005), in both of which annotators had been trained and labeled a bigger number of sentences, had lower results (respectively, $\kappa = 0.789$ and $\kappa = 0.77$, average between three annotators). For now, these results suggest that we are dealing, again, with a subjective notion.

2.2.3 Relation between Formality and Attitude

Journalism, law, science, fiction... every textual genre has its own linguistic characteristics. These characteristics generally come by convention, and there are numerous style manuals that have been written for different languages so that all writers can follow general recommendations on grammar, vocabulary, correction or style in a specific genre or, sometimes, in general usage of the language. Many of these genres are formal by nature, such as journalism or academia. In fact, most manuals are aimed at language used in formal environments; informal genres, such as personal letters, e-mails or chat messages, do not in general need to follow special rules or recommendations. Examples of such manuals are *Elements of Style* (Strunk, 2007), *A Dictionary of Modern English Usage* (Fowler et al., 1926), *OSCOLA* (Meredith, 2010), the *Publication Manual of the American Psychological Association* —better known as APA style, designed for academic texts in behavioural and social sciences– (American Psychological Association, 2010), *Chicago style* (Turabian et al., 2009), *The BBC News Style Guide* (*BBC News style guide*), *Scientists must write* (Barrass, 2005)... In such manuals of style we can find several extracts that encourage the writer to avoid subjectivity and opinion:

- "Place yourself in the background", "Write in a way that draws the reader's attention to the sense and substance of the writing, rather than to the mood and temper of the author", "Unless there is a good reason for it being there, do not inject opinion into a piece of writing.", "Similarly, to air one's views at an improper time may be in bad taste" (Strunk, 2007).

- "We can't replace tested knowledge and hard-won understanding with personal opinion, a relativistic view of truth, (...)" (Turabian et al., 2009).

- "*Dictator* is a term generally to be avoided (except in a historic context) because it is too subjective.", "Our responsibility is to remain objective" (*BBC News style guide*).
- "Statements should be objective (based on evidence), not subjective (based on the imagination or unsupported opinion). So, avoid excessive qualification." (Barrass, 2005).

If these recommendations are followed by writers, we would expect formal texts to be more neutral in terms of attitude than informal texts. In other words, we expect formality and attitude levels to be correlated.

2.2.4 Collocations

The term "collocation" was first introduced by Firth (1975) to refer to combinations of words that are frequently used together instead of other equivalent and plausible word combinations. A very typical example of a collocation would be the phrase *to give a talk*, which is preferred over combinations with other verbs, such as *make*, *hold* or *deliver* —which can be naturally used with the similar word *speech* in *deliver a speech*—. Other examples of collocations are *to break a promise*, *to drive someone crazy...*

As opposed to other types of Multiword Expressions, collocations are semantically transparent, which means that they are in principle unequivocally understandable but, since the preference for a specific word combination over another one is conventional, they can be produced incorrectly by L2 speakers and need to be learned (Nunberg et al., 1994). It can be argued that the preference for a specific word in a collocation is given by a restricted, differentiated sense of that word. For instance, many dictionaries would have an extra entry for the special sense of *stiff* in *stiff drink*.

Collocations are also one of the dimensions that can distinguish near-synonyms, with the example of a *daunting task* vs. a *daunting assignment*. In fact, collocational information of near-synonyms can reveal much more than just preferences in co-occurrence patterns: it can potentially also tell us about the denotational information that differs between two words. Cruse (1986) gives the words *land* and *ground* or *flesh* and *meat* as examples: the words in the pairs mean (almost) the same, but *land* "describes the dry surface of the earth in contrast with *sea*, while *ground* describes the dry surface of the earth in contrast with *air*". In the second case, "*flesh* (is) profiled against the frame/domain of the body's anatomy and *meat* (is) profiled against the frame/domain of food". This information can be implicitly encoded in text if we capture the fact that *flesh and bones* and *meat and potatoes* are good collocations but *meat and bones* and *flesh and potatoes* do not collocate so well, or could mean something else.

While sometimes words can be attracted to each other in text, there are cases in which words appear to be repelling each other, that is, they "co-occur less often than would be expected by chance" (Evert, 2005); two words with such a negative attraction between them have been termed "anti-collocations". There are two main reasons why anti-collocations exist, what could be understood as there being two types of anti-collocations. The first reason can be the fact that there is already a competitor that collocates well with a word, such as in the above mentioned case of *give* vs *deliver a talk*; and these are the anti-collocations that are typically of most research interest. The other reason may lie in some sort of incompatibility between the words, be it semantic, pragmatic or stylistic (Evert, 2005). This last type can pose a problem to the detection of the more interesting first type of anti-collocations, because they may be the majority of cases (Evert, 2005).

2.3 Paraphrasing

Two phrases, sentences or texts are a paraphrase of each other if they express the same —or almost the same— meaning with a different surface form. In this sense, paraphrases are the equivalent of synonyms beyond the lexical level. As an example, consider the sentences:

> *Today, almost a thousand visitors came to the building.*
>
> *The building was visited by nearly a thousand people today.*

Here, there is a typical case of syntactic paraphrase, with active and passive voice. Other lexical items have been moved and/or replaced: *Almost* and *nearly* are near-synonyms, and *people* is replacing *visitors.* However, this last pair of words are in a hyponymy relation, rather than in a near-synonymy one. The concept of entailment is worth mentioning here. Paraphrases can be thought of as sentences that entail each other reciprocally, that is, if one of the sentences is true, then the other one must be true as well. But, as we have seen, we can adopt a notion of paraphrase that is less strict, an unidirectional paraphrase. This is typically the case when hypernyms, rather than near-synonyms, are involved. If there are two sentences, sentence A and sentence

B, and A entails B but not viceversa, B can still be a correct paraphrase of A. For example:

A: *I own a cat.*

B: *I own an animal.*

A entails B, because if I own a cat, it is true that I own an animal. The reverse is not true: the fact that I own an animal does not mean I own a cat. In this case, B would be an acceptable paraphrase of A, but A would not be a correct paraphrase of B.

Ambiguity —both lexical and syntactic— does not allow to make judgments on whether two expressions are paraphrases of each other (Androutsopoulos & Malakasiotis, 2010). As an example, consider the sentence *He saw her duck.* Before we can create or choose a paraphrase for this sentence, a decision on its meaning has to be made: either she has a duck, or she moved her head down. If this is undecidable, it is also undecidable what is a correct paraphrase of it.

Paraphrases can differ in the viewpoint or evaluation a speaker wants to convey: in the place of emphasis, in style, connotation, etc (Hirst, 2003), in a similar way as near-synonyms can. Consider these examples taken from Hirst (2003):

1a. *At least 13 people were killed by a suicide bomber on a bus in downtown Jerusalem this morning.*

1b. *A suicide bomber blew himself up on a bus in downtown Jerusalem this morning, killing at least 13 people.*

2a. *The U.S.-led invasion of Iraq ...*

2b. *The U.S.-led liberation of Iraq ...*

2c. *The U.S.-led occupation of Iraq ...*

In sentences 1a and 1b there is a change in focus. Sentence 2a emphasises the action of killing and its object, while sentence 2b places the focus on the suicide bomber. Sentences 2a, 2b and 2c implicitly reflect different possible viewpoints.

Finally, we can talk about three different levels of paraphrases: lexical, phrasal and sentential paraphrases (Madnani & Dorr, 2010). Lexical paraphrases refer directly to synonyms (or near-synonyms) but, as shown above, it is not restricted to them: words in a hypernymy relation can also involve paraphrases. Phrasal and sentential paraphrases are related to less trivial syntactic changes, even if they can make use of lexical paraphrases, too.

3 Literature Review

In this section, we review relevant work done on lexical choice, which is the most direct application of near-synonym modelling, as well as on the specific dimensions of near-synonymy that this thesis is concerned with. We also present a few approaches to paraphrasing; concretely, to paraphrasing with a specific change in style or attitude.

3.1 Near-synonymy and Lexical Choice

Lexical choice is the problem of choosing the appropriate word for a given context, typically from a list of words that are near-synonyms. In this section, we review approaches that model (some of) the dimensions of difference between near-synonyms. We have seen that near-synonyms can differ in many different dimensions, sometimes in several at the same time. There have been approaches that aim at modelling these dimensions, since being aware of any of them can potentially improve human-assisted lexical choice (Inkpen, 2007a) as well as automatic lexical choice (Inkpen, 2007b).

Edmonds and Hirst (2002) created a computational, clustered model of lexical knowledge in which sets of near-synonyms were grouped in structured clusters that represented their fine-grained differences. This model, however, was built by hand and for only nine sets of near-synonyms, since it is very expensive.

Inkpen and Hirst (2006) built a lexical knowledge base that distinguished Edmonds's (1999) four coarse-grained distinctions. The knowledge base was mainly made of automatically acquired information from near-synonym dictionaries written for humans, such as Choose the Right Word (Hayakawa & Ehrlich, 1994). Using a small list of "seed words" that are typically used in dictionaries to refer to the characteristics of a near-synonym with respect to a specific dimension (for example, *suggests* for denotation and *favorable* for attitude), an algorithm learned linguistic patterns that were used to describe the differences among near-synonyms. These patterns were then classified into classes that represent more fine-grained distinctions, and a knowledge-extraction module was used to automatically extract and classify near-synonyms based on these patterns. The collocational information did not come from the book; instead, it was obtained from the statistical co-occurrence of words in corpora. They incorporated this knowledge base into Xenon, an NLG system, as a lexical choice component.

Gardiner et al., (2013) explored the attitudinal dimension of near-synonyms and showed that near-synonyms that differ in expressed attitude behave differently from others in a lexical choice task. They argue that the whole text or document should be taken as features, rather than the immediate context of the near-synonym only. This was based on the intuition that opinionated texts keep a stable, coherent sentiment throughout.

Some of these and other approaches have been evaluated on the so-called Fill-in-the-Blanks (FITB) task (Edmonds, 1997). The FITB task was designed in order to test the lexical choice of near-synonyms in context. In Natural Language Generation, lexical choice is typically described as the process of linking a representation of a concept to a specific linguistic form (Elhadad, 1993). Ideally, a lexical choice system would take into account both linguistic context and the differences between all possible candidates. The original design of this task consisted of sentences from the 1987 Wall Street Journal from which several selected words had been removed. The system had to select the word that was originally chosen for a slot from a set of near-synonyms (or near-synset). Many different approaches to lexical choice have been proposed and tested on the FITB task with the same data, which ensures comparability, and most of them do not take into account any dimensions of difference among near-synonyms (or account only for one). Approaches have been based on collocations statistics (Edmonds, 1997, 1999), on Pointwise Mutual Information (Inkpen, 2007a), on a language model (A. Islam & Inkpen, 2010), on Latent Semantic Analysis (Wang & Hirst, 2010), among others. The best result was obtained by Islam (2011), with an accuracy of 75.4% (averaged over groups of near-synonyms) with an n-gram based approach.

The FITB task, however, and the way it was designed, has some weak aspects: Firstly, it has traditionally been tested with only 7 near-synsets. While this can be easily overcome by adding more, this would reduce comparability with other works. Secondly, the data extracted from the Wall Street Journal reflects only newspaper usage. Apart from being domain specific, newspaper language has particular characteristics that could, possibly, be fixing both formality and attitude of near-synonyms across the whole task. Thirdly, a choice is considered correct only if the chosen word is the same word the original author chose. This leads to pessimistic results, since more than one candidate word could in principle be acceptable in a sentence. It has actually been proved that multiple options can often be correct, as Inkpen (2007) showed with a 78.5% of agreement between two native speakers on the task. Another reason why having only one correct option can cause bad results is that, as Edmonds notes, authors do not always reflect typical usage. In fact, Edmonds says, results are rather reflecting the proportion of typical usage in the newspaper.

In this thesis, the information extracted on the three dimensions of focus (formality, attitude and collocations) is used and evaluated on a modified version of the FITB task that seeks to overcome its drawbacks.

3.2 Formality

Traditionally, style has been studied at the text level with both language-independent and language-dependent variables (Biber, 1991). Despite the well-noted tendency to exclude content words of computational stylistic analysis, given their topic dependency (Argamon & Koppel, 2010), formality has also been studied at the lexical level in a more recent approach that we present here.

At the document level, Sheikha and Inkpen (2010) performed document classification into formal and informal by designing a series of linguistic features including passive and active voice, type-token ratio, average word length, and other language-dependent features, such as abbreviations and phrasal verbs. Heylighen and Dewaele (1999) created a formality score, called F-score, that was based mainly on the frequency of different part-of-speech (POS) tags, distinguishing between deictic and non-deictic categories of words, the latter being expected to be more present in formal texts. This score was used for sentences in Lahiri et al. (2011).

Formality has also been studied at the lexical level by Brooke (2014), and this approach will be described in more detail in Section 4 because it is the starting point of our work, for being lexical and continuous (not categorical). Brooke (2014) assigned words formality scores that can range from 1 (completely formal) to -1 (completely informal). His work builds upon other approaches that aimed at generating sentiment lexicons (Turney & Littman, 2003). These scores were calculated in different ways that were later evaluated and compared. The two baselines were scores based on word length and on the presence of latinate affixes in the word. One approach is based on the relative appearance of words in corpora of different formality, and it is based on the assumption that a word that appears mostly in formal corpora is formal, and viceversa. Another approach, which obtains better results, is based on two manually constructed lists of "seed words", one with words that are taken to be informal, the other one with formal words. Scores for words are then generated taking into account word co-occurrence using Pointwise Mutual Information (PMI) or Latent Semantic Analysis (LSA). The best result is achieved by a voting system that combines up to 5 different strategies. Our work applies one of Brooke's methods for formality induction to a much smaller corpus, obtaining comparable results.

Modelling formality can help researchers investigate and test linguistic hypotheses, and from a more practical point of view, formality at the document level can be, in some environments, such as the medical one (Sheikha & Inkpen, 2010), an indicator on what texts can be trusted.

3.3 Attitude

The term "attitude" is used in the near-synonym literature, but many studies and practical applications have been rather focused on the analysis of "sentiment" or "subjectivity". While these three are different concepts, there is certain overlap between them, and work done for sentiment and subjectivity can inspire and benefit the study of attitude.

Sentiment Analysis, also called opinion mining, aims at determining the emotion that a text expresses, very often with respect to a specific entity. The applications of such a field are plenty: from ranking products and merchants (McGlohon et al., 2010) to improving service or product quality, or even predicting election results (Tumasjan et al., 2010), to name just a few. With the increase of publicly available reviews and postings about products and other entities on the Internet, it has become easier to do research on it. Sentiment analysis typically classifies documents "polarly" into positive or negative, with or without varying degrees of polarity. This task can be addressed at different levels. Turney and Littman (2003) distinguished three: classification of words by sentiment, classification of documents by sentiment, and recognition of subjective fragments in text. The third level corresponds to subjectivity analysis, which consists of telling apart sentences or other fragments of text that express opinions from sentences that present objective facts (Wiebe, 2000). Subjectivity analysis can be used as a previous step to sentiment analysis by which objective sentences are discarded (Wiebe, 2000).

As it has been mentioned, sentiment can be present in lexical items; it is referred to as the "semantic orientation" of a word or a phrase. It is important to note that the analysis of sentiment should not rely on such cues alone to determine the polarity of a whole document. For instance, Turney (2002) noticed that positive movie reviews could easily include negative words referring to characters (such as *evil* or *foolish*) and these were confusing the system. Also, Kennedy and Inkpen (2006) take into account the so-called valence shifters in their analysis. Valence shifters are words that can change the semantic orientation of another word, such as the prototypical *not.* They found that including them in the analysis improved the results of a unigram approach.

Most work on the semantic orientation of lexical items has been aimed at inducing sentiment lexicons based on lists of seed words. This can be achieved from co-occurrence information or from lexical relations in existing lexicons (Andreevskaia & Bergler, 2006), such as WordNet (Fellbaum, 1998). An example of such an approach is that of Hatzivassiloglou and McKeown (1997), who looked at adjectives that are joined in text by the conjunctions *and* or *but.* Their hypothesis was that adjectives connected by *and* are of the same polarity, but are of opposite polarity if joint by the conjunction *but.*

Finally, one of the best known approaches to inducing the sentiment of lexical items is that of Turney and Littman (2003), who learned the semantic orientation of words from their statistical association with sets of positive and negative words. They used Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA). Similar work had been carried out by Turney (2002), who obtained the semantic orientation of phrases, rather than words, by calculating their PMI with the words *excellent* and *poor* and subtracting the first from the second. An approach that deals with phrases might have an advantage over a word-based approach when it comes to finding the overall polarity of a text, since words may have multiple senses with different polarity or can, without changing meaning, change their polarity (*unpredictable steering* vs. *unpredictable plot*) (Turney & Littman, 2003).

In this thesis, our goal is not to distinguish words by their polarity. More precisely, we want to tell apart neutral words (those which do not express any attitude or opinion) from those which do —no matter if they are negative or positive—.

Numerous sentiment corpora and lexicons have been created in the past years. We briefly present some of them, which have been used in this thesis, here:

- The Scale dataset v1.0[1] (Pang & Lee, 2005) contains around 5000 Internet movie reviews by four authors with normalised sentiment ratings derived from their original star rating.

- SentiWordNet 3.0 (Baccianella et al., 2010) is a set of WordNetsynsets that were automatically annotated with sentiment, assigning each of them a score for positivity, neutrality and negativity.

[1] Available at http://www.cs.cornell.edu/people/pabo/movie-review-data/ .

- The MPQA Subjectivity Lexicon (Wilson et al., 2005) is a set of more than 8000 words that were automatically classified as weakly or strongly subjective, and assigned a polarity label.
- The Harvard General Inquirer (Stone et al., 1966) is a lexicon with different kinds of information, including class information on polarity.

3.4 Relation between Formality and Attitude

Formality and attitude, as we saw in Section 2.2.3, can be directly related. Words can be described stylistically in many different aspects, and particular styles tend to be present with others. As Brooke (2014) noted, style is not only continuous, as opposed to discrete, but it is also multidimensional. He posed the word *cute* as an example: even though it is not slang, it is usually judged as being more informal than other slang words. This, according to Brooke, is because apart from formality, this word has other stylistic dimensions that may be correlated with informality: it is subjective and it is commonly used in spoken language. Brooke and Hirst (2013) induced a lexicon for three pairs of stylistic dimensions: colloquial vs. literary, concrete vs. abstract, and subjective vs. objective. Note that the left extremes share characteristics, and so do the right ones. Brooke divides them into a "situational" and a "cultural" pole, following previous work by Leckie-Tarry (1995). Indeed, their correlation analysis among styles in different genres —as calculated with an LDA topic model— showed that situational styles (colloquial, concrete and subjective) were positively correlated, and the same happened with the cultural styles. The two opposite poles, situational and cultural, were negatively correlated.

3.5 Collocations

There have been multiple approaches dealing with the extraction of collocations from corpora (Dias et al., 1999; Lin, 1998; Smadja, 1993; Kermes & Heid, 2003), some of which focus on specific POS combinations (Blaheta & Johnson, 2001; Breidt, 1996; Daille, 1994; Krenn et al., 2001). Most extraction methods are based on statistical measures of association between words that use co-occurrence data, or on a combination of measures (Pecina & Schlesinger, 2006). A good and extensive compendium of statistical measures with their description and properties can be found in Evert (2005). Some of these measures are Mutual Information (K. W. Church & Hanks, 1990), t-score (K. Church et al., 1991) or Chi-square.

Evert and Krenn (2011) suggested a way of evaluating the quality of the extracted candidates to collocations that is based on experts' judgments. However, the notion of collocation appears to be subjective, which is shown by Pecina and Schlesinger's (2006) first inter-annotator agreement with Cohen's κ ranging from 0.29 to 0.49.

To the best of our knowledge, only two approaches have related collocation extraction to near-synonymy in some way or another. Inkpen and Hirst (2002) extracted collocational preferences of near-synonyms with the intention of expanding an already existing knowledge-base of near synonym differences (Inkpen & Hirst, 2001), based partially on Pearce (2001). Pearce extended the extraction of collocations by looking at synonyms on WordNet. Concretely, he wrote, "A pair of words is considered a collocation if one of the words significantly prefers a particular lexical realisation of the concept the other represents." As an example, *emotional baggage* is considered a good collocation because *emotional* and *luggage*, a near-synonym of *baggage*, do not collocate well; they are in fact an anti-collocation. Instead of WordNet, Inkpen and Hirst (2002) applied this principle but used their own near-synsets.

Work done on collocations has applications both in Computational Linguistics as well as in lexicographic studies. Collocational or co-occurrence knowledge can be useful for both monolingual (Heid, 2004) and bilingual dictionaries (Smadja, 1993); for lexical choice in Natural Language Generation (Edmonds, 1997; Terra & Clarke, 2004) and for Machine (assisted) Translation (Smadja et al., 1996), among others.

3.6 Paraphrasing

Androutsopoulos and Malakasiotis (2010) proposed a classification of paraphrasing methods into three types: recognition, generation or extraction methods.

A paraphrase recognizer receives two linguistic expressions as input and has to decide whether they are paraphrases or not, or with what probability. A paraphrase generator takes an expression, usually a sentence, and has to suggest as many paraphrases of the sentence as possible. Finally, a paraphrase extractor takes a corpus as input and it needs to output pairs of paraphrases. Of course, these methods may be used in combination.

Paraphrases have shown to be very useful for Question Answering systems, where an expression can be phrased differently from a question but contain the answer to it (Riezler et al., 2007; Duboue & Chu-Carroll, 2006). Information extraction can be improved in a similar way by expanding the query with equivalent paraphrases (Shinyama & Sekine, 2003). Text summarisation and simplification can also take benefit of paraphrasing methods: sentences to be used for summarisation should not be paraphrases of each other, and simplification is itself a particular type of paraphrasing.

3.6.1 Paraphrase Generation Techniques

Our focus will be on generation techniques. Many generation techniques get their inspiration from Statistical Machine Translation (SMT). The commonalities between both tasks are evident: They both change text using a different language (considering language here as being both a specific system of communication and a particular way of expression) while keeping its meaning. In more technical terms, both tasks share basic methodological principles. The fundamental elements of an SMT system are a language model and a translation model. The purpose of the language model is to ensure fluency, whereas the translation model is responsible for the preservation of meaning. A translation model consists of a list of words or phrases and their equivalents in another language, as well as a translation probability. Such a model is typically trained on parallel corpora, that is, corpora that have the same content but in different languages. If the translation model could be trained to generate equivalent expressions in the same language, and not in a different one, it could be used for paraphrasing. For this, paraphrase extraction is needed. Monolingual parallel corpora, however, which would be needed to train such a model, are not as common as multilingual or bilingual parallel corpora, and usually not of the necessary size. Quirk, Brockett, and Dolan (2004) made an SMT-based paraphraser and obtained their paraphrase model from similar sentences coming from news articles that describe the same event. Zhao et al. (2008) and Bannard and Callison-Burch (2005) made use instead of a pivot language, which allowed them to use bilingual parallel corpora. Such an approach consisted of obtaining the translation of the input into the pivot language and then translating it back to the original language. When the expressions to paraphrase are fixed templates, such as *X wrote Y* or *X is the author of Y*, a monolingual corpus can be used to extract a paraphrase model (Ravichandran & Hovy, 2002).

3.6.2 Paraphrasing with a Change in Style

The described approaches perform paraphrasing in its strictest sense, without intending to change any characteristics of the original text in particular. To our knowledge, there has not been much work on paraphrasing with a change in style or attitude, let alone with a change in formality.

Xu et al. (2012) perform paraphrasing of modern English into Shakespeare-style language and viceversa. Paraphrase pairs are obtained from a parallel corpus of his original plays and their modern translations. They also propose automatic evaluation metrics that can account for style based on cosine similarity, a language model, and logistic regression. These metrics, which correlated with human judgments, capture the similarity of an output with one of the styles, or the probability that an output belongs to a style.

Valence shifting is another special type of paraphrasing (M. Gardiner & Dras, 2012). It consists of a change in the expressed attitude of a text. Guerini et al. (2008) built Valentino, a tool for valence shifting. They created vectors of negatively and positively valenced words based on WordNet and SentiWordNet, and extracted possible modifiers for them from a corpus. As an example, there is *wrong* and *incorrect*, *victory* and *triumph*. Paraphrasing was then performed with different strategies. As argued in Section 2.2.3, formality and attitude can be related, and in examples of Valentino outputs there is, even if involuntarily, a change in formality (*likely* = *with considerable certainty*).

Sheikha and Inkpen (2011) made a summary of the linguistic characteristics of formal and informal style and used them, together with parallel lists of formal and informal words, to adapt a Natural Language Generation (NLG) system to generate sentences of a given formality. This is not exactly a paraphrasing system because its input is a subject, verb and complement in any style, instead of a sentence.

In this thesis, we create a paraphraser that changes the formality of an input text in any direction: to more formal or to more informal.

3.6.3 Evaluation of Paraphrase Generation

There is no unique way to evaluate paraphrase generation. One possible method of automatic evaluation is to have a fixed reference set of all desired outputs for the given inputs, based on which precision and recall can be obtained (Callison-Burch et al., 2008). This set would contain numerous correct paraphrases, but the possible number of correct paraphrases is in principle unknown. Another available option of automatic evaluation would be to assess the effect of the generator in an application. Theoretically, paraphrase and entailment recognizers could also be used for evaluation, but their state of the art may still not be good enough for such purposes.

Finally, a more accurate but expensive option is manual evaluation. Native speakers can evaluate many different linguistic aspects of the output; for instance, Zhao et al.

(2009), in addition to asking speakers about the commonly evaluated in SMT features of Adequacy and Fluency, also ask them to evaluate the impact of a generator on an application.

4 Modelling Formality

In this and the next sections we present the work carried out for this thesis. In this section we present our work on the process of obtaining formality scores for words. Following successful work on style by Brooke (2014), we assigned words a formality score based on their distributional similarity with "seed" words of which we assume we know their formality. Scores range from +1 (extremely formal) to -1 (extremely informal): formality is conceptualised as a continuous property. The procedure required applying Latent Semantic Analysis (LSA) to a term-document matrix to extract cosine similarities between target words and seed words from it. Our target words were all words in the vocabulary, and their score is based on their similarity with seed words.

We first present the resources needed (seed words and corpora, Sections 4.1 and 4.2), the technique applied (Latent Semantic Analysis, Section 4.3), the calculation of the scores (Section 4.4) and finally show and discuss the results of an intrinsic evaluation (Section 4.5).

4.1 Seed words

Two word lists that represent both ends of the formality spectrum are needed: one that contains words that are assumed to have very high formality, and another one with words that are considered to be of very low formality. These are called "seed lists" and their content "seed words", because they are, in a way, the training set the scores are based on. They are later also used for evaluation, which can only be valid applying leave-one-out cross-validation.

Brooke (2014) provides the two seed lists that he used in his study[2], and we used these lists for our work as well. The informal seed list contains 138 words that were extracted from an on-line slang dictionary, whereas his 105 formal seed words were taken from a list of discourse markers and a list of adverbs from a sentiment lexicon, with the intention of avoiding specific topics and having a balance in sentiment in both seed lists. Part of speech is not balanced, because informal words are mostly nouns and interjections, among others.

4.2 Corpora used

We used a combination of corpora that included texts at both ends of the formality spectrum to get the necessary co-occurrence data to obtain the scores. For reasons of

[2] The seed lists used can be found on his website: `http://www.cs.toronto.edu/~jbrooke` .

availability, limitations on computer memory and time, two very big corpora used by Brooke (2014) were not used here: this is the case of Dev-Blog (216 million word tokens from English blogs), which was not available, and the much bigger ICSWM 2009 Spinn3r Dataset (Burton et al., 2009) (1.3 billion word tokens), with a varied combination of genres. However, as results show, it was possible to obtain very good scores with a much smaller combination of corpora. The corpora we used were the following:

The **British National Corpus**[3] (The British National Corpus, (2007)), BNC, consists of 100 million words from contemporary British English texts sampled from various sources and genres, such as newspapers, academic books, transcriptions of informal conversations, business meetings, etc. We included both its written (90%) and its spoken part (10%).

The **Brown corpus** (Francis & Kučera, 1979) is much smaller than the BNC (1 million words) and it contains written portions of different genres in modern American English sampled from printed text during 1961, including news, fiction, humour, etc. Brooke (2014) used this as a development corpus to experiment with various methods to obtain the formality scores. The Brown corpus is available as part of the Natural Language Toolkit (NLTK) (Bird et al., 2009).

The **Switchboard corpus**[4] (Godfrey et al., 1992) is a compilation of around 2400 American English telephone conversations that contains over 2.6 million words.

The **Scale dataset** v1.0 movie review data set (Pang & Lee, 2005) consists of slightly more than 2 million words from 5000 short movie reviews by four different authors, accompanied by the original score that the author assigned the movie, which we do not need for our purposes. Our reason for including this corpus, even if it was not used by Brooke, was first to ensure and increase the presence of polar vocabulary —because it is necessary for attitude scores, as explained in Section 5—; and second, to increase the final corpus size.

A small part of the **Blog Authorship Corpus** (BAC) (Schler et al., 2006) was also used; in fact, only 4000 blog posts with a total of nearly 28 million words, some of which disappeared in the filtering step described in the next section. This corpus is made of blog posts collected in August 2004 from *blogger.com.* Texts come with information about their author but this was not needed for our purposes. BAC is bigger than the BNC in size (140 million words), and we decided to include a part of it to compensate for

[3] The BNC can be freely downloaded from http://www.natcorp.ox.ac.uk/ .

[4] The Switchboard corpus can be freely retrieved on-line from https://catalog.ldc.upenn.edu/ldc97s62 .

our lack of blog text in comparison with Brooke's, but not all of it for various reasons:

- as we could observe, it is a noisy corpus with some blog posts in other languages, or with code-switching;
- to not have mostly blog posts;
- for memory limitation reasons.

From now on, we will refer to the collection of corpora we used in this work as "the Mixed Corpus".

4.2.1 Previous Corpora Filtering

In order to avoid computer memory problems and accelerate the computationally intensive process of Latent Semantic Analysis, while trying to stay as close as possible to Brooke's (2014) steps, these corpora were filtered and processed in the following two stages:

4.2.1.1 General pre-processing and filtering The general filtering stage has this name because it is both used for obtaining the formality and attitude scores. After the general filtering, a specific filtering has to be applied depending on the goal.

As a first step, all words that are not entirely in lower case were removed from the corpus, since we do not want to obtain scores for proper nouns or acronyms, which also result in a larger vocabulary.

Next, the corpus was tokenised with the NLTK tokeniser (Bird et al., 2009). At first, with the aim of reproducing Brooke's work as close as possible but with fewer data, only individual words were taken into account. However, in English many phrasal verbs have one-word near-synonyms. These close equivalents very often differ in formality, the phrasal verb typically being the more informal option (*find out* vs *discover*, *put up with* vs *tolerate*), with some exceptions (*pass away* vs *die*). Therefore, we wanted to obtain scores for phrasal verbs as well. In a second round of filtering, we tokenised the corpus merging those phrasal verbs that were not separated in text and explored the effect that adding them can have in accuracy, since they represent a considerable increase in the vocabulary. An extensive list of 2363 phrasal verbs was obtained from the reference section of a specialised website for English as Second Language[5]. A small set of repeated phrasal verbs in the list was removed. Then, phrasal verbs of more than 3 tokens (very

[5] https://www.usingenglish.com/reference/phrasal-verbs/list.html .

few) were dismissed, and phrasal verbs that offered a variation (such as *Colour (Color) up* were consequently added once for each of their possible versions. At the end of this process, 2350 phrasal verbs remained. The verbs were automatically conjugated with the Pattern Python module (Smedt & Daelemans, 2012) so that past, participle, gerund and 3rd person singular forms of the verb would be captured, as well as the infinitive. These forms were added to an instance of NLTK's Multi-Word Expression Tokenizer (MWETokenizer) and the text was tokenised accordingly.

In a next step, all words containing digits were removed, as well as all symbols other than the full stop ".". At this point, and not before for memory reasons, the filtered separate corpora were merged into one, and hapax legomena —that is, words that only appear once in the corpus— were removed.

Lemmatisation was considered as a possible previous step but this led to a small decrease in accuracy in Brooke's work. Another of his findings was that documents were a better context unit than paragraphs; therefore, we did not use paragraphs.

4.2.1.2 Specific filtering for formality In Brooke's work, only documents with more than 100 words were kept. Moreover, documents that do not contain any of the seeds, or which only contain seeds among the most common 20, were also removed.

Since the size of the Mixed Corpus is much smaller than Brooke's, we decided to apply only two of these filtering criteria: we removed documents with less than 100 words and documents that did not contain any of the seeds.

Because weights in the term-document matrix would be 1 for presence and 0 for absence —that is, frequency is not taken into account—, all repetitions of words in a document were removed at this point.

4.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997), also called Latent Semantic Indexing (LSI) is a commonly used technique in Natural Language Processing, especially useful for Information Retrieval or Document Classification. LSA is based on the semantic Distributional Hypothesis (Z. S. Harris, 1954), which states that words with similar meanings tend to occur in the same linguistic contexts. This co-occurrence is represented in the form of a matrix where, typically, word types are in rows and documents take up columns. Other linguistic units can be chosen: lemmas or pairs of words for the rows; and paragraphs or sentences can be used as text passages in columns instead of whole documents. In what follows, we will refer to them as words (in rows) and documents (in columns). The indices of the matrix may also contain different information, the most straightforward kind of information being raw frequencies of every word in the corresponding documents. Another option, which we used here, is to take into account only presence (1) or absence (0) of a word in a document. More informative indices are based on more advanced ideas, such as the widely used *term frequency-inverse document frequency*, *tf-idf* (Salton & McGill, 1983).

In such a matrix, distributional similarities between words or the similarity of two documents can be obtained with different measures that compare their vectors. One of the most widely used measures is the cosine, which takes the angle between two vectors into account and ranges between -1 and 1. However, in such a setting, similar documents —those that deal with the same topic— need to share (content) words in order for the similarity measure to capture this similarity. Precisely because of the use of synonyms or near-synonyms, which, despite having the same or similar meaning, have their own separate rows in the matrix, two documents that share a topic may end up having a low cosine similarity. Imagine the following artificial example made with near-synonyms:

Doc1: *My dad was super-tired after the bike trip*

Doc2: *My father was exhausted after the bicycle tour*

$\cos(\text{doc1}, \text{doc2}) = 0.0$

In such a game-sized setting, the cosine similarity between both documents would be 0 if we only take into account content words, which are the ones researchers are typically interested in. The same situation arises with words: two words that are similar in meaning must appear in the same documents in order to have a high similarity score.

LSA provides a solution to this problem by applying a mathematical dimensionality reduction technique called Singular Value Decomposition (SVD). By reducing the dimensionality of the matrix, not only does SVD reduce the required memory and time to work with it, but it also projects, respectively, co-occurring (= similar) terms and similar documents onto the same dimensions. In such a way, similar documents no longer need to share exactly the same words in order to have a high similarity score: it suffices with containing words that are similar in meaning.

The dimensionality of the new matrix is a parameter that can be chosen and dimensions have no direct interpretability. They can be regarded as “latent topics”: “latent” because they are not visible in the original matrix, and “topics” because they come from a combination of similar words or documents.

Among the advantages of LSA are the facts that it overcomes sparseness and handles synonymy. It is language-independent, but it needs a big corpus. It must be taken into account that this technique is computationally expensive. Crucially, it does not handle polysemy or homonymy: words can have multiple different meanings, which could make them belong to different topics at the same time, but latent topics in LSA are orthogonal: they do not have anything in common. An alternative to this is Probabilistic LSA (PLSA).

4.3.1 Singular Value Decomposition

Take a matrix A with terms (t) in the rows and documents (d) in the columns. A is of dimensionality $(t \times d)$. SVD decomposes matrix A into a product of three matrices:

$$\hat{A} = TSD^t \tag{1}$$

With minimal $||\hat{A} - A||^2$ (Â is as close as possible to the original A), and where:

T is a $t \times n$ orthogonal matrix;

S is a $n \times n$ diagonal matrix;

D is a $d \times n$ orthogonal matrix;

$n = min(t, d)$.

Matrix S contains the so-called "singular values" of A in descending order. In order to obtain the reduced matrix, a new number of dimensions, or topics k $(k < n)$ must be chosen. Then, the first k columns of T or D (depending on whether the object of interest are terms or documents) are multiplied by the left- and upper-most $k \times k$ submatrix of S, in the following way:

$B_{k\times d} = S_{k\times k}D^t_{k\times d}$ (document-topic matrix)

$C_{t\times k} = T_{t\times k}S_{k\times k}$ (term-topic matrix)

Document and term similarities can be obtained from matrices B and C respectively with the same cosine measure.

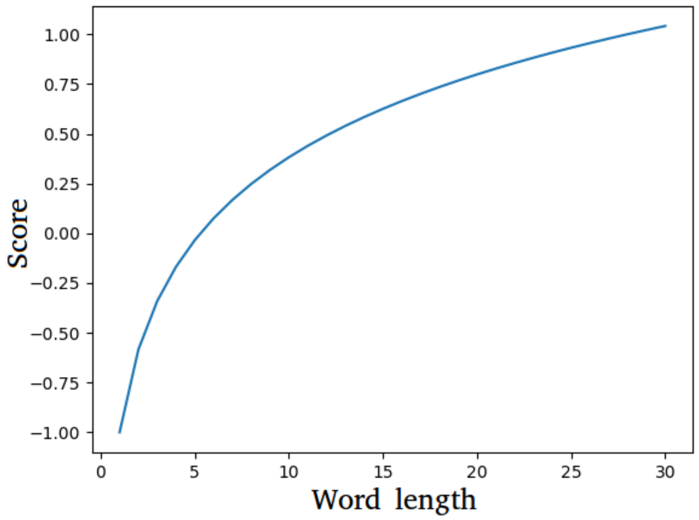

Figure 1: The word length baseline

4.4 Calculation of scores

4.4.1 Word length Baseline

We used one of Brooke's (2014) baselines, based on word length, which has been used in the past as a sign of formality. The assumption is that longer words tend to be more formal. The following formula assigns a word a formality score based on its length in number of characters, l:

$$FS(w) = -1 + 2\frac{\log l}{\log L} \tag{2}$$

L is the maximum possible word length, an upper bound that is set to 28 characters, the length of the word *antidisestablishmentarianism*, one of the longest words in English, disregarding scientific words.

In the case of words joint by a hyphen, they were treated as one word and the average length between their components was taken, following Brooke.

Figure 1 shows the shape of the word length baseline. The boundary between categorical informality and formality (that is, a score of 0) is between 5 and 6 characters: all words with 6 letters or more will be considered formal.

4.4.2 LSA-based score

The method we will describe in this section was chosen from a list of methods that Brooke (2014) applied for its good results and ease of implementation. Under this method, formality scores are based on the distributional similarities between target words and seed words. First, a term-document matrix was created from our filtered Mixed Corpus. We chose document as context motivated by the assumption that formality is kept equal across a document. The weights in this matrix were 1 for presence of a word in a document and 0 for absence. LSA was applied to this matrix with a resulting dimensionality k of 3, compressing and reducing the data considerably. These choices are partially based on Brooke's work and on availability of computer memory. Assuming formal seeds have a score of +1 (extremely formal) and informal seeds a score of -1 (extremely informal), a preliminary formality score (FS') is calculated for every word w in the vocabulary using cosine similarity with the following formula:

$$FS'(w) = \sum_{s \in S | s \neq w} W_s \times FS(s) \times \cos(w, s) \tag{3}$$

Where w is the target word and s is a seed word belonging to the complete set of seeds S. $FS(s)$ is the assumed formality score of the seed (1 or -1). W_s is a weight that compensates for the fact that there are more informal seeds than formal seeds, and seeks to balance this difference. It is just the ratio of informal or formal seeds, and so its value depends on whether s is formal or informal. For formal seeds:

$$W_s = \frac{\sum_{s \in I} 1}{\sum_{s \in F} 1 + \sum_{s \in I} 1} \tag{4}$$

For informal seeds:

$$W_s = \frac{\sum_{s \in F} 1}{\sum_{s \in F} 1 + \sum_{s \in I} 1} \tag{5}$$

Where F is the set of formal seeds and I is the set of informal seeds. $\sum_{s \in C} 1$, where C is a class (formal or informal), is equivalent to the amount of seeds of the given class. This is divided by the total amount of seeds. In Equation 3, then, in order to obtain a preliminary score for a word, the cosine similarity of this word with every seed word (leaving this word out of the seed set if it is there[6]) is multiplied by 1 or -1, and weighted to account for the difference in size of the two seed lists. The resulting scores for every seed word are then added up.

Once the preliminary score for each word is calculated, the final formality score is refined to take into account any possible bias towards formality or informality in

[6] This is a form of leave-one-out cross-validation which ensures that training and test set are not equal: the score assigned to a seed word, which will be evaluated, cannot be based on the assumption that the very same seed word is formal/informal.

the corpus. This is done by taking the preliminary score of a function word —which should be neutral in terms of formality— as a reference, and recalculating the scores so that they range between -1 and 1 with the reference word as a midpoint. Similarly as in Brooke (2014), the function word chosen was *and*, because it appears in most documents without being related to their formality. The final formality scores (FS) are obtained with the following formula:

$$FS(w) = \frac{FS'(w) - FS'(r)}{N_w} \tag{6}$$

Where r is the reference word and N_w is a normalisation factor. For words that have been preliminarily assigned a score higher than the preliminary reference score, that is, words that are at this stage considered formal, the normalisation factor N_w is:

$$N_w = \max_{w \in F'}(FS'(w) - FS'(r)) \tag{7}$$

(Where $F' = \{w|FS'(w) > FS'(r)\}$, that is, F' represents the set of all words whose preliminary score is higher than the reference score).

For words that have been preliminarily assigned a score lower than the reference score, N_w is:

$$N_w = \max_{w \in I'}|FS'(w) - FS'(r)| \tag{8}$$

(Where $I' = \{w|FS'(w) < FS'(r)\}$, that is, I' represents the set of all words whose preliminary score is lower than the reference score).

The normalisation factor is equivalent to the preliminary scores that are farther away from the reference in both directions, that is, the maximum (Equation 7) and minimum (Equation 8) preliminary scores. Equation 6, then, normalises formality scores to fall between 1 and -1.

4.5 Experiments and Evaluation

We ran two different experiments: Experiment 1 does not account for phrasal verbs, whereas Experiment 2 does, as explained in Section 4.2.1. In addition to the baseline, we compare our results with one of Brooke's best systems, which uses LSA with $k = 20$ on the (filtered) ICSWM corpus.

Three different types of accuracy were calculated for all systems:

- **Class-based accuracy on the seeds**: The percentage of covered words in the two seed lists that were assigned an appropriate individual score; that is, a positive score in the case of formal seeds and a negative score in the case of informal seeds. **Coverage** indicates the percentage of covered words, that is, seed words that appeared in the corpus.

- **Pairwise accuracy on the seeds**: For every pair of informal and formal seeds, resulting from an exhaustive pairing between both lists, the percentage of covered pairs where the informal word was assigned a lower score than the formal word.

- **Pairwise accuracy on the CTRW set**: The Choose the Right Word (CTRW) set is a list of pairs of near-synonyms that differ in formality. This list was extracted from a book with the same name (CTRW, Hayakawa 1994) and Brooke (2014) made it available on his website[7]. The accuracy on this list of near-synonyms is of special interest to this thesis because it directly tests the applicability of these scores to near-synonyms. It is not possible to calculate a class-based accuracy on this list of near-synonyms pairs because they were not explicitly described as being formal and informal counterparts, but rather one of the near-synonyms in a pair is *more* formal than the other.

4.6 Results and Discussion

	Accuracy on Seeds			Accuracy on CTRW	
Method	Coverage	Class-based	Pairwise	Coverage	Pairwise
Exp 1	97.1	93.6	**98.9**	97.7	77.4
Exp 2	97.1	**94.1**	**98.9**	97.7	76.9
Brooke's system	100	93	98.4	99.7	**81.9**
Word length	100	86.4	91.8	100	63.7

Table 1: Accuracy of formality scores on the seed set and the CTRW set

Results are shown in Table 1. As a first remark, we observe that there is no big difference between both of our systems (Exp 1 and Exp 2). This comes as no surprise since their only difference (the presence of phrasal verbs as individual words) is not present in the seeds nor in the CTRW word set. The fact that Experiment 2 is better, even if not by much, can, on the contrary, be surprising at first. With phrasal verbs, the size of the vocabulary increased considerably (173.439 words in Exp 1 and 181.265 in

[7] `http://www.cs.toronto.edu/~jbrooke/`

Exp 2), which makes some verbs have fewer observed contexts to build a semantic space from. There is, though, one advantage in considering phrasal verbs. Phrasal verbs are meaningful lexical items which typically have a different meaning than the one of their main verb. To illustrate, take the verbs *turn* and *turn up*. They have the same main verb, but their meaning is quite different. *Turn* is therefore an ambiguous word. If we account for phrasal verbs, we remove one of the meanings of the form *turn* and place it into another form, *turn_up*, reducing ambiguity, specialising senses —which have, after all, distinct contexts— and, as a result, having better defined vectors, which could be responsible for this small increase in accuracy.

All systems performed better on the seed sets than on the bigger CTRW. A possible explanation for this is that whereas formality differences between seed lists are extreme (*cop*, *furthermore*) —or at least this is what they were designed for—, the CTRW list has words that differ in formality more slightly (for example, *unchangeable* and *immutable*). Both Exp 1 and Exp 2, which outperform the selected Brooke's system on the seeds, perform more poorly than Brooke's on CTRW. It is also true that our system's coverage is lower; but if we bear in mind that our scores were generated from a much smaller corpus (around 133 million words against Brooke's 1.3 billion words in their unfiltered form), its good performance is still noteworthy. In spite of the high accuracy of the very simple word length baseline on seeds, it remains below all other methods.

An exploration of the errors showed that a number of them took place in seeds that can have another meaning of a different formality. This is the case of some informal words, for example *con*, which can be a disadvantage (as in "pros and cons"), apart from a swindle; *buck*, which can be an animal, especially in singular (its meaning of dollars would most typically be expected in plural), or *scab*, which, apart from being "someone who does not take part in a strike", is also a crusted wound. Seed lists should be designed very carefully so as to avoid words with multiple meanings that have different formality. It is true, however, that, unless we account for different senses, formality-ambiguous words are inevitably going to have an inappropriate score.

These results confirm that it is possible to induce formality at the lexical level with fairly good results, even with a relatively small corpus and taking phrasal verbs into account.

5 Modelling Attitude

After having assigned words a formality score, we want to characterise their orientation in terms of attitude. We do this based on the concept of polarity, as described in Section 2.2.2.

Polarity and formality may behave similarly in some aspects: it is likely that formal texts tend to have words of a more neutral polarity than informal texts. However, a big difference between them is that, whereas formality tends to be kept constant within a document, because the situational aspects that require a specific level of formality do not —or should not— change, this is not necessarily the case with polarity. Polarity tends to be studied at the sentence level, rather than at the document level. For example, a movie review can outline both its positive and negative aspects. And even within a sentence, polarity can be mixed.

We obtained scores in a similar way as for formality, by also assigning them a score based on their distributional similarity with seed words that are extremely positive (+1) and extremely negative (-1). We made use of the concept of polarity, expecting polar words to carry more attitude and non-polar words (those with a score closer to 0) to be attitude-neutral. Although we ultimately wanted to capture the distinction between polar and non-polar, we did not use non-polar seeds because it is not clear what an "extremely non-polar" word would be.

We first applied the same method as described in the last Section (4) to see if the hypothesised similarity between formality and polarity allows us to obtain good results despite using documents as context. After that, we applied the same method but with sentences as contexts. Because of memory limitation reasons, the corpus had to be further reduced for that experiment. In order to allow for comparison, we repeated the experiment for documents with the reduced corpus.

In Section 5.1, we describe the creation of seed lists. Section 5.2 explains the experiments that were carried out. We describe the evaluation method in Section 5.3. Results are presented and discussed about in Section 5.4. Finally, in Section 5.5 we investigate the quality of our seed words.

5.1 Seed words

Positive and negative seeds were created based on the Subjectivity Lexicon (SL) (Wilson et al., 2005), a collection of subjective words annotated with the strength of their subjectivity (a word can be strongly subjective, if it can be considered subjective in most contexts; or weakly subjective otherwise) and their "prior polarity", that is, their polarity out of context. Its contents were partially manually extracted. This dataset fits our needs because it provides us with information on the a priori polarity direction of a word as well as on whether this word is typically used subjectively. We selected 30 words for each seed list, all being considered as strongly subjective by SL, and with a negative or positive polarity, accordingly. Seeds can be found in Appendix A.

5.2 Experiments

Note that we applied the method described in Section 4.4 in three different settings, which required a different preprocessing. Given the small but positive effect on the results that we observed when using phrasal verbs (Section 4.6), we included them as part of the vocabulary. After applying the general filtering described in Section 4.2.1, some more specific procedures were needed:

First, we applied the same method as for formality (**Experiment 3**). Therefore, we followed the same filtering as much as possible: Documents with fewer than 100 words were removed, and documents which did not contain any of the seeds were also removed.

In the second setting (**Experiment 4**), we applied the method to sentences, because they are more likely to keep polarity stable than whole documents. Text was split into sentences before punctuation and symbols were removed. Since units in this setting were sentences, we did not remove any "document" because of its length or the absence of seeds, and included all sentences as they were. Due to memory limitations, it was not possible to carry out this experiment as it had initially been designed. Therefore, further filtering was required: The corpus was lemmatised and stop words —including frequent words such as prepositions and pronouns— were removed, with the exception of the word *and*, which, as explained in Section 4.4.2, was used as a reference word.

Since Experiments 3 and 4 had originally been designed to be comparable, differing only in context type (documents in comparison to sentences), and the corpus had to finally be altered in the second setting due to memory limitations, we added a new experiment, **Experiment 5**, in which we used the same corpus as in Experiment 4 (lemmatised and without stop words), but with documents as units, instead of sentences.

Scores for attitude were calculated analogously to those for formality (see Section 4.4.2).

5.3 Evaluation

We compare our systems to the majority baseline (which is equivalent to the proportion of the bigger test word list, 50% in all cases) and to an upper-bound based on the General Inquirer lexicon (GI), which (Stone et al., 1966) was built in the following way:

In the cases where seed words were divided by senses in the GI, we only considered the seed as correct when all of its senses had the same assumed orientation. That is, if, according to the GI, one sense of a word was positive and another sense was not, this was counted as an incorrect seed. After all, seeds must be unambiguous representatives of their polarity. Since GI does not assign polarity scores to words, but only classes, only class-based accuracy can be calculated with it.

Apart from class-based and pairwise evaluation on the seed lists, analogously to experiments on formality (as described in Section 4.5), a second set of words was created to evaluate attitude scores in the polar vs non-polar distinction, in addition to the positive-negative distinction. As explained in Sections 2.2.2 and 3.3, polarity is not necessarily a dimension of difference among near-synonyms per se; rather, we hypothesise, polarity extremes correspond to words with an unneutral attitude, whereas words without polarity are not attitude-carrying. This second evaluation setting consists of two lists: a polar list (including all 60 positive and negative seeds), and a non-polar list. The non-polar list was built based on SentiWordNet (Baccianella et al., 2010), taking words that had the maximum neutral score (1) and a positive and negative score of 0 in all of their senses. 60 words were then manually selected out of those to keep a balance in the test set. The neutral word list can be found in Appendix A together with polar seed lists. Class-based accuracy is not available for this set of words because no number range has been specified, or hypothesised, to classify words into polar and non-polar. In order to evaluate scores with these two new word lists, their absolute value was taken, with 1 corresponding to a polar word (negative or positive) and 0 corresponding to a non-polar, neutral word.

5.4 Results and Discussion

As shown in Table 2, performance on the polar seeds was remarkably low in all cases, but especially pairwise, reaching a value as low as 5.3% in Experiment 5. Surprisingly, documents (Exp 5) had a slightly better class-based accuracy than sentences (Exp 4), even if all class-based accuracies were disappointingly very close to the simple majority baseline (50%). With 91.8% accuracy, the GI upper bound is far beyond our three

	Polar seeds			Polar vs non-polar	
Experiment	Coverage	Class-based	Pairwise	Coverage	Pairwise
Exp 3	100	50	10.3	100	86.6
Exp 4	86.7	53.8	8.6	87.5	65.6
Exp 5	86.7	55.8	5.3	87.5	82.1
Majority baseline	100	50	50	100	50
GI upper bound	81.7	91.8	-	-	-

Table 2: Accuracy of attitude scores

systems.

Interestingly, accuracy on the polar/non-polar set of words is instead much higher: except for Experiment 4, with sentences as context, the accuracy of our other two systems is remarkable; and all experiments outperformed the baseline.

Altogether, results seem to indicate that, even though scores can capture whether a word is polar or neutral, they fail to represent the actual (direction of) polarity of words. Furthermore, it is not confirmed that sentences are more useful than documents for our goal.

There are many possible reasons for performance being so low in the positive/negative distinction: badly chosen seeds (we will have a closer look at that in Section 5.5), an insufficient number of dimensions in LSA, too little data, a wrong conceptualisation of attitude as polarity, among others. It is possible that contexts in which negative and positive words appear do not differ as much as formal and informal contexts and, therefore, other approaches are needed.

As a first exploration to understand what was happening, we observed the histograms of scores (Figures 2, 3 and 4), where it is obvious that something was not working as expected. In them, the x-axis represents scores and the y-axis, their frequency (the number of words that had a score falling in that interval).

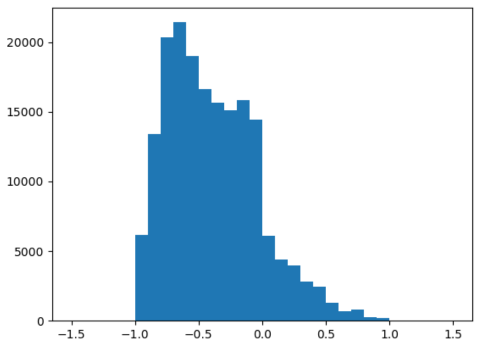

Figure 2: Exp 3 (documents, whole corpus)

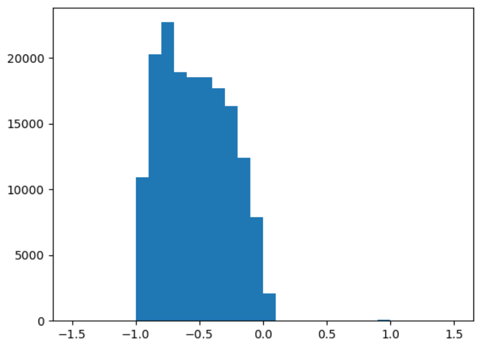

Figure 3: Exp 4 (sentences, lemmatised)

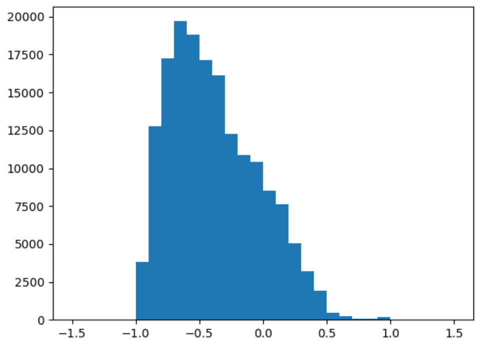

Figure 4: Exp 5 (documents, lemmatised)

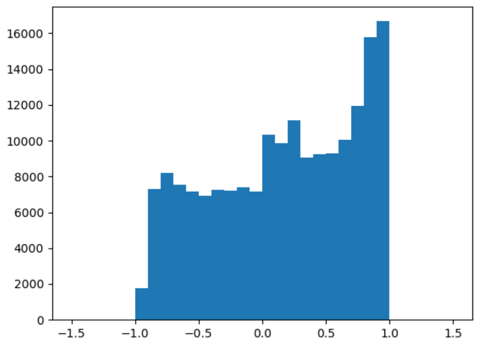

Figure 5: Exp 2 (formality, lemmatised)

It becomes clear from the histograms that attitude scores from all experiments (Figures 2, 3, and 4) are biased towards negative. They all have a different shape than formality scores (Figure 5). With most scores being negative, most words were assigned negative polarity and this explains why polar class-based accuracy was close to 50%. Given the good results on polar/non-polar words, we expect most extreme words —positive or negative— to have the lowest scores.

The reason of this bias remains unclear. A similar approach using LSA for capturing polarity had been implemented by Turney and Littman (2003), who used a smaller set of seeds (14 in total), a larger number of dimensions (in the order of hundreds), and a different corpus. It is possible, but out of our possibilities for testing, that a much bigger number of LSA dimensions could improve these results.

We can, however, explore the appropriateness of our seed lists by treating them as clusters: this is explained in the next Section (5.5).

5.5 Comparison of seeds

As it could be observed, the quality of the results for formality and for attitude differ substantially. The reason may lie in different factors, as we discussed, or in a combination of them. We decided to look at the quality of the chosen seed words objectively to try to clarify the origin of the problem. In order to do so, we treated seed lists as word clusters and computed a metric of clustering evaluation. We hypothesise that a set of seed lists is a good one if seeds in one list are similar among each other and remarkably different from seeds in the other list(s). In other words, we prefer to have low intra-cluster distances and high inter-cluster distances.

After running LSA (Section 4.3), we obtained a reduced matrix where all words receive a condensed vector, which in our case is of dimensionality $k = 3$. This will allow us not only to compare the metrics of cluster cohesion, but also to visualise seeds in the reduced semantic space. The measure used is described in Section 5.5.1 and results are presented in Section 5.5.2.

5.5.1 Davies-Bouldin index

The Davies-Bouldin index (DB) (D. L. Davies & Bouldin, 1979) is a metric of internal cluster evaluation. Evaluating a clustering internally means that the evaluation is solely based on the clustered data, without relying on any external, ground truth data or using the clustering results in an application. Internal evaluation metrics measure the quality of a clustering, rather than its particular usefulness in a desired setting. This metric was chosen for giving better (lower) scores to clusters with low intra-cluster distances and high inter-cluster distances. It is calculated in the following way:

First, S_i is needed for every cluster:

$$S_i = \left(\frac{\sum_{j=1}^{T_i} |X_j - C_i|^p}{T_i} \right)^{1/p} \tag{9}$$

Where T_x is the size of a cluster x, C_x is the centroid of cluster x, and X_k is a data point belonging to this cluster. p is generally set to 2, which makes the expression be equivalent to a Euclidean distance. Overall, with $p = 2$, S_i is the square root of the average distance of data points in this cluster i to its centroid.

Next, $M_{i,j}$ has to be calculated:

$$M_{i,j} = ||C_i - C_j||_p = \left(\sum_{k-1}^{n} |c_{k,i} - c_{k,j}|^p \right)^{1/p} \tag{10}$$

Where, again, C_x is the centroid of cluster x, and $c_{k,i}$ is the k-th component of the centroid C_i. $M_{i,j}$ is, essentially, the square root of the distance between two cluster centroids. Again, setting p to 2 makes this a Euclidean distance. Ideally, between two clusters, $M_{i,j}$ has to be maximised, whereas S_i should be minimised. A good cluster separation measure, $R_{i,j}$, that satisfies this and other basic requirements, is:

$$R_{i,j} = \frac{(S_i + S_j)}{M_{i,j}} \tag{11}$$

The lower the value, the more optimal the clusters are according to these criteria. The final index is then calculated as follows:

$$DB = \frac{1}{N} \sum_{1=N}^{N} max_{j \neq i} R_{i,j} \tag{12}$$

Where N is the number of clusters, 2 in our case.

5.5.2 Results and Discussion

Experiment	Brief description	DB-index
1	Formality, no phrasal verbs	1.742
2	Formality, with phrasal verbs	**1.729**
3	Attitude, same corpus as 2	3.249
4	Attitude, sentences, lemmatised	4.428
5	Attitude, documents, lemmatised	4.348

Table 3: Clustering cohesion results (the lower, the better)

From what numbers show (Table 3), it is clear that formal seeds have been better clusters in their two respective experimental settings than polar seeds in any of their three different settings, because both of them have a lower DB score. Vectors of seeds in Experiment 2 were the best clustering of all, whereas the worst experiment was Experiment 4, which used sentences as context. Since there is no direct interpretation for the DB index —although relative scores are themselves very informative—, and since 3 dimensions are still graphically representable, we decided to have a look at how the seeds look like in space, to have a more intuitive idea of how good and bad these clusters really are. In the 3D plots (Figures 6 and 7), the three axes represent the three uninterpretable "latent topics" that were chosen as the number of dimensions.

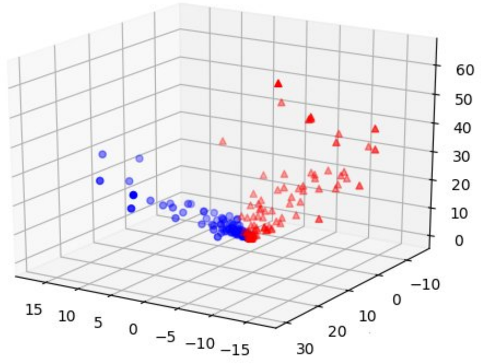

Figure 6: Formality seeds of experiment 2 on the compressed semantic space. Informal seeds are blue, formal seeds are red.

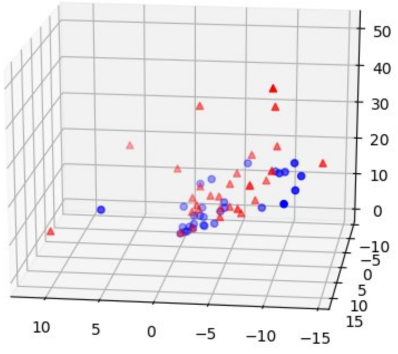

Figure 7: Polar seeds of experiment 3 on the compressed semantic space. Negative seeds are blue, positive seeds are red.

By looking at the plots, one can visualise the substantial difference between formality and polar seeds. Formal and informal seeds form two well enough differentiated groups that can be quite well separated by a straight line. Although many seeds from different formality levels have the tendency to clump together, some of them are more far apart, being probably better representatives of their formality level. It would be possible to identify those which are in the center and remove them. We would expect that such a seed set, even if smaller, could lead to better scores. We leave this, however, for future work.

With such results on clustering cohesion, it is not surprising that the assigned scores for attitude are much worse than those for formality: according to our assumption, seed words have to be good clusters in order to obtain good scores. If the reference words on which the score assignation is based are bad clusters, that is, if there is no clear distinction between them in their vector representation, they cannot lead to clearly distinguished scores. At the same time, we do not rule out that a different number of dimensions could have led to better clusters, but limitations in computer memory did not allow us to try the experiment with a higher dimensionality as has been done in the past (Turney & Littman, 2003).

Again, this adds evidence to the previously mentioned idea (see Section 5.4) that words with different polarity may not appear in such different contexts as formal and informal words do, possibly because they can be used to describe the same, in principle neutral, situations (*the party was great* and *the party was horrible*).

6 The Relation between Formality and Attitude

As we suggested in Sections 2.2.3 and 3.4, there are reasons to believe formality and attitude interact with each other. With a formality and an attitude lexicon enriched with scores, we can check this idea by running a correlation analysis between the two. While it is true that the quality of our attitude lexicon is very low, results on its intrinsic evaluation showed that they are not so bad at distinguishing polar from non-polar words.

A Spearman's correlation was run to determine the relationship between formality scores and the absolute values of attitude scores. There was a strong, negative and statistically significant correlation $r_s = -0.64, p = 0.0$) (Figure 8). Lexicons of experiments 2 and 3 were used, because they had the highest accuracy and both modelled phrasal verbs. These results seem to point to the direction of our hypothesis: informal words are more polar (are more extreme in sentiment), and formal words are more neutral.

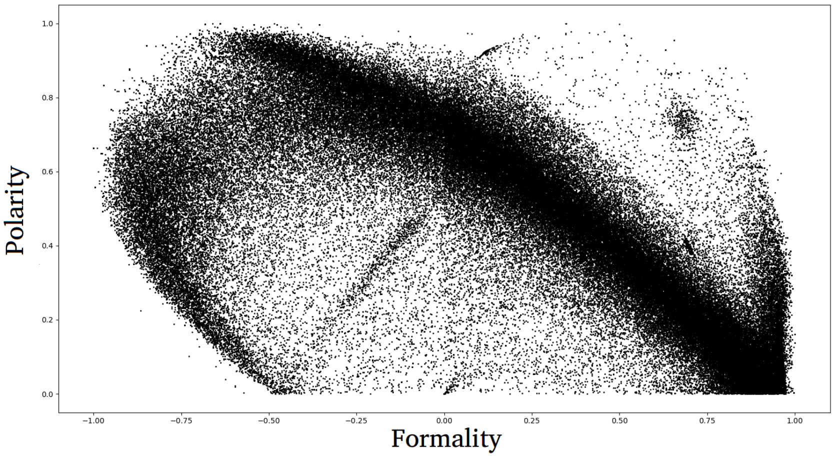

Figure 8: Formality (x axis) and Absolute polarity (y axis).

By looking at the plot, one can see a thick line that pictures the captured tendency. The other interesting shape is the one covering the informal words that are most informal (from -1 to -0.5, approximately). Interestingly, this range of words does not seem to reach levels of polarity as high as those seen in less extremely informal words.

When assessing these results, one must take into consideration as drawbacks that even weak relationships can show up as significant with such a big amount of data points; in addition to the uncertain quality of the attitude lexicon. This inconvenience can be overcome by using a different lexicon. This analysis was therefore repeated with SentiWordNet (SWN) (Baccianella et al., 2010), which is believed to have more reliable scores. Since SWN deals with senses and our formality lexicon does not, rather than mixing the formalities of all senses in one score, it was necessary to merge the SWN scores for different senses of a word. Although senses are typically ordered by frequency of usage, this is difficult to determine rigorously, so scores were averaged across senses. For every word, its final score was the average of positive scores minus the average of negative scores. In such a way, a word that has both positive and negative senses will end up having a score that is closer to neutral, as would be expected in our scores. Objective scores were not used for considering them, in most cases, redundant.

The resulting relation[8] does not show any kind of linearity, as we can observe in Figure 9. SWN scores, even though continuous, tend to be round numbers, which is, we think, reflected in the horizontal lines of the plot. Such a data configuration does not fit with our hypothesis. It is important to remark, however, that a more sophisticated way of assigning them continuous scores could have been used, and that SWN is, after all, not a perfect resource.

The main conclusion to draw from both analyses together is that we do not have strong evidence for the hypothesised relation between formality and attitude.

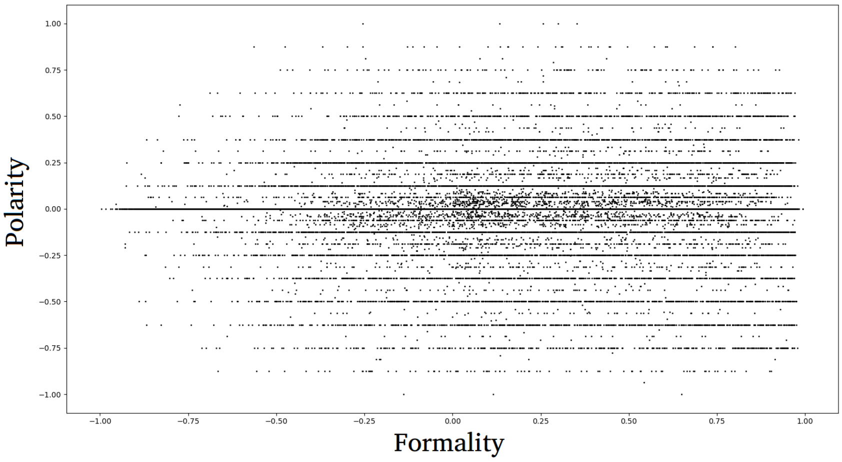

Figure 9: Formality (x axis) and Polarity (y axis) from SentiWordNet.

[8] Note that, since we are not using the absolute value here, the expected shape (if the hypothesis was true) would be rather triangular.

7 Modelling Collocations

Collocations are the third and last dimension that we wanted to induce in near-synonyms. As said in Section 2.2.4, words can have "preferences" when it comes to co-occurrence: a combination of words that would otherwise be completely meaningful can turn out to be unused and even sound strange for native speakers. Collocations are especially difficult for second language learners in production, rather than in comprehension, because of their conventionality.

Collocational preferences were extracted from the Mixed Corpus in its general filtered form (see Section 4.2.1). Collocations were induced in the form of rankings based on three different association measures that we will describe in detail in the next section. Every measure gives a different ranking of collocations, emphasising different aspects, and having three different rankings to consider allows us to make the most of the advantages of the three, therefore making better decisions when it comes to choosing the best collocation between two options.

7.1 Procedure and measures

Collocations were extracted from our corpus following Inkpen and Hirst (2002). First of all, proper names were removed from the corpus because they were not of our interest. They could easily show up as relevant collocations, even though they should not be able to make a difference between near-synonyms. After the general filtering, Pedersen's Ngram Statistics Package[9] (NSP) was used (Banerjee & Pedersen, 2003). NSP can automatically extract n-grams and their counts from corpora and compute their statistical association with various association measures. It outputs a rank of collocations ordered by the strength of their association.

The association measures we used were the Log-likelihood ratio, the Odds ratio and Pointwise Mutual Information. These measures were chosen for having been used in the past to extract collocations or for related uses with good results (Dunning, 1993; Pecina, 2005; Blaheta & Johnson, 2001), and for being non-parametric. Non-parametric measures are more suitable for this task than parametric measures because the latter assume a particular distribution of the data that does not typically fit the real Zipfian distribution of linguistic data. Pearson's Chi-square was considered at a first stage but it was dismissed for not being as suitable for sparse data as the log-likelihood ratio (Evert, 2005). We will now discuss each measure in detail.

[9] NSP is available at `http://www.d.umn.edu/~tpederse/nsp.html`.

7.1.1 Log-likelihood ratio

The likelihood ratio between the observed and the expected is one of the most straightforward ways to compute association strength between words. Due to the typical Zipfian distribution of linguistic data, which can make many expectation values much smaller than 1, it is usually computed on a base-2 logarithmic scale.

Words in bigram	word 1	¬ word 1	Totals
word 2	O_{11}	O_{12}	O_{1-}
¬ word 2	O_{21}	O_{22}	O_{2-}
Totals	O_{-1}	O_{-2}	O_{--}

Table 4: Contingency table, where the indexed O's are placeholders for frequency counts

In our particular example (Table 4), O_{11} holds the number of times the bigram (word 1, word 2) has been observed. O_{12} represents the number of times word 2 has been seen forming a bigram with a word other than word 1, and so on. O stands for *observed*, and underscores represent all possible rows or columns (1 and 2), whose values are added up. The respective expected values (E) for each cell are calculated from the row and column totals in the contingency table:

$$E_{ij} = \frac{O_{i-} \times O_{-j}}{O_{--}} \tag{13}$$

Finally, the log-likelihood ratio statistic G is calculated in the following way:

$$G = 2 \times \sum_{ij} \left(O_{ij} \times \log \frac{O_{ij}}{E_{ij}} \right) \tag{14}$$

7.1.2 Odds ratio

The odds ratio is an association measure that indicates the degree to which one variable influences another. In our setting, it compares the odds that a word A appears given another word B to the odds of word A occurring in the absence of word B. In the above contingency table, the odds ratio would be calculated as follows:

$$OR = \frac{\frac{O_{11}}{O_{21}}}{\frac{O_{12}}{O_{22}}} \tag{15}$$

Which is equivalent to:

$$OR = \frac{O_{11}O_{22}}{O_{12}O_{22}} \tag{16}$$

7.1.3 Pointwise Mutual Information

Pointwise Mutual Information (PMI) is a measure that comes from information theory. It ultimately quantifies the dependence between two variables, or “the amount of information” that we can learn from one variable about the other. In case of total independence between variables, PMI is 0.

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \tag{17}$$

In Equation 17, x and y stand for two different words, and $p(x)$ is the probability of encountering word x.

Extracted information on collocational preferences was only extrinsically evaluated in the Fill-in-the-Blanks task (see Section 8).

8 The Fill-in-the-Blanks task

The Fill-in-the-Blanks (FITB) task (Edmonds, 1997) is a way of evaluating the lexical choice capacity of a system which consists of choosing, from a limited set of choices, a word to fill in a blank space in a sentence. The correct option is the word the author of the sentence chose when writing it. In its first version, the task was performed with 7 different sets of near-synonyms on the 1987 portion of the Wall Street Journal corpus (WSJ), and it has been mostly replicated with the same settings ever since. As discussed in Section 3.1, the task has some weak aspects that can limit its usefulness as an evaluation method. We decided to apply some changes to the original task, which would limits comparability with other works but at the same time makes the task more appropriate for our specific setting. These changes are described in Section 8.1. Our goal is to discover if using information on specific dimensions of difference of near-synonyms, like the information we acquired in previous sections on formality, attitude and collocations, can help in this task. The baselines used in all experiments are presented in Section 8.2. Sections 8.3 and 8.4 describe our systems and their results: respectively, those which use the information about each dimension separately, and those which use it in combination. Finally, we tested the best combined systems on a bigger list of near-synonym sets extracted from WordNet (Section 8.5).

8.1 Changes to the original task

The first change we made was to replace the corpus from which sentences are extracted. The WSJ corpus has the disadvantage of representing mainly one genre, namely newswire, and only one variety of English, American English. Ideally, we wanted a corpus that represents a wider range of genres and varieties, especially because we want to experiment with different levels of formality, but also because our Mixed Corpus, the corpus the extracted information is based on, is more representative of the varieties of English. We therefore decided to use the British and American freely available portions of the GlowBe[10] (M. Davies & Fuchs, 2013) corpus, which is made of informal blogs (60%) and other types of texts found on the web (newspapers, company websites, etc.). The two mentioned portions make up a total of around 800,000 words together.

The 7 original near-synonym sets are a very reduced test set and they do not necessarily correspond to the three dimensions we work with. We decided not only to find near-synonym pairs or triples that would account for our distinctions, but also to in-

[10] `http://corpus.byu.edu/glowbe/`

Formality-based	Attitude-based	Collocations-based
prize, award	fat, overweight	difficult, hard, tough
basic, fundamental	conspiracy, arrangement	error, mistake, oversight
busy, occupied	cowardly, cautious	job, task, duty
whole, entire	reject, decline	responsibility, commitment, obligation, burden
discover, find out	worry, concern	material, stuff, substance
lucky, fortunate	acknowledge, admit	give, provide, offer
trip, voyage	drunk, alcoholic	settle, resolve
sort, type	debate, discuss	customer, client
hire, employ	bias, prejudice	tell, say
goal, objective	ignored, overlooked	color, colour
remember, recall	stubborn, persistent	forest, woods
disclose, reveal	anger, rage	command, order
chat, dialogue	threat, warning	skinny, thin, slim
speed, velocity	funny, hilarious	wedding, marriage
teach, educate	tasty, delicious	dad, father
huge, enormous	tragedy, incident	honest, sincere
purchase, buy	violence, force	rich, wealthy
quick, rapid	beautiful, gorgeous	outgoing, friendly
sight, vision	confident, proud	valid, legitimate
stick, adhere	battle, conflict	weak, fragile

Table 5: The 60 chosen near-synsets.

crease the test size to 20 near-synonyms per dimension (that is, 20 for formality, 20 for attitude and 20 for collocations), in addition to the 7 original near-synonym sets.

The quality of the three kinds of extracted information on near-synonyms was evaluated on the FITB task both separately —dimension by dimension— (Section 8.3) and together, in a combined setting (Section 8.4). In order to create test sets that focus on one dimension only, the new near-synonym pairs were manually chosen from (M. E. Gardiner et al., 2013; Edmonds, 1999; Hayakawa & Ehrlich, 1994). After an exploration of the corpus, two basic criteria were established for accepting a set of near-synonyms:

1. Every near-synonym has to be present at least 10 times in the corpus;

2. None of its members must appear more than 90% of the total frequency of the near-synonym set. This filter was applied to ensure that there was a real lexical choice difficulty between near-synonyms and that it was represented in the corpus.

Of course, near-synonyms do not necessarily differ in one dimension only. However, lists were especially created to focus on at least one of the three. These lists of near-synonyms can be found in Table 5.

As a first step after obtaining the lists, the corpus was lowercased and segmented into sentences. For every set of near-synonyms, the relevant sentences —those that contain the near-synonyms in question— were extracted. The space the near-synonym was occupying is left blank for the system and a decision for every sentence has to be made.

We will proceed to present the baselines used and to describe how our system makes decisions.

8.2 Baselines

Two very basic baselines were created to compare our results with. The first baseline is a simple **majority class baseline**, where the most common near-synonym in the set —the one which appears in most sentences— is selected for every sentence. The second baseline is based on a **language model**. A trigram language model based on the Mixed Corpus was built with the SRILM tool (Stolcke et al., 2002) with Kneser-Ney smoothing. The language model-based baseline consists of calculating the probability of a sentence with every possible candidate and selecting the candidate that maximises sentence probability.

8.3 Separate Setting

In the separate setting, we test systems that use only one of the three kinds of information that we have extracted. Each system is then evaluated on the corresponding set of near-synonyms.

8.3.1 Formality

For every sentence that has a blank, an average formality score is calculated. In order to do so, all words of the sentence for which a formality score is available have their scores summed up and divided by the total number of the same words, obtaining the average formality FS of the sentence s:

$$FS(s) = \frac{\sum_{w \in (F \cap S)} FS(w)}{\sum_{w \in (F \cap S)} 1} \tag{18}$$

F is the set of words for which we have a formality score, and S is the set of words in the sentence. $F \cap S$ is their intersection, that is, it represents the words that are in the sentence for which a formality score is available. To determine the most suitable filler, the difference between its score and the average is calculated. We select the word whose score is closest to this average score.

8.3.2 Attitude

The selection method for near-synonyms differing in attitude is analogous to the method for formality, with the only difference being that the absolute values of the attitude scores are taken, disregarding their sign, instead of their real value. The reason for this is that, as shown in Section 5.4, these scores had an acceptable quality only when evaluating the polar and non-polar distinction in absolute values.

8.3.3 Collocations

In order to choose a candidate based on collocational preferences, the system uses the scores of those collocations that each near-synonym forms with the words that immediately precede and follow the blank. It is true that syntactic information could be used as well, but we wanted to base our method on purely lexical resources. Every association measure available (log-likelihood ratio, Odds ratio, and PMI see Section 7) emits one vote, corresponding to the near-synonym used in the highest-scoring collocation among all possibilities. For illustration, in the sentence:

> *our hope is large and sweet , so we wait with great expectation as a bride awaits her ---------- day.*

The possible near-synonyms for this sentence are *wedding* and *marriage*. The preceding and following words are *her* and *day*.

The three collocation measures emit their judgments (Table 6). The solution proposed by the system is therefore *wedding*, because it has more votes.

	Log-likelihood ratio	PMI	Odds ratio
her wedding	800.1	3.87	15.28
her marriage	**1210.31**	3.62	12.74
wedding day	1094.44	**5.49**	**46.97**
marriage day	10.67	-3.08	0.12
Vote	*marriage*	*wedding*	*wedding*

Table 6: Votes of the different collocations' systems

The actual solution, that is, the word that was present in the original text, is also *wedding*. In this case, the system's choice was successful. The advantages of using a voting system become clear in this example, where relying only on log-likelihood estimates would have led the system to the wrong solution.

In the case of a blank being at the beginning or the end of a sentence, only collocations for the previous or next word, accordingly, are compared. Similarly, when there is no score available for a collocation formed with either the preceding or the following word, only the existing collocations are compared. In the uncommon cases of there not being any collocation score available, the system backs off to the majority class baseline. The same strategy is applied in the very rare case of a tie between the three measures, which can only happen where there are at least three near-synonyms in a set. There were only two sentences with such cases amongst all the sentences for the 60 near-synonym sets, and they were both in the collocations' list of near-synsets.

8.3.4 Balanced data

We found the distribution of near-synonyms to be very skewed, which made the majority class baseline very good and difficult to beat: one of the near-synonyms of a set was usually much more frequent than the others, which is a normal phenomenon. In order to see how well our system would perform in a less biased setting, even if unnatural, a balanced test set was created where the number of sentences was kept the same for each near-synonym.

8.3.5 Results and Discussion

	Average accuracies		
Class	System	Majority Baseline	LM Baseline
Collocations	0.648	0.696	**0.759**
Formality	0.627	0.659	**0.764**
Attitude	0.578	0.685	**0.784**

Table 7: Accuracy of all systems on each dimension

	Average accuracies		
Class	System	Majority Baseline	LM Baseline
Collocations	0.534	0.500	**0.673**
Formality	0.640	0.438	**0.704**
Attitude	0.500	0.500	**0.711**

Table 8: Accuracy of all systems on each dimension with balanced data.

As shown in Table 7, none of our systems beats any baseline. The language model has the highest average accuracy in all three test sets, followed by the majority class baseline. Among our three systems, the system with the best result is the collocations system. This is not surprising considering it is taking into account the surrounding context, similarly to the language model. In addition, we must take into account that it is the only near-synonym list where there were sets with more than 2 near-synonyms, which increases difficulty. The fact that attitude is the worst system is also not surprising: we already know from Section 5.4 that attitude scores had a poor quality. Overall, using information on these dimensions separately does not seem to be a useful approach to this task, at least in the way we designed the different systems. Of course, the quality of the extracted information also plays a role. With balanced data, all systems that are somehow related to frequency see their accuracy decreased (collocations and the two baselines). Only the formality system performs slightly better. In any case, the language model baseline remains the most effective one.

No meaningful pattern could be found when looking at the individual decisions for each near-synonym set. Table 9 presents the best and worst results of each of our systems, that is, the near-synonym sets where each system obtained its highest and its lowest accuracy.

Class	System's best prediction	System's worst prediction
Collocations	command, order (0.829)	give, provide, offer (0.478)
Formality	huge, enormous (0.859)	sight, vision (0.302)
Attitude	stubborn, persistent (0.867)	tasty, delicious (0.188)

Table 9: Best and worst results on near-synsets of every system.

8.4 Combined Setting

In the separate settings presented previously, near-synonyms are assumed to differ at least in the corresponding dimension for which they are being tested in the FITB task. However, in a more natural situation of lexical choice, we would not have information about the nature of differences between specific near-synonyms. Therefore, a combined system —a system using all 3 kinds of information at once— was created and used on the 60 near-synonyms together (Section 8.4.1). Finally, and this time with the aim of improving our system's accuracy, a weighted system was created in order to diminish the impact of the weakest predictor, attitude. This weighted system was later separately enriched with frequency and with the language model, taking advantage of their quality (Section 8.4.2).

8.4.1 Simple combined system

In the simplest combined system, the individual systems of formality, attitude and collocations are used for every set of near-synonyms. Every system emits a vote for their most likely candidate and the near-synonym that received more votes is chosen. In the case of a tie between the three systems, which can only happen in the not so common cases where there are at least 3 near-synonyms, the collocations' candidate has priority. Priority is given to collocations because they have the best results in the individual systems' setting. Results of this system are found in Table 10.

8.4.2 Weighted system

Being conscious of the big impact of frequency and the usefulness of the language model, while trying to create a more accurate system, we created a weighted system that takes advantage of this information. The vote that frequency emits corresponds to the most frequent near-synonym in a set. Seeing the poor performance of attitude scores, we assumed that giving its votes a lower weight would increase the system's performance. We manually experimented with several combinations on the test set, the best ones being:

Formality = 0.3, Collocations = 0.3, Attitude = 0.1, Frequency = 0.3

Formality = 0.3, Collocations = 0.3, Attitude = 0.1, LM = 0.3

Results of weighted systems are found in Table 10.

8.4.3 Results and Discussion

Method	Accuracy
LM baseline	**0.769**
Weights + LM	0.738
Weights + Frequency	0.702
Majority baseline	0.680
Simple combined	0.658

Table 10: Accuracy of all systems in the combined setting on 60 near-synonym sets.

Results of the different combined settings are presented in Table 10. Again, the language model baseline outperforms all other systems. The simplest combination, or equal voting system, is the poorest performer of all, possibly because of attitude, but it does better than any of the three separate systems that we saw in the last section. Both the weighted system that includes frequency and the system complemented with the language model beat the majority baseline this time. It comes as no surprise that the latter has a higher accuracy than the former, because the language model performs better than all other systems in all settings.

A paired-samples t-test revealed that there is a significant difference in the performance of the weighted system with frequency ($M = 0.702, SD = 0.109$) and the performance of the majority baseline ($M = 0.68, SD = 0.110$), $t(59) = -2.11$, $p = 0.039$ on the list of 60 near-synonym sets. In other words, the weighted system with frequency performs significantly better than the majority class baseline. A similar analysis was run to assess the difference between the language model baseline and the weighted system helped by the language model. Again, a paired t-test showed that there is a significant difference between them ($M(lm) = 0.769$, $SD(lm) = 0.103$ and $M(wlm) = 0.738$, $SD(wlm) = 0.112$, $t(59) = 3.503$, $p < 0.001$). Therefore, the language model baseline performs significantly better than our weighted system with the language model.

8.5 WordNet setting

In order to test our best systems on a bigger amount of data, we extracted 1000 suitable synsets from WordNet. This list of synsets is not only much larger than our previous list, but it is also randomly created without any constraint on dimensions of difference between near-synonyms. Our WordNet synsets contain near-synonyms that can differ, in principle, in any dimension, or in several at the same time. In addition, the selected WordNet synsets have more words, on average, than the 60 near-synsets (2.29 against 2.13), which makes the task more difficult. With this experiment, we want to see how generalisable our results on the 60 near-synsets are. We expect our systems to have a slight decrease in performance, due to the fact that dimensions may not be as very well represented in this test set.

This WordNet-based list of synsets was obtained by first shuffling all noun, adjective and verb synsets from WordNet. Then, synsets were picked according to the following criteria:

1. All members of a synset must appear in our corpus with a frequency of at least 5;
2. No member of a synset can have more than 90% of all occurrences of members of that synset together.

These criteria were applied until the desired number of synsets of each part of speech (500 nouns, 250 adjectives and 250 verbs) was reached. See Appendix B for the WordNet list of synsets.

8.5.1 Results and Discussion

Method	Accuracy
LM baseline	**0.816**
Weights + LM	0.791
Weights + Frequency	0.714
Majority baseline	0.676

Table 11: Accuracy of all weighted systems on the WordNet set of near-synonyms.

Results of combined systems on the WordNet dataset are presented in Table 11. We can observe the same pattern of results as for the 60 near-synsets (Table 10), showing that our previous results generalise well on different data that were not manually selected. On them, the language model has the highest accuracy (0.816) of all settings where it was applied. Contrary to our expectations, all methods except for the majority baseline performed better on the 1000 WordNet synsets than on our carefully selected 60 near-synonym sets.

It is important to point out that, even though the language model is better at the FITB task than our systems, there might be situations where the latter are more relevant. For instance, in a situation where L2 learners have to do the FITB task, a possible application that teases the different dimensions apart could be useful in order to give the learners feedback on their mistakes.

9 The paraphraser

In addition to the evaluations on the FITB task we wanted to test the usefulness of the formality scores in an application. We decided to build a paraphrasing tool that produces paraphrases with a change in formality based on lexical substitution. Concretely, for any given input sentence, the paraphraser should create a more formal or more informal version. The idea that we had in mind is that such a tool, if it works properly, could help L2 learners of English who are not yet proficient enough to adapt their written language to the appropriate situation.

Section 9.1 presents the external resources needed to build the paraphraser and Section 9.2 describes the necessary preprocessing of the available resources. Section 9.3 explains the way the paraphraser works and its different modules. In 9.4 we describe the evaluation method that we chose for the paraphraser, and in Section 9.5 the results of this evaluation are presented and discussed. Finally, Section 9.6 introduces the main limitations of the system and proposes some improvements for future work.

9.1 Resources

The paraphraser has to produce a grammatical and fluent text that is equivalent in meaning to the input with a change in formality. The following are the elements used that contribute to this goal:

The **Paraphrase Database 2.0**, PPDB (Pavlick et al., 2015), is an automatically constructed collection of paraphrase pairs originally built by Ganitkevitch et al., (2013). Paraphrases were extracted from bilingual parallel corpora, using pivot languages to find different expressions with a shared translation. Pairs of paraphrases are ranked by quality based on a regression model that is fit to human judgments of paraphrase quality on a subset of the database. The paraphrase was automatically enriched with other types of information, such as entailment relations and scorings for style. However, after a closer examination of the database, we decided not to use this information for reasons of limited reliability, and in the case of style scores, for not being present for all paraphrase pairs.

The database comes in different types and sizes. Types correspond to lexical, phrasal or syntactic paraphrases. For the paraphraser, we used only lexical and phrasal paraphrases, because we wanted to investigate the effect of near-synonym substitution with a change in formality. Lexical paraphrases are single word paraphrases. The phrasal paraphrase pack includes not only phrasal verbs, which are of our interest, but also lexical substitutions with a larger word context, which can contribute to making the appropriate decision when, for example, substituting words with multiple meanings. Larger sizes of the database include paraphrases with a lower place in the ranking, which are of lower quality but lead to a higher recall. Since we were to further prune the database for formality and similarity, which would considerably reduce its size, we opted for using the largest paraphrase packs (XXXL size).

As a last remark, it should be mentioned that the database is available in 23 different languages (Ganitkevitch & Callison-Burch, 2014), which makes this application potentially reproducible for many other languages.

Google pre-trained vectors[11] were used for similarity judgments. They are a set of 3 million 300-dimensional word and phrase vectors trained on 100 billion words of the Google News dataset with an approach described in Mikolov et al., (2013). These vectors were used along with the Gensim software (Řehůřek & Sojka, 2010) to calculate cosine similarity scores for all paraphrase pairs that were left after the formality filter that will be described in the next section.

9.2 Preprocessing

Several steps of preprocessing are needed before the mentioned resources are used as part of the paraphraser:

9.2.1 Formality filtering

All paraphrase pairs for which we do not have formality information were removed from the PPDB. That is, only paraphrases whose words were in the formality lexicon we created were kept. The goal of the paraphraser is to create a change in formality, and no uninformed substitutions should be applied. For the remaining paraphrase pairs, a Formality Difference (FD) score was computed. In the case of lexical paraphrases, this was simply the difference in formality of both words:

$$FD(w_1, w_2) = FS(w_1) - FS(w_2) \tag{19}$$

If the second word w_2 is more formal than the first word w_1, $FS(w_2)$ is bigger than $FS(w_1)$ and the FD score is therefore negative. This means that, in order to paraphrase

[11] The archive can be found on `https://code.google.com/archive/p/word2vec/` .

to a more formal sentence, only paraphrases with a negative score should be chosen. For phrasal paraphrases, the FD score is:

$$FD(p_1, p_2) = FS(p_1) - FS(p_2) \tag{20}$$

Where $FS(p)$ is the average formality of all words in the phrase p:

$$FS(s) = \frac{\sum_{w \in p} FS(w)}{\sum_{w \in p} 1} \tag{21}$$

9.2.2 Similarity filtering

After the formality filtering, a similarity filtering was applied, where only paraphrases for which there was similarity information were kept. That is, a paraphrase was kept only if all words in the paraphrase pair had a vector in the Google vectors set. In the case of phrasal paraphrases, the words *of*, *and* and *a*, as well as punctuation marks were ignored because of their little semantic content, and there was no vector for them. Cosine similarity between phrasal paraphrases was calculated with the Gensim utility *n_similarity*, which takes the average vector of the words of each phrase and calculates the cosine similarity between them.

9.2.3 Named-Entity and other filterings

When inspecting the initial versions of the paraphraser, it was observed that it was making consistent errors on specific sets of words, namely, in countries, continents, days of the week, months, person names and surnames and personal pronouns. Not only were the paraphrases incorrect because there was an important change in meaning, but also because they were not really different in formality (except probably for cases of name diminutives, such as *Alex* instead of *Alexander*). Paraphrases including any of these elements were removed from the PPDB for these reasons.

A directly usable list of countries was obtained from the PyCountry Python package, and an extensive list of names and surnames was extracted from a website[12] (concretely, the lists of most popular first names and last names were used). Pronouns in all of their cases (*I*, *me*, *my*, *mine*, *myself*) and persons were also removed.

[12] `http://deron.meranda.us/data/`

9.3 Procedure

In this section, we explain the way the paraphraser works. In broad terms, this paraphraser follows the overgenerate-and-rank paradigm in Natural Language Generation. The process of paraphrasing can be divided into 4 main steps: Preprocessing of the input, generation of candidates, ranking of candidates and choice.

9.3.1 Input preprocessing

When a text is given to the paraphraser, it is first segmented into sentences and tokenised taking phrasal verbs into account. As input, the user should also indicate the direction of the paraphrasing: "to (more) formal" or "to (more) informal". Each sentence of the input is paraphrased separately.

9.3.2 Generation

There are two kinds of generation: word-based and phrase-based. In word-based generation, the system searches the pruned PPDB for word equivalents which have the desired change in formality. This is done word by word and the original word itself is always considered as well a candidate for the final paraphrase. In order to keep the search space smaller, options are limited to 3 possible candidates per word. In this way, with all candidates of every input word $(w_1, w_2, w_3, ..., w_n)$, the structure in Figure 10 is generated[13].

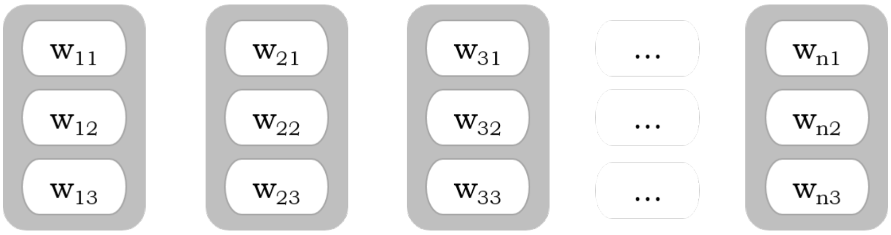

Figure 10: Word candidates

When a word has more than 2 candidates in the PPDB, the two candidates with highest similarity are kept, as long as they have a similarity of at least 0.6 with the input word. To generate all possible word-based paraphrases, we take the Cartesian product of this structure, as Figure 11 illustrates.

Phrase-based generation is more complicated because the length of phrases may vary, and the members of a phrasal paraphrase pair can have different lengths, which

[13] It is important to mention that the representation in Figure 10 assumes all words have at least two possible paraphrases, which in reality is not always the case.

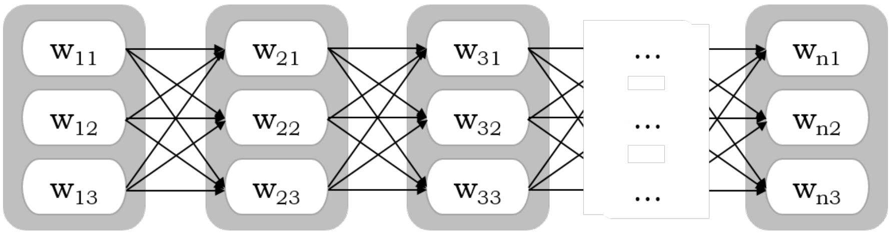

Figure 11: Generation of all word-based paraphrases

can lead to a difference in length between input and output.

In phrase-based generation, the system first partitions an input sentence into all possible and sensible phrase and word combinations. Sensible here means that no phrase can be longer than 5 words, which is the longest phrase available. A partition is, then, a way of segmenting a sentence into chunks. Only partitions are kept where all its phrases are present in the PPDB and have an equivalent paraphrase in the desired direction of paraphrasing (to formal or to informal). Similarly as in the word-based paraphrasing, for every phrase, only the two (if available) most similar paraphrases (with similarity ≥ 0.6) are kept as candidates. The very same phrase is also a candidate to take its own place in the sentence. Paraphrases for single words are reused from the word-based paraphraser. As a final step, for every remaining partition, the Cartesian product of all word or phrase candidates is applied, creating many candidate sentences. Finally, all different candidate sentences are put together and joint with the candidates that come from the word-based paraphraser in order to perform the next step: ranking.

It is easy to see how large the space of possible paraphrases becomes even for short sentences. As we will mention in Section 9.6, one of the main limitations of this system is that, with the computer used, it can only process sentences of up to 22 words. A clear improvement for the future would be to limit the space by preliminarily rejecting some partial paraphrase candidates before full sentences are completely formed.

9.3.3 Ranking

The ranking of all sentence candidates is based on what the desired characteristics of the paraphraser's output are. The crucial idea behind this ranking system is that a good paraphrase must:

- Have a change in formality in the requested direction;
- Have the same meaning as the input sentence;

- Sound natural and correct.

There are precisely three components that take care of each of these features. The acquired formality information is responsible for ensuring that there is a change in formality and for quantifying this change. The similarity information extracted from Google word embeddings compensates for the inexactitude of the PPDB, and forces words to have a more similar meaning. It is true that similarity does not imply preservability of meaning (*cats* and *dogs* are distributionally similar, but mean different things), but we believe this to be the best possible available approximation to it. Finally, our language model (built from the Mixed Corpus as described in Section 8.2) looks after the naturalness and fluency of the output.

Every candidate paraphrase is assigned a 3-dimensional vector where dimensions correspond to scores related to these three mentioned characteristics:

9.3.3.1 Formality The formality of a sentence is calculated as the average formality of those words of the sentence for which there is a formality score, exactly in the same way it was done for the FITB task (Section 8.3). This average formality is calculated both for the input and the candidate output sentence. The final score is the absolute difference between both of them:

$$FS(s) = |FS(i) - FS(o)| \tag{22}$$

Where i stands for input sentence and o for input sentence. By taking the absolute value we ensure that, regardless of the paraphrasing direction, the higher the score, the larger the difference in formality between input and output. Having a higher difference increases the chances of being able to obtain a perceived changed in formality. We are interested in quantifying this because, from what can be seen by looking at the formality scores, a small change in the formality scores does not always carry a perceived change in formality (two (near-)synonyms that are in principle equal in formality will very rarely have the exact same score).

9.3.3.2 Preserved Meaning The calculation of the similarity score of a candidate paraphrase is based on the average of the available pairwise similarities between the input sentence and the candidate sentence. Every word or phrase of the candidate sentence that has an available similarity score with its corresponding word or phrase in the input sentence (that from which it was generated) is used. It is true that this method is not sensitive to reordering, but based on the way the generation step works, phrase by phrase, we do not expect this to be a problem.

Of course, with such an approach, the best candidate sentence according to the similarity measure would be exactly the very same input sentence, because it would have the maximum similarity possible (1). Since no change in the input is generally not desirable, but still possible, two further impediments are later used to avoid this to happen: A lower weight for similarity in the weighted system, and the addition of a specific constraint in the choice module.

9.3.3.3 Fluency The fluency score is based on the perplexity of the sentence according to the language model. Perplexity is a measure that gives an indication on how well a language model can predict a sentence. It was chosen for being, in principle[14], immune to differences in sentence length. In general, the higher the perplexity, the less fluent a sentence is. We change the sign of perplexity to invert this, so that a higher score indicates a better paraphrase, just as in formality and similarity, the other vector components. Queries to the language model were carried out with the KenLM software (Heafield, 2011).

As a next step, in order to make the three different kinds of scores comparable among them, we normalise them to a score ranging between 0 and 1, where the minimum value is mapped to 0 and the maximum value to 1. This was implemented using the MinMaxScaler function available in the Scikit-learn Python package (Pedregosa et al., 2011).

Finally, a weighting scheme was applied. The weights were assigned manually, partially based on the author's personal linguistic judgments on the output of several trial sentences —different from those in the test set—, which is, admittedly, a weakness of the system. Another criterion was the fact that both similarity and formality information had already been used in the paraphrase generation module, which partly justifies a higher weight in the so far unused language model. It was also taken into account that, as has been said, similarity scores benefit sentences that are (almost) word-to-word equal to the input. Moreover, out of the three scores, the similarity-based score is the only one that is arguably representing the feature it has been assigned to, i.e. preserved meaning, which makes it less reliable.

A weighted sum of the scores gives the final ranking score, and the paraphrase p with the highest score is the "winning candidate":

[14] Of course, an unnaturally long sentence will leave a language model very perplexed.

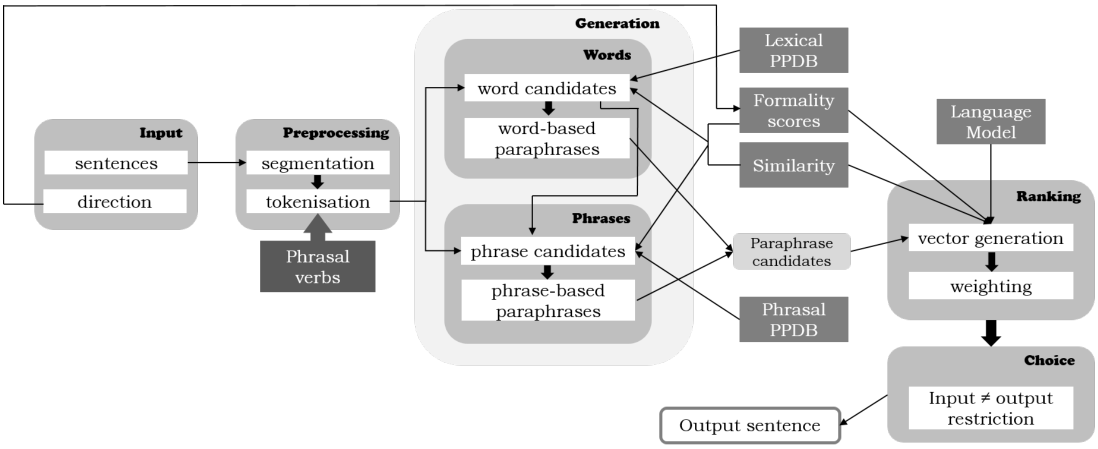

Figure 12: Overall structure of the paraphraser

$$p_{winner} = argmax_p\ \lambda_1 FS(p) + \lambda_2 SS(p) + \lambda_3 LMS(p) \tag{23}$$

Where FS, SS and PP are the formality score, similarity score and the language model score, respectively. The final chosen weights λ were 0.3 for formality, 0.2 for similarity, and 0.5 for fluency.

9.3.4 Choice

Once paraphrases are ranked, in the general case, the paraphrase that is selected to be the system's output is almost always the paraphrase with the highest ranking score. There is one exception: If the winning paraphrase candidate is exactly the same as the input sentence, it is only finally selected if there is a big enough difference with respect to the score of the second candidate in the ranking.

In practical terms, this means that the paraphraser will try to make changes to the input, and will only leave it as is if the next best option is a much worse candidate. After a manual examination of typical scores on a held-out set of sentences, the minimum difference allowed between the input sentence score and the score of the second sentence in the ranking in order for the input sentence to be chosen was set to 0.2.

A general overview of the whole structure of the paraphraser can be seen in Figure 12.

9.4 Evaluation

In this section, we present the criteria on which the evaluation of the paraphraser was based 9.4.1 and explain how the evaluation was designed 9.4.2.

9.4.1 Evaluation criteria

In the ranking section (9.3.3), we described the characteristics we want the output of the paraphraser to have, i.e. what we believe makes a good paraphrase, as well as the elements of the system which are responsible for each of them. A good paraphrase in our context is one that has the desired change in formality, preserves the meaning of the input sentence and sounds natural and correct. Therefore, we decided to evaluate the system on these terms as well. The results of such an evaluation allows us to detect the specific element of the system that is not working properly: formality, similarity (possibly the PPDB itself) or the language model.

(Automatic) paraphrasing has a lot in common with (machine) translation, with the exception of the change in formality, which is the added novelty in this work. Translations should also retain the original meaning and be "fluent". Statistic Machine Translation (SMT) systems are typically intended to satisfy both features with the so-called translation model and a language model, respectively. This similarity between the paraphrasing task and translation allows us to base our evaluation methodology on established machine translation evaluation patterns, at least partially.

It was clear to us that a manual evaluation with help of native speakers would be necessary: these three aspects are not only virtually impossible to judge automatically in an accurate way, but they are also highly subjective. In MT, human evaluation commonly involves assigning translations fluency and adequacy judgments. Fluency refers to the correctness of the translation, whereas adequacy is concerned with the level of meaning preservation with respect to the original text or to a reference translation. These measures were created on the occasion of the NIST Machine Translation Evaluation Workshop (LDC et al., 2005).

We believe fluency to be a complex and unclear feature: it has even been circularly defined as "how fluent the translation is" (Callison-Burch et al., 2007). Having in mind that language models take care of this feature, there is not only one important linguistic aspect to be evaluated. A text can be more or less grammatically correct; and this is not necessarily the same as sounding natural. By natural, we mean whether the chosen combination of words could have been used by a native speaker. For example, a text could be meaningful and flawless, but use strange word combinations: *Let's drink a drink* is grammatically correct, but it is not what a native speaker would typically say, therefore, it is not natural. Consequently, we will distinguish what we will call grammaticality from naturalness. A low score in any of them could be an indicator of the language model not doing its job properly, but a low score on naturalness could also be a sign of the similarity information or the PPDB proposing inadequate word candidates (and the language model not being good enough to account for them).

The evaluation, therefore, takes into account four different factors: grammaticality, naturalness, difference in formality and preserved meaning.

9.4.2 Evaluation design

16 short texts of different topics and genres were selected from the Internet or from films, eight formal texts and eight informal ones, as judged by the author. We made sure topics were as neutral as possible, meaning it would be natural to talk about them in both a formal and an informal way. Some of them were food, animals, health or celebrations, among others. There are descriptions, dialogs, short stories and explanations. The technical requirement for a text in order to be chosen was for all of its sentences to have at most 22 words, in order for the paraphraser to be able to process them. With the exception of the two dialogs, texts had roughly three to four sentences in total. See Appendix C for the original texts, together with their source and paraphrase.

Eight of these texts, four formal texts and four informal ones, were selected to make a control version of them. These eight texts were chosen randomly while avoiding repetitions of topics and/or genres. A native and proficient speaker of English was asked to paraphrase them with the indicated change in formality ("to formal" for informal texts and viceversa). The goal of creating human paraphrases was to have paraphrases that present an upper bound to those of the system. The native speaker was therefore asked to do the same our paraphraser does, with the following instructions:

> *There are 8 short texts in total that have to be paraphrased with a change in formality. The first 4 texts should be made more formal, and the last 4 more informal. The resulting paraphrase should sound natural (native), be grammatical, and have the same meaning as the original text (don't add or subtract meaning).*
>
> *Only make changes that, to your judgment, change the formality of the text to the required direction (even if only slightly). Don't make changes that you think are not changing formality. There is no need to transform the text completely: leave some parts as they are if you can't think of a way of making them more (in)formal.*
>
> *When paraphrasing, try to prioritise word substitution over syntactic*

changes (for example, changing an active voice to a passive voice), but do apply them if necessary.

Try to paraphrase sentence by sentence separately, without mixing sentence content.

With control paraphrases, we have a total of 24 paraphrases (eight human paraphrases and 16 system paraphrases, eight of which can be directly compared to their human version). Control paraphrases are also found in Appendix C.

We posed ourselves a series of questions, the answer to which was obtained during the analysis of the evaluation results:

- To what extent do raters agree with respect to the different variables?
- Are our texts properly selected in terms of formality?
- Does the paraphraser perform significantly differently from a human in any of the aspects that are evaluated?
- Does the paraphraser produce a significant change in formality, and in the desired direction?
- Is there a(n inverse) correlation between the formality of an original text and the formality of its paraphrase?
- Is there a correlation between grammaticality and naturalness? (were native speakers able to tell them apart, or were these aspects influenced by each other?)

Two different questionnaires (Q1 and Q2) were created, each with 16 sections, each section corresponding to a pair of texts (original and paraphrase). In four out of 16 cases, the shown paraphrases were human paraphrases. The human paraphrases were different in each questionnaire. When possible, texts were distributed avoiding repetition of topics. The questionnaires were created with Google's utility Google Forms. Questionnaires were given to a total of four native speakers of English to answer, two for each questionnaire. Sections of the questionnaires were structured in the following way: First, the native speaker is presented with a system or a human paraphrase. They have to rate it answering the following questions:

`Grammaticality`

Question How grammatical do you find the text above?

Answer 1: Full of mistakes - 5: Flawless English

Description By grammatical, we mean how grammatically correct the text is: could this text have been produced by a proficient native? Are there (m)any mistakes? Choose 1 if mistakes make the text incomprehensible.

`Naturalness`

Question How natural-sounding do you find the text above?

Answer 1: Completely unnatural - 5: Native English

Description By natural, we mean how adequate the combination of words that are chosen to express the given meaning is. Could a native speaker have chosen these words? Does their combination sound weird?

`Achieved Formality`

Question How formal or informal is the language used in the text above?

Answer 1: Very informal - 5: Very formal

Description Regardless of the topic of the text, and regardless of the formality you would expect in the situation it evokes, how formal or informal is the language used? Are the words used formal?

Only then, for comparison, a second text, the original one, is presented next to the paraphrase.

`Compared Formality`

Question Which one of the two texts is more formal, in terms of the language used?

Answer The first one, the second one, equally formal/informal

`Original Formality`

Question How formal or informal is the language used in this second text?

Answer 1: Completely different - 5: Exactly the same

`Preserved Meaning`

Question To what extent do the first and the second text mean the same?

Answer 1: Completely different - 5: Exactly the same

Description Do the first and the second text have the same meaning? Do their meanings vary slightly? Are they saying completely different or unrelated things?

9.5 Results and Discussion

Different statistical analyses were carried out to answer our questions about the quality of the paraphraser's output. All tests were run using the R (R Core Team, 2017) programming language. When required, normality was tested with Shapiro-Wilk tests. In cases of paired tests where normality was not met, a Sign test was run instead of a Wilcoxon test due to the abundance of ties in the data. The significance level was set to .05 for all tests, and all p-values were two-tailed. Except for those cases in which comparison of individual responses was necessary, tests were carried out across questionnaire items using the average of all given answers for an item. We will now discuss each of the questions previously introduced.

- To what extent do raters agree with respect to the different variables?

We calculated the inter-rater agreement (IRA) between respondants. Since not all questions had been answered by the same (number of) subjects, three separate IRAs had to be conducted. Weighted Cohen's kappa κ could not be applied to all items, because some of them had four respondants. Others had only two respondants because of the distribution of human and automatic paraphrases over the questionnaires. Therefore, Krippendorff's alpha for ordinal variables was chosen, because it can be used for any number of raters, so as to keep measurements as comparable as possible.

Results are presented in Table 12. Q1 and Q2 stand for questionnaires 1 and 2, respectively. Lowest results are shown in bold. According to Krippendorff (2004), if the risks of drawing incorrect conclusions are not known, 0.667 is considered to be the minimum acceptable value for alpha. While raters seem to agree fairly well when it comes to grammaticality, agreements are weaker for other variables, and most values are below the established limit. After grammaticality, original formality was the only variable where at least two values are higher than this limit. Preserved meaning was the variable with the lowest agreement by difference, featuring even a negative alpha between the two Q2 raters, which possibly indicates a strong tendency to disagreement. However, when looking more closely at the specific answers that gave a negative result (see Table 13) we found that the reason of this apparent disagreement is mainly the lack of variation of rater 1. As Krippendorff (2004) points out, insufficient variation, even if most results are close or even equal, makes data unreliable; but not necessarily because of a lack of agreement between raters. Something similar happened with Q1 raters. Their scores actually coincided in 5 out of 8 cases, but the high amount of 5's in the ratings make them unreliable. In the 4 raters setting, where there were no human paraphrases, variation is higher, and so is alpha, but agreement is also low.

We must, then, take these results into account before drawing conclusions from our statistical analysis. A lower agreement indicates, in general, low reliability, which prevents strong conclusions.

Variable	4 raters	2 raters (Q1)	2 raters (Q2)
Grammaticality	.688	.684	.787
Naturalness	**.293**	.608	.524
Preserved meaning	**.163**	**.054**	**-.244**
Achieved Formality	.526	.614	.716
Original Formality	.44	.905	.764

Table 12: IRA results

Rater 1 (Q2)	5	5	5	5	5	5	5	5
Rater 2 (Q2)	**3**	5	5	**4**	5	**3**	5	**4**

Table 13: Pairs of answers of 2 raters (Q2) on Preserved meaning

- Are our texts properly selected in terms of formality?

The original texts that were manually selected for this evaluation were assumed, and classified, to be either formal or informal. The suitability of texts can be verified or refuted with statistical tests. We ran one-sample tests on originally formal (Sign test) and informal texts (t-test) to determine if original formality (as answered when prompted with the system's paraphrase) was significantly different from a neutral formality level ($\mu_0 = 3$, the middle point in the Likert scale). Results are shown in Table 14, and they seem to indicate that subjects did consider the assumed formal texts to be formal ($p < .05$), but this was not the case for informal texts: the value of original formality is not significantly different from 3 ($p > .05$), which does not allow us to conclude that informal texts were properly selected.

- Does the paraphraser perform significantly differently from a human in any of the aspects that are evaluated?

Texts	Mean ± SD	statistic	p-value	Cohen's d (ES)[15]
Formal texts	4.16 ± 0.44	S = 8	**.008**	2.62 (L)
Informal texts	2.44 ± 0.79	t(7) = -2.02	.083	

Table 14: Results of one-sample tests on original formality ($\mu = 3$)

A series of paired tests were run between human paraphrases (C) and their system counterparts (S) on the different variables. Table 15 summarises results of these tests[16] . The variable formality change is the difference, in absolute value, between achieved formality and original formality. There were significant differences in grammaticality and naturalness between human and system paraphrases, with paraphrases produced by the system being notoriously worse, as the respective differences in means and the large effect sizes reveal. In the other variables, preserved meaning and formality change, human controls and system did not differ significantly ($p > .05$). According to these results, the paraphraser seems to achieve a comparable change in formality to that of the human. It seems to be the case as well for preserved meaning, but we must bear in mind that results on this variable have to be taken with a grain of salt because of their unreliability revealed by the IRA.

Variable	Mean ± SD	statistic	p-value	Cohen's d (ES)
Grammaticality	H: 4.31 ± 0.65 S: 2.50 ± 0.89	t(7) = 4.66	**.002**	1.65 (L)
Naturalness	H: 4.56 ± 0.42 S: 2.94 ± 1.02	t(7) = 4.08	**.005**	1.44 (L)
Preserved meaning	H: 4.75 ± 0.38 S: 4.38 ± 0.58	t(7) = 1.655	.142	
Formality Change	H: 0.56 ± 0.42 S: 0.69 ± 0.53	S = 2	.688	

Table 15: Results of tests comparing Human paraphrases with their equivalent System paraphrases.

- Does the paraphraser produce a significant change in formality, and in the desired direction?

Two one-sample tests were run in order to determine if paraphrases had a significant change in formality, one for the paraphraser (t-test) and one for the human (Sign test). The null hypothesis in both cases was that there was no change in formality ($\mu_0 = 0$).

[16] All tests were paired t-tests except for two sign tests for achieved formality and formality change. H stands for Human and S stands for System.

As results show (Table 16), both the paraphraser and the human paraphrases deviated significantly from the null hypothesis, i.e. they produce a significant change in formality.

Formality Change	Mean ± SD	statistic	p-value	Cohen's d (ES)
Paraphraser	0.63 ± 0.47	t(15) = 5.37	**<.001**	1.34 (L)
Human	0.56 ± 0.42	S = 7	**.016**	1.35 (L)

Table 16: Results of one-sample tests on Formality Change ($\mu = 0$)

Paired tests (a t-test and a Sign test, respectively) were run between achieved and original formality separately in formal and informal texts, in order to see if the change in formality was achieved in the requested direction. Results (Table 17)[17] showed that whereas this holds for texts that were originally formal, the difference is not significant in originally informal texts, whose paraphrases are actually slightly more informal ($M_{OF} = 2.44, M_{AF} = 2.31$). This could be due to the fact that, as we saw earlier in this section, these texts had not been properly selected in terms of formality. In the case of human paraphrases, the direction was correct, but sign tests showed that differences were not significant, probably partly because of the small sample size (only 4 paraphrases in each direction).

Setting	Originally	Mean ± SD	statistic	p-value	Cohen's d (ES)
Paraphraser	Formal	OF: 4.16 ± 0.44 AF: 3.60 ± 0.76	t(7) = 2.909	**.023**	1.02 (L)
	Informal	OF: 2.44 ± 0.79 AF: 2.31 ± 0.83	S = 7	1	
Human	Formal	OF: 4.5 ± 0.71 AF: 3.75 ± 0.65	S = 4	.125	
	Informal	OF: 1.5 ± 0.41 AF: 1.88 ± 0.63	S = 0	.25	

Table 17: Results of direction of the paraphraser's change in formality

- Is there a(n inverse) correlation between the formality of an original text and the formality of its paraphrase?

From the way the paraphraser was designed, we would expect that, in an ideal setting, Original and Achieved Formality would be inversely correlated: an originally informal text would end up being formal, and viceversa. However, plots (Figure 13)[18] seem to indicate that the tendency is in fact the reverse. Correlation analyses (Table 18) confirmed that there was, in both cases, a strong, positive correlation.

[17] OF stands for original formality and AF for achieved formality.

[18] Size of dots indicates frequency.

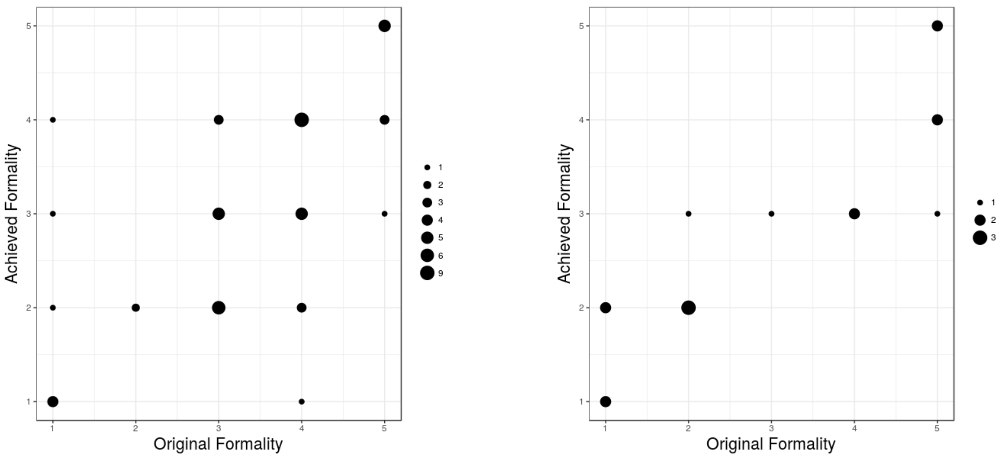

Figure 13: Original and achieved formality of the system (left) and the human (right).

At first, by looking at the plots, one might think that this indicates that there is no change in formality. However, it must be noted that our expectations were actually based on a categorical view of formality, where scores 1 and 2 correspond to informal and 4 and 5 correspond to formal. In fact, one does see a tendency in the correct direction —just not a big enough tendency to change to a different "category" of formality. What this means is that formal texts, once paraphrased, were indeed more informal, but the overall impression the raters had of the paraphrase was that it was still a formal text. The reverse would apply for informal texts. The same happened with human paraphrases.

Set	Variable 1	Variable 2	Type	p-value	ρ
Paraphraser	Original F.	Achieved F.	Spearman	$<$ **.001**	.647
Human	Original F.	Achieved F.	Spearman	$<$ **.001**	.880
Paraphraser	Grammaticality	Naturalness	Spearman	$<$ .001	.786

Table 18: Correlation results

- Is there a correlation between grammaticality and naturalness?

As explained in Section 9.4, grammaticality and naturalness are two variables that result from the disambiguation of the typically used concept of fluency. We argued that they are different, but did raters understand this difference? If they are not correlated, we could conclude that people did distinguish between both. If instead they are correlated, it could either mean that they did not make a difference between them, or that most paraphrases deserved a similar rating on these two aspects.

Results are shown in the last row of Table 18: There is a strong and significant positive correlation between grammaticality and naturalness. The idea that these two concepts were badly defined and explained, and that they were not distinguished by raters, cannot be rejected.

A visual comparison of grammaticality and, especially, of naturalness between formal and informal paraphrases (as output by the paraphraser) seemed to indicate that formal paraphrases were, in general, worse than informal ones (see Figure 14). No valid statistical test, to our knowledge, could be meaningfully applied to check for the significance of this difference —data are not independent and not paired—, but this could possibly be indicating a tendency for formal sentences to be perceived as more unnatural. We hypothesise that native speakers could have understood the term naturalness as a sort of "familiarity", and that formal sentences, because of their more restricted situations of usage, sound, in this sense, less "familiar".

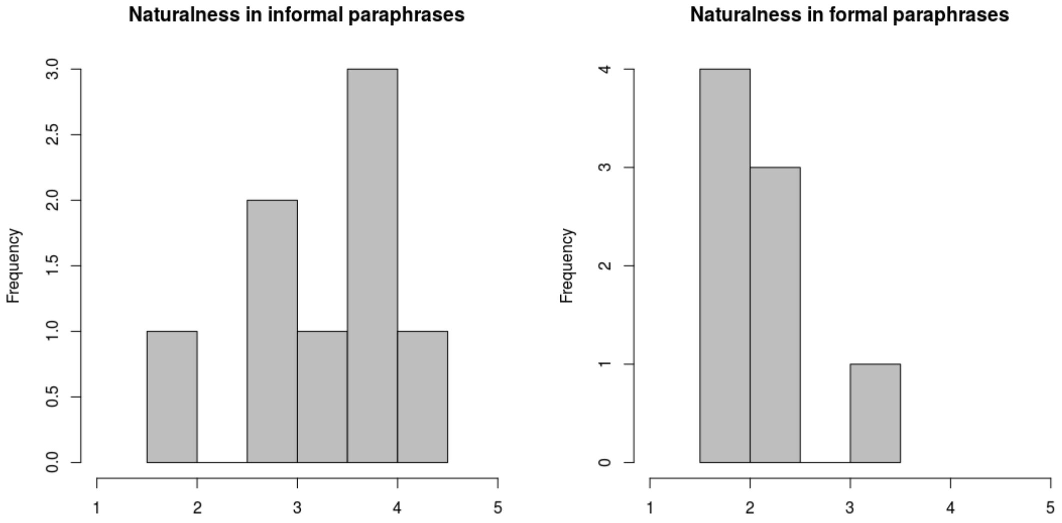

Figure 14: Naturalness in informal and formal automatic paraphrases

To sum up, the main findings were:

Subjects did not agree very well in all variables. Grammaticality was the only variable with the minimum accepted value of Krippendorff's alpha. Preserved meaning showed a very low agreement, mainly due to the unreliability of data derived from the lack of variation. These results do not allow us to draw many strong conclusions from the paraphraser's evaluation results.

Another finding was that subjects did not consider assumed informal texts to be significantly different from a middle, neutral formality level, which affects the balance of the evaluation set as we had designed it, as well as the perceived changes in formality.

In the face of these facts, if we were to repeat such an evaluation, we would ideally have not only more subjects and more texts, but we would also carry out a preliminary evaluation of the formality of the chosen texts.

One of the most important conclusions to draw from the evaluation is that the paraphraser is still very far from being as good as a human in the same task. In the next section we discuss some ways in which we think the performance of the paraphraser could have been improved. In general, automatic and human paraphrases are only comparable when it comes to the change in formality, which was significant for both, but the desired direction in formality was only significantly achieved by the paraphraser when paraphrasing to informal, possibly due to the inadequate selection of informal texts. We believe that, in the case of human paraphrases separated into formal and informal, sample size was too small to obtain significant results.

Original and achieved formality were in a strong positive correlation both for the automatic paraphrases and for human paraphrases, which led us to think that the resulting paraphrases, even if they could feature a formality change in the desired direction, that change was small and not enough to make the text change drastically in formality.

Finally, it was observed that grammaticality and naturalness were also in a strong positive correlation, possibly indicating that speakers did not distinguish between both concepts.

Overall, it seems that formality scores, which are used by the paraphraser, are doing their job well enough. The language model, which is the main responsible for grammaticality and naturalness, is not producing the effect that we had expected. Of course, low results on grammaticality can also be due to the fact that the paraphraser does not use any explicit syntactic information. In addition, it is also possible that the language model is just not assigned a big enough weight in the ranking module. When it comes to similarity scores, as we have seen, we cannot make a definite conclusion: Preserved meaning was quite high in the automatic paraphrases, but low IRA points out the unreliability of judgments on this variable.

9.6 Limitations and Future Work

In this section, we discuss the drawbacks and disadvantages of the paraphraser that we found, and suggest solutions to overcome them in future work.

Efficiency limitations and the Moses-based paraphraser. One of the most important drawbacks of this paraphraser is the maximum of 22 words per sentence. More time and expertise would have been necessary to reduce the search space while at the same time keeping the best possible paraphrase candidate, and even to increase the number of substitute candidates per word and phrase, which is currently limited to 2 —or 3 if we count the very same word or phrase—. The way of dealing with the sentence length limitation that we chose in the end was to limit the length of the test sentences to that maximum. At an earlier stage, however, we considered another solution: building a Moses-based paraphraser instead. Moses (Koehn et al., 2007) is an open source toolkit for Statistical Machine Translation (SMT). We discussed similarities between translation and paraphrasing in Sections 3.6 and 9.4, partly because they inspired the paraphraser's evaluation methodology. These similarities are one of the justifications for using this toolkit, the main reason being its efficient and fast implementation.

Minimally, we would need to provide the MT system with a language model and a translation model. The translation model is a table with words or phrases in one language and their equivalents in the other, together with the corresponding translation probability. Two additional elements than can be added are the distortion model, which ensures that there is not an excessive reordering when translating, and the word penalty, which is responsible for a reasonable sentence length. The final translation is then decided by reducing the translation cost, which is a weighted combination of these four elements. Optimal weights are typically set by using a development set.

We have started to design and implement a system that takes advantage of Moses' speed for paraphrasing purposes. A brief description of the design will follow.

As a language model, we would use the language model we built from the Mixed Corpus. Its function is the same as in the paraphraser: ensuring correct and natural output. The translation model would be the key component: Instead of linking expressions in two languages, it would consist of paraphrases as extracted from the PPDB. In order to account for formality direction, two different paraphrase tables would have been created, one table for paraphrasing to formal, and the other one to informal. The direction of the paraphrasing is given by the difference in formality as calculated in Section 9.2.1. Every word or phrase would be included as a paraphrase of itself. In order to obtain the equivalent to translation probabilities, it is important that real probabilities are created, for which we need to take into consideration the following:

In SMT, translation probabilities are learnt from parallel corpora. Due to the noisy channel model, probabilities are the inverse of what we ultimately want: they encode the probability of the original word given a translation. As a simple example, imagine we have a sentence in Spanish (the original language, "s"): *Tengo hambre.* And we want to translate it into English ("e"): *I'm hungry.* In the translation table, we will have:

Tengo hambre	*I'm hungry*	p(*Tengo hambre* | *I'm hungry*)

In the paraphrasing case, parallel corpora are not available. We have, however, other indicators of how good a paraphrase can be: these are similarity and the change in formality. A weighted combination of them could be used, which would be turned into a probability by distributing its value among the corresponding possible origins for a specific paraphrase ensuring it ranges between 0 and 1.

The distortion model would be one which strongly penalises even the minimum distortion: the goal of our paraphraser is to apply lexical substitution only, with no reordering at all.

Word penalty could in principle be left as it is set by default in Moses, or even dismissed.

One aspect that could be less trivial to deal with, however, is how to avoid the Moses-based paraphraser outputting the very same input sentence.

We leave this on how to improve this paraphraser for future work.

Manual intervention. Another major drawback of the paraphraser is the manual optimisation of the weights assigned to similarity, formality and the language model by observing the performance on a selected group of short sentences containing words or expressions that have near-synonyms that differ in formality. In addition to being a subjective method, the author is not a native English speaker, for which judgments on grammaticality and naturalness may not have been ideal. Had there been more time for optimisation, a native speaker would have been asked to create a small development dataset by paraphrasing several sentences themselves, which would be used as ground-truth for a similar manual, or even automatic, optimisation of the weights.

Absence of syntactic information. Our initial intention was to build a paraphraser that worked uniquely by lexical substitution, where a language model would take care of the output being grammatically correct. However, as results show, grammar, as well as naturalness, are still far from being acceptable. In the future, it would be a good idea to analyse the concrete linguistic errors of the paraphraser and include a syntactic module to it.

PPDB and distributional similarities. PPDB is not perfect but was, to our knowledge, the best option available when it comes to a list of paraphrases. We decided not to use its scores after a manual examination and use distributional similarity instead, which does not guarantee equivalence of meaning either. Entailment information could have also been very useful, and we do not rule out using it in the future if it is improved.

Two language models. Another possibility that remains to be explored is that of using two separate direction-specific language models, one created from informal corpora only (to paraphrase to informal) and another one based on formal corpora. This could potentially increase the naturalness of the system, because the language model would reward especially those paraphrases that are correct and in the required formality level.

10 Conclusions

In this thesis, we have extracted lexical information related to formality, attitude and collocational preferences in an almost unsupervised way with the goal of partially characterising near-synonyms and assisting in the task of lexical choice. Information on formality was then used as one of the main ingredients to build a paraphraser that produces paraphrases with a change in formality based on lexical substitution.

First, formality information was extracted in the form of lexical scores with existing methods that make use of the distributional similarity between a word and words of known formality (or "seed words"). An intrinsic evaluation showed that high-quality information on formality could be obtained from a much smaller corpus than used in previous work, and that the addition of phrasal verbs as units in the vocabulary proved beneficial, possibly thanks to the verb disambiguation they give rise to.

Next, attitude information was extracted in terms of polarity. Two main methods were compared, a document-based method and a sentence-based method. Results were very poor in terms of distinguishing positivity from negativity, but decent when considering the distinction between polar and non-polar (neutral) words. Surprisingly, the document-based method performed much better than the sentence-based method on this distinction. After an exploration of seed words as clusters, we concluded that bad results were partly due to a bad set of seeds, but also possibly to an insufficient number of dimensions in our LSA approach, among others. More dimensions could have led to a vector representation of the seed words that was more adequate to our purposes, but this was out of our computational possibilities.

We explored the interaction between formality and attitude using the extracted information, but due to the low reliability of the extracted attitudinal scores and conflicting results in a similar study with SentiWordNet, we could not draw meaningful conclusions from it.

Collocational preferences were obtained by means of three different measures of statistical association. This information was only tested extrinsically in the Fill-in-the-Blanks (FITB) task, together with formality and attitude.

The FITB task was modified to overcome its drawbacks and to fit better to our needs by changing the corpus and near-synonym sets, losing comparability but gaining in exchange a much more informative insight. The experiments on the FITB task showed that a system using the obtained information for lexical choice was not able to outperform a language model or the majority class baseline. However, the baseline was outperformed when frequency information or the language model were added to a voting system combining the three kinds of information. A separate setting showed that, out of the three, collocations proved to be the most useful kind of information.

Finally, an automatic paraphraser with a change in formality based on lexical substitution was built making use of the obtained lexical formality information and other freely available resources. A human evaluation was carried out with four native speakers of English on several paraphraser outputs and human paraphrases written by another native speaker. The low inter-rater agreement does not allow us to draw strong conclusions, but it was clear that the paraphraser is far from human performance in terms of fluency. The paraphraser did produce a change in formality, but this change showed to have the desired directionality on formal texts only, possibly because the originally assumed informal texts had not been properly selected in terms of formality, as perceived by human raters. We pointed out a number of possible future improvements to the paraphraser, including higher efficiency and the use of syntactic information or separate language models to increase grammaticality and naturalness, among other suggestions.

Answering the main research questions we posed in the introduction, we conclude that:

1. Information on the different dimensions of difference between near-synonyms can be useful, especially if used in combination with frequency or a language model. Even though the language model has better results, an application that requires distinguishing between dimensions could probably benefit from our system.

2. It is not clear that there is a correlation between lexical formality and attitude. Obtaining better scores for attitude would allow us to look at their relation in better conditions.

3. The paraphraser does produce a perceived change in formality, but mechanisms other than word and phrase substitution might be needed in order to produce output that is more correct.

A major pitfall of this work has been the insufficient, even if already big, computer power and memory to increase the number of LSA dimensions, as well as the sentence length limit of the paraphraser —where, admittedly, lack of expertise also played a role—.

Overall, we could extract two different useful kinds of information that can potentially distinguish between near-synonyms: formality and collocations. Certainly, near-synonyms can differ in many other dimensions that were not taken into account in this work. It remains to be tested whether extraction of other kinds of information (dialectal or denotational variation, for example) can be helpful in the task of lexical choice. Formality proved to be useful for a simple automatic paraphraser. That still needs to be improved in future work to gain a performance that is comparable to human performance.

Appendices

A Attitude seeds and non-polar words

Positive	excellent, stunning, awesome, prodigious, gorgeous, fabulous, hilarious, perfect, delighted, handsome, amazing, delicious, astonishing, heroic, happy, lovely, magnificent, smartest, cautious, brilliant, marvelous, fascination, eloquent, satisfactory, fantastic, gifted, charming, genius, exquisite, sensational
Negative	horrible, pity, barbarous, stupidly, awful, humiliating, hideous, scandal, terrifying, deceive, stink, weird, dishonor, obsessively, stingy, incompetent, bastard, bore, fiasco, insolent, arrogant, impulsive, silly, hysteria, nasty, coward, stubborn, tragic, ridiculous, importunate
Non-polar	external, here, formed, basis, finished, operation, empty, full, first, appointment, initial, green, squared, circular, always, anywhere, apply, add, union, corresponding, situation, outside, inside, afternoon, transfer, visiting, built, sometimes, watch, concept, age, task, form, half, assign, continue, sequence, multiple, forty, rice, person, absent, yellow, red, blue, examine, image, extension, mother, published, ago, daily, category, together, nine, social, available, fifteen, eight, either

B WordNet synsets used on the FITB task

Due to the big amount of near-synsets, which make up to around 2300 words, we show here only part of them. The whole list is available for consultation on the Internet[19].

Nouns	Verbs	Adjectives
arrangement, organization, organisation, system	clean, pick	electric, electrical
region, realm	figure, enter	attached, committed
repository, secretary	vacation, holiday	exclusive, sole
touch, spot	read, register, show, record	secret, private
effect, force	guide, run, draw, pass	capable, open, subject
pledge, toast	bid, call	decreased, reduced
hearing, audience	sit_down, sit	dear, good, near
profit, gain	box, package	careful, deliberate, measured
education, training, breeding	justify, warrant	substantial, substantive
concession, grant	interview, question	fiscal, financial
topic, subject, issue, matter	research, search, explore	crude, rough
area, region	blame, fault	false, mistaken
days, years	wish, bid	distinct, decided
assimilation, absorption	feed, eat	violent, wild
appearance, show	wheel, roll	former, late, previous
publication, publishing	format, arrange	nasty, tight
mention, reference	reserve, hold, book	plus, positive
distance, length	represent, interpret	dangerous, unsafe
population, universe	hit, strike, come_to	quick, warm
pitch, delivery	comment, notice, remark, point_out	big, heavy
treatment, intervention	piece, patch	like, same
finish, destination, goal	sweep, sail	insecure, unsafe
determination, purpose	react, oppose	loose, open
mount, setting	gap, breach	adequate, equal
program, programme	spy, sight	gentle, soft
offer, offering	rule, find	standard, received
part, piece	repeat, take_over	hurt, wounded
press, pressure, pressing	concern, interest, occupy, worry	hard, strong

[19] https://drive.google.com/file/d/0BynZF6WncRqRTlplNWJPa0Y3M1k/view?usp=sharing

C Paraphrased texts

Source https://en.wikipedia.org/wiki/Christmas

Questionnaire Automatic in Q1, Human in Q2

Formality Formal

Original Christmas cards are purchased in considerable quantities, and feature artwork, commercially designed and relevant to the season. The content of the design might relate directly to the Christmas narrative, with depictions of the Nativity of Jesus, or Christian symbols such as the Star of Bethlehem.

Automatic Christmas cards are bought in enormous quantities, and feature artwork, commercially designed and relevant to the season. The content of the design might directly related to the Christmas narrative, with portrayals of the Nativity of Jesus, or Christian symbols such as the Star of Bethlehem.

Human Christmas cards are bought in large numbers, and feature artwork, commercially designed and relevant to the season. The design's content might relate directly to the Christmas narrative, with depictions of the Nativity of Jesus, or Christian symbols like the Star of Bethlehem.

Source Film "the Young Victoria"

Questionnaire Q1 and Q2

Formality Formal

Original
"Lord Melbourne. The Liberal leader who'll probably be in power
when the Princess succeeds. He may be troublesome."
"Why?"
"He puts the interests of England above those of Europe."
"Which is bad?"
"Which is not useful to us. He wouldn't spill one drop of English
blood to save a foreign throne."
"Why would he save a foreign throne if it wasn't in England's
interest?"
"That is just the kind of thinking your Uncle Leopold is afraid of.
Which is why he's content to find his niece is the future Queen
of England."

Automatic	“Lord Melbourne. The liberal leader who’ll maybe be in power when the princess succeeds. He may be bothersome.” “Why?” “He pulls the interests of England above those of Europe.” “Which is horrible?” “Which is anymore useful to us. He wouldn’t spill one drop of English blood to save a foreign throne.” “Why would he save a foreign throne if it ain’t in England’s interest?” “That is maybe the really of dreaming your dad Leopold is scared of. Which is why he’s content to find his sister is the future queen of England.”
Source	http://www.who.int/elena/titles/fruit_vegetables_ncds/en/
Questionnaire	Automatic in Q2, Human in Q1
Formality	Formal
Original	Fruits and vegetables are important components of a healthy diet. Reduced fruit and vegetable consumption is linked to poor health and increased risk of noncommunicable diseases (NCDs). An estimated 5.2 million deaths worldwide were attributable to inadequate fruit and vegetable consumption in 2013.
Automatic	Fruits and veggies are important parts of a healthy diet. Lowered pear and potato consume is linked to poor health and higher risk of noncommunicable disease (NCDs). The roughly 5.2 million dead globe were due to inadequate fruit and vegetable consume in 2013.
Human	Fruits and vegetables are important parts of a healthy diet. Eating fewer fruits and vegetables is linked to bad health and a higher risk of noncommunicable diseases (NCDs). Around 5.2 million deaths worldwide were due to inadequate fruit and vegetable consumption in 2013.
Source	http://www.dailymail.co.uk/health/article-154824/Why-eat-soya.html
Questionnaire	Q1 and Q2
Formality	Informal
Original	I always enjoy visiting my sister and her family, partly because they’re vegetarians and Annie is a whiz at transforming soya mince into delicious meals for her boys - lasagne and bolognese, for instance.

Automatic	I always enjoy visiting my mother and her family, partly due they're vegetarians and Annie is a whiz at transforming soya mince into delectable meals for her boys - lasagne and bolognese, thus for instance.
Source	https://mymerrychristmas.com/forum/threads/christmas-plans-preparations-2015.63579/
Questionnaire	Q1 and Q2
Formality	Informal
Original	Once September arrives I start daydreaming about Christmas a lot more. I start thinking about gift ideas, menus, decorating and everything else. With the days moving fast, it's impossible not to start putting some thought into the approaching holidays!
Automatic	Once September arrives I start daydreaming about Christmas a lot of people more. I begin thinking about gift ideas, menus, decoration and all else. With the days moving fast, it's inconceivable necessarily to begin placing some believed into the approaching holidays!
Source	https://thecatsite.com/threads/how-do-i-make-my-cat-open-up.344916/
Questionnaire	Automatic in Q1, Human in Q2
Formality	Informal
Original	I have 2 cats and they're all really great! But, my 2nd cat is often under the bed and very wary about new people. Toys also sometimes scare her a lot too... Any ways to make my cat calm and easier to approach?
Automatic	I have 2 felines and they are all obviously tremendous! Although, my 2nd pup is frequently under the bed and extremely cautious about new people. Playthings additionally occasionally scare her a lot too... Any ways to make my pup calm and simpler to approach?
Human	I own 2 cats and they're all really wonderful! But, my 2nd cat is often under the bed and very wary about new people. Toys also sometimes frighten her a lot too... Any suggestions to make my cat calm and easier to approach?
Source	https://en.wikipedia.org/wiki/Cat
Questionnaire	Q1 and Q2
Formality	Formal

Original	Cats, like dogs, are digitigrades. They walk directly on their toes, with the bones of their feet making up the lower part of the visible leg. Cats are capable of walking very precisely because, like all felines, they directly register; that is, they place each hind paw (almost) directly in the print of the corresponding fore paw, minimising noise and visible tracks.
Automatic	Kittens, like puppies, are digitigrades. They walk directly on their toes, with the bones of their feet making up the lower part of the visible knee. Kittens are able walking pretty exactly because, like everyone cats, they directly register; that is, they place each hind paw (almost) directly in the print of the corresponding fore paw, lessening noise and visible tracks.
Source	Film “The Big Lebowski”
Questionnaire	Automatic in Q2, Human in Q1
Formality	Informal
Original	“I gotta tell you, Tone, man, earlier in the day, I was really feeling shitty, man. Really down in the dumps. Lost a little money... Hey, you know what? forget about it, huh? Forget about it. Yeah, fuck it, man. I can’t be worried about that shit.”
Automatic	“I got da tell you, Tone, man, previously in the day, I was indeed feeling lousy, man. Obviously down in the dumps. Loses a little money... Maybe, you known what? Forgotten all about it, huh? Forgotten about it. Suppose, fuck it, man. I isn’t possible concerned about that shit.”
Human	“I must tell you, Tone, man, earlier in the day, I was really feeling awful, man. Really down in the dumps. Lost some money... well, you know what? forget about it, huh? Forget about it. Yeah, fuck it, man. I can’t be worried about that shit.”
Source	https://www.google.com/policies/terms/
Questionnaire	Automatic in Q2, Human in Q1
Formality	Formal
Original	Using our Services does not give you ownership of any intellectual property rights in our Services or the content you access. You may not use content from our Services unless you obtain permission from its owner or are otherwise permitted by law. These terms do not grant you the right to use any branding or logos used in our Services. Don’t remove, obscure, or alter any legal notices displayed in or along with our Services.

Automatic Using our service does not give you ownership of any intellectual property rights in our service or the content you access. you may not use content from our service if you obtain permission from its owner or are otherwise forbidden by law. These terms do not grant you the right to use any branding or logos used in our service. Don't remove, obscure, or change any legal notices displayed in or together with the our service.

Human Using our Services doesn't mean that you own any intellectual property rights in our Services or the things you access. You may not use content from our Services unless you get permission from its owner or are otherwise allowed by law. These terms do not give you the right to use any branding or logos used in our Services. Don't remove, obscure, or change any legal notices shown in or along with our Services.

Source https://www.tripadvisor.com/ShowTopic-g293921-i8432-k10474754-Trip_report_My_trip_to_Vietnam_3_weeks-Vietnam.html

Questionnaire Q1 and Q2

Formality Informal

Original Saying that we loved Hoi An is an understatement. Everything was amazing, the hotel, the vibe of the town, the My Son tour, the beach... just wow. The food was also our favorite in Vietnam.

Automatic Asserting that we loved Hoi An constitutes an understatement. Whatever was incredible, the hotel, the ambience of the village, the My Son tours, the beach... just wow. The food was additionally our favorite in Vietnam.

Source https://www.biography.com/people/salvador-dal-40389

Questionnaire Automatic in Q1, Human in Q2

Formality Formal

Original Consequently, Dalí was subjected to furious acts of cruelty by more dominant students or his father. The elder Salvador wouldn't tolerate his son's outbursts or eccentricities, and punished him severely. Their relationship deteriorated when Salvador was still young, exacerbated by competition between he and his father for Felipa's affection.

Automatic Therefore, Dalí was subjected to livid acts of cruelty by more dominant classmates or his uncle. The elder Salvador wouldn't tolerate his son's outbursts or eccentricities, and punished him badly. Their relationship deteriorated when Salvador was still young, exacerbated by competion between he and his uncle for Felipa's affection.

Human	As a result, Dalí faced furious acts of cruelty by more dominant students or his father. The elder Salvador wouldn't stand his son's outbursts or eccentricities, and punished him severely. Their relationship deteriorated when Salvador was still young, worsened by competition between him and his father for Felipa's affection.
Source	http://www.harrywinston.com/en/collection/962/winston-icons
Questionnaire	Q1 and Q2
Formality	Formal
Automatic	Creating stunning jewels from precious diamonds is an incredible art unto itself. Master jewelers must have and not just a sound knowledge of the finest material, but the creative curiosity and technical virtuosity to awaken an unimagined brilliance at each gem.
Original	Creating spectacular jewels from precious gemstones is an extraordinary art unto itself. Master jewelers must possess not only a sound knowledge of the finest materials, but the creative curiosity and technical virtuosity to awaken an unimagined brilliance in each individual gem.
Source	https://www.reddit.com/r/GalaxyS7/comments/5liz6l/this_is_the_best_phone_ive_ever_had/
Questionnaire	Human in Q1, Automatic in Q2
Formality	Informal
Original	I've had it for a week now and couldn't be happier, my note 4 was great except for the awful battery life. The difference between the note 4 and the s7 is astounding, the speed and the battery blows the note 4 out of the water!
Automatic	I've had it for per week now and wasn't able be happier, my note 4 was tremendous except in cases of the horrendous battery life. The disparity between the note 4 and the s7 is remarkable, the speed and the battery blows the note 4 out of the groundwater!
Human	I've had it for a week now and couldn't be happier, my note 4 was great except for the awful battery life. The difference between the note 4 and the s7 is astounding, the speed and the battery are far better than on the note 4!

Source	http://www.sciencedirect.com/science/article/pii/S0271531717302440
Questionnaire	Q1 and Q2
Formality	Formal
Original	Hypertension is one of the most important preventable causes of premature death. Studies have been conducted assessing the impact of pomegranate on blood pressure, with varying results. The aim of this review was to critically appraise and evaluate the effect of pomegranate on blood pressure in adults, using evidence from randomized clinical trials (RCTs).
Automatic	Hypertension is one of the most important preventable causes of untimely death. Studies have been done analyzing the effect of pomegranate on blood pressure, with different results. The purpose of this review was to critically appraise and analyze the effect of pomegranate on blood pressure in teens, used proof from randomized clinical trials (RCTs).
Source	http://www.auntpeaches.com/2017/06/finding-plaid.html
Questionnaire	Q1 and Q2
Formality	Informal
Original	Sometimes people like to ask me why I like painting. It is my least favorite thing ever. This will sound weird, but that question is sort of like asking a mother why they like their child. Which, depending on the child, might not be an unreasonable question.
Automatic	Occasionally people like to ask me why I like paintings. The latter is my least favorite thing ever. This will sound bizarre, although that question arises sort of like requesting a mother why they like their child. Which, depends upon the child, might necessarily becomes an unjustified question.
Source	https://www.tripadvisor.co.uk/ShowUserReviews-g187147-d3192219-r484717412-La_Petite_Rose_des_Sables-Paris_Ile_de_France.html
Questionnaire	Automatic in Q1, Human in Q2
Formality	Informal
Original	The place was absolutely amazing! One of the best restaurant nights ever. The owner couple was so lovely and the atmosphere was so welcoming. Even though they barely spoke any English (and we don't speak French) everything was so easy and we understood each other very well. The food was also good, felt like I was at grandma's.

Automatic	The place was absolutely incredible! One among the best restaurant evenings ever. The owner few was thus delightful and the atmosphere was thus welcomed. Even although they scarcely spoke whatsoever English (and we don’t speak French) whatever was thus easy and we understood each other extremely well. The foodstuffs was additionally good, felt like I was at grandmother ’s.
Human	The place was absolutely splendid! One of the best restaurant nights yet. The owner couple was really lovely and the atmosphere was very welcoming. Even though they barely spoke any English (and we don’t speak French) everything was really easy and we understood each other very well. The supper was also delicious, felt like I was at my grandmother’s.

References

American Psychological Association. (2010). *Publication manual of the american psychological association*. Washington, DC: American Psychological Association.

Andreevskaia, A., & Bergler, S. (2006). Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *Eacl* (Vol. 6, pp. 209–216).

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, *38*, 135–187.

Apresjan, J. D. (1973). Synonymy and synonyms. In *Trends in soviet theoretical linguistics* (pp. 173–199). Springer.

Argamon, S., & Koppel, M. (2010). The rest of the story: Finding meaning in stylistic variation. In *The structure of style* (pp. 79–112). Springer.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, pp. 2200–2204).

Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the ngram statistics package. In *International conference on intelligent text processing and computational linguistics* (pp. 370–381).

Bannard, C., & Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 597–604).

Barrass, R. (2005). *Scientists must write: A guide to better writing for scientists, engineers and students*. Taylor & Francis. Retrieved from `https://books.google.com.mt/books?id=PgqCAgAAQBAJ`

Batchelor, R., & Offord, M. (1993). *Using french synonyms*. Cambridge University Press. Retrieved from `https://books.google.com.mt/books?id=yzN4jwEACAAJ`

Bbc news style guide. (n.d.). `http://www.bbc.co.uk/academy/journalism/news-style-guide`. (Accessed: 2017-03-30)

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Blaheta, D., & Johnson, M. (2001). Unsupervised learning of multi-word verbs. In *Proc. of the acl/eacl 2001 workshop on the computational extraction, analysis and exploitation of collocations* (pp. 54–60).

Bloomfield, L. (1933). Language. *New York, Holt.*

Breidt, E. (1996). Extraction of vn-collocations from text corpora: A feasibility study for german. *arXiv preprint cmp-lg/9603006*.

The british national corpus, version 3 (bnc xml edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from `http://www.natcorp.ox.ac.uk/`

Brooke, J. (2014). *Computational approaches to style and the lexicon* (Unpublished doctoral dissertation). University of Toronto.

Brooke, J., & Hirst, G. (2013). A multi-dimensional bayesian approach to lexical style. In *Hlt-naacl* (pp. 673–679).

Burton, K., Java, A., & Soboroff, I. (2009). The icwsm 2009 spinn3r dataset. In *Third annual conference on weblogs and social media (icwsm 2009).*

Callison-Burch, C., Cohn, T., & Lapata, M. (2008). Parametric: An automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd international conference on computational linguistics-volume 1* (pp. 97–104).

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (meta-) evaluation of machine translation. In *Proceedings of the second workshop on statistical machine translation* (pp. 136–158).

Chafe, W. L. (1971). Directionality and paraphrase. *Language*, 1–26.

Chierchia, G., & McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics.* MIT press.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. *Lexical acquisition: exploiting on-line resources to build a lexicon*, *115*, 164.

Church, K., Gale, W., Hanks, P., Hindle, D., & Moon, R. (1994). *Lexical substitutability.* In Atkins and Zampolli (eds.) Computational Approaches to the Lexicon. Oxford: Oxford University Press, pp. 153-177.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22–29.

Cruse, D. A. (1986). *Lexical semantics.* Cambridge University Press.

Daille, B. (1994). *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques* (Unpublished doctoral dissertation). Université Paris 7.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*(2), 224–227.

Davies, M., & Fuchs, R. (2013). *Corpus of global web-based english:(glowbe): 20 coun-*

tries, 1.9 billion words. BYE, Brigham Young University.

Dias, G., Guilloré, S., & Lopes, J. G. P. (1999). Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In *Proceedings of conférence traitement automatique des langues naturelles (taln).*

DiMarco, C., Hirst, G., & Stede, M. (1993). The semantic and stylistic differentiation of synonyms and near-synonyms. In *Aaai spring symposium on building lexicons for machine translation* (pp. 114–121).

Duboue, P. A., & Chu-Carroll, J. (2006). Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *Proceedings of the human language technology conference of the naacl, companion volume: Short papers* (pp. 33–36).

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, *19*(1), 61–74.

Edmonds, P. (1997). Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics* (pp. 507–509).

Edmonds, P. (1999). *Semantic representations of near-synonyms for automatic lexical choice* (Unpublished doctoral dissertation). University of Toronto.

Edmonds, P., & Hirst, G. (2002). Near-synonymy and lexical choice. *Computational linguistics*, *28*(2), 105–144.

Elhadad, M. (1993). *Using argumentation to control lexical choice: a functional unification implementation* (Unpublished doctoral dissertation). Columbia University.

Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations* (Unpublished doctoral dissertation).

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th annual meeting on association for computational linguistics* (pp. 188–195).

Farrell, R. (1977). *Dictionary of german synonyms.* Cambridge University Press. Retrieved from `https://books.google.com.mt/books?id=8h6SDAEACAAJ`

Fellbaum, C. (1998). *Wordnet.* Wiley Online Library.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Fowler, H. W., et al. (1926). *Dictionary of modern english usage.* The Clarendon Press.

Francis, W. N., & Kučera, H. (1979). *Manual of information to accompany a standard corpus of present-day edited american english, for use with digital computers.* Brown University, Department of Lingustics.

Ganitkevitch, J., & Callison-Burch, C. (2014, May). The multilingual paraphrase database. In *The 9th edition of the language resources and evaluation conference.* Reykjavik, Iceland: European Language Resources Association. Retrieved from `http://cis.upenn.edu/~ccb/publications/ppdb-multilingual.pdf`

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013). Ppdb: The paraphrase database. In *Hlt-naacl* (pp. 758–764).

Gardiner, M., & Dras, M. (2012). Valence shifting: Is it a valid task. In *Australasian language technology association workshop 2012* (p. 42).

Gardiner, M. E., et al. (2013). *Natural language processing methods for attitudinal near-synonymy* (Unpublished doctoral dissertation). Sydney, Australia: Macquarie University.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *1992 ieee international conference on acoustics, speech, and signal processing, 1992, icassp-92* (Vol. 1, pp. 517–520).

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Guerini, M., Strapparava, C., & Stock, O. (2008). Valentino: A tool for valence shifting of natural language texts. In *Lrec.*

Harris, R. (1973). *Synonymy and linguistic analysis.*

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on european chapter of the association for computational linguistics* (pp. 174–181).

Hayakawa, S. I., & Ehrlich, E. (1994). *Choose the right word.* Harper Perennial.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation* (pp. 187–197).

Heid, U. (2004). On the presentation of collocations in monolingual dictionaries. In *Proceedings of the 11th euralex international congress* (Vol. 2, pp. 729–738).

Hemchua, S., Schmitt, N., et al. (2006). An analysis of lexical errors in the english compositions of thai learners. *PROSPECT-ADELAIDE-*, *21*(3), 3.

Heylighen, F., & Dewaele, J.-M. (1999). Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center Leo Apostel, Vrije*

Universiteit Brüssel.

Hirst, G. (2003). Paraphrasing paraphrased. In *Keynote address for the second international workshop on paraphrasing: Paraphrase acquisition and applications.*

Inkpen, D. (2007a). Near-synonym choice in an intelligent thesaurus. In *Hlt-naacl* (pp. 356–363).

Inkpen, D. (2007b). A statistical model for near-synonym choice. *ACM Transactions on Speech and Language Processing (TSLP)*, *4*(1), 2.

Inkpen, D., & Hirst, G. (2001). Building a lexical knowledge-base of near-synonym differences. In *Proceedings of the workshop on wordnet and other lexical resources, second meeting of the north american chapter of the association for computational linguistics (naacl 2001)* (pp. 47–52).

Inkpen, D., & Hirst, G. (2002). Acquiring collocations for lexical choice between near-synonyms. In *Proceedings of the acl-02 workshop on unsupervised lexical acquisition-volume 9* (pp. 67–76).

Inkpen, D., & Hirst, G. (2006). Building and using a lexical knowledge base of near-synonym differences. *Computational linguistics*, *32*(2), 223–262.

Islam, A., & Inkpen, D. (2010). Near-synonym choice using a 5-gram language model. *Research in Computing Sciences*, *46*, 41–52.

Islam, M. A. (2011). *An unsupervised approach to detecting and correcting errors in text* (Unpublished doctoral dissertation). University of Ottawa.

Jackson, H. (1988). *Words and their meaning.* Longman. Retrieved from `https://books.google.es/books?id=5qIrAAAAMAAJ`

Jackson, H. (2014). *Words and their meaning.* Routledge.

Katz, J. J. (1972). *Semantic theory.*

Kempson, R. M. (1977). *Semantic theory.* Cambridge University Press.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, *22*(2), 110–125.

Kermes, H., & Heid, U. (2003). Using chunked corpora for the acquisition of collocations and idiomatic expressions. *Proceedings of COMPLEX 2003*.

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on computational linguistics* (p. 1367).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... others (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the acl on interactive poster and* *demonstration sessions* (pp. 177–180).

Kreidler, C. W. (1998). *Introducing english semantics.* Psychology Press.

Krenn, B., Evert, S., et al. (2001). Can we do better than frequency? a case study on extracting pp-verb collocations. In *Proceedings of the acl workshop on collocations* (pp. 39–46).

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Sage.

Lahiri, S., & Lu, X. (2011). Inter-rater agreement on sentence formality. *arXiv preprint arXiv:1109.0069*.

Lahiri, S., Mitra, P., & Lu, X. (2011). Informality judgment at sentence level and experiments with formality score. In *International conference on intelligent text processing and computational linguistics* (pp. 446–457).

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

LDC, et al. (2005). *Linguistic data annotation specification: Assessment of adequacy and fluency in translations* (Tech. Rep.). Revision 1.5. Technical report.

Leckie-Tarry, H. (1995). Language and context: A functional theory of register (d. birch, ed.). *London: Pinter*.

Lin, D. (1998). Extracting collocations from text corpora. In *First workshop on computational terminology* (pp. 57–63).

Madnani, N., & Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, *36*(3), 341–387.

McGlohon, M., Glance, N. S., & Reiter, Z. (2010). Star quality: Aggregating reviews to rank products and merchants. In *Icwsm.*

Meredith, S. (2010). *Oscola: Oxford standard for the citation of legal authorities.* Faculty of Law, University of Oxford.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mosquera, A., & Moreda, P. (2011). Enhancing the discovery of informality levels in web 2.0 texts. *Proceedings of the 5th Language Technology Conference (LTC 2011).*

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, 491–538.

Palmer, F. (1981). *Semantics.* Cambridge University Press. Retrieved from `https://books.google.com.mt/books?id=UWJSaxH9GiMC`

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 115–124).

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2015, July). Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (pp. 425–430). Beijing, China: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/P15-2070`

Pearce, D. (2001). Synonymy in collocation extraction. In *Proceedings of the workshop on wordnet and other lexical resources, second meeting of the north american chapter of the association for computational linguistics* (pp. 41–46).

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the acl student research workshop* (pp. 13–18).

Pecina, P., & Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the coling/acl on main conference poster sessions* (pp. 651–658).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Quine, W. V. (1961). From a logical point of view (2nd ed.). In (p. 20-46). Cambridge, MA: Harvard University Press.

Quirk, C., Brockett, C., & Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 conference on empirical methods in natural language processing.*

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ravichandran, D., & Hovy, E. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 41–47).

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. (`http://is.muni.cz/publication/884893/en`)

Richards, J., Platt, J., & Platt, H. (1997). *Longman dictionary of language teaching and applied linguistics.* Longman. Retrieved from `https://books.google.es/books?id=k8xVtwAACAAJ`

Rickford, J. R., & McNair-Knox, F. (1994). Addressee-and topic-influenced style shift: A quantitative sociolinguistic study. *Sociolinguistic perspectives on register*, 235–276.

Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., & Liu, Y. (2007). Statistical machine translation for query expansion in answer retrieval. In *Annual meeting-association for computational linguistics* (Vol. 45, p. 464).

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval.* McGraw-Hill. Retrieved from `https://books.google.es/books?id=7f5TAAAAMAAJ`

Saussure, F. d. (1959). *Course in general linguistics.* (Ed. by C. Bally and A. Sechehaye. Trans. by W. Baskin.) New York: Philosophical Society. (Originally published as "Cours de linguistique générale", 1915)

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *Aaai spring symposium: Computational approaches to analyzing weblogs* (Vol. 6, pp. 199–205).

Shalaby, N. A., Yahya, N., & El-Komi, M. (2009). Analysis of lexical errors in saudi college students' compositions.'. *Ayn, Journal of the Saudi Association of Languages and Translation*, *2*(3), 65–93.

Sheikha, F. A., & Inkpen, D. (2010). Automatic classification of documents by formality. In *Natural language processing and knowledge engineering (nlp-ke), 2010 international conference on* (pp. 1–5).

Sheikha, F. A., & Inkpen, D. (2011). Generation of formal and informal sentences. In *Proceedings of the 13th european workshop on natural language generation* (pp. 187–193).

Shinyama, Y., & Sekine, S. (2003). Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on paraphrasing-volume 16* (pp. 65–71).

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics*, *19*(1), 143–177.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, *22*(1), 1–38.

Smedt, T. D., & Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, *13*(Jun), 2063–2067.

Somasundaran, S., Ruppenhofer, J., & Wiebe, J. (2007). Detecting arguing and sentiment in meetings. In *Proceedings of the sigdial workshop on discourse and dialogue* (Vol. 6).

Stede, M. (1993). Lexical choice criteria in language generation. In *Proceedings of the sixth conference on european chapter of the association for computational linguistics* (pp. 454–459).

Stolcke, A., et al. (2002). Srilm-an extensible language modeling toolkit. In *Interspeech* (Vol. 2002, p. 2002).

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Strunk, W. (2007). *The elements of style.* Penguin.

Terra, E., & Clarke, C. L. (2004). Fast computation of lexical affinity models. In *Proceedings of the 20th international conference on computational linguistics* (p. 1022).

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*(1), 178–185.

Turabian, K., Booth, W., Colomb, G., Williams, J., & Staff, W. (2009). *A manual for writers of research papers, theses, and dissertations, seventh edition: Chicago style for students and researchers.* University of Chicago Press. Retrieved from `https://books.google.com.mt/books?id=i6aXJLeZ2OMC`

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424).

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, *21*(4), 315–346.

Wang, T., & Hirst, G. (2010). Near-synonym lexical choice in latent semantic space. In *Proceedings of the 23rd international conference on computational linguistics* (pp.

1182–1190).

Warren, B. (1988). Semantics: word meaning. *Johannesson Nils-Lennart*, 61–95.

Wiebe, J. (2000). Learning subjective adjectives from corpora. In *Aaai/iaai* (pp. 735–740).

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, *39*(2), 165–210.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347–354).

Xu, W., Ritter, A., Dolan, B., Grishman, R., & Cherry, C. (2012). Paraphrasing for style. *Proceedings of COLING 2012*, 2899–2914.

Zhao, S., Lan, X., Liu, T., & Li, S. (2009). Application-driven statistical paraphrase generation. In *Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp: Volume 2-volume 2* (pp. 834–842).

Zhao, S., Wang, H., Liu, T., & Li, S. (2008). Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Acl* (Vol. 8, pp. 780–788).