universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„What's in a pun? Assessing the relationship between phonological and semantic distance and perceived funniness of punning jokes"

verfasst von / submitted by

Anna Palmann, BSc BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 013
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Joint Degree Programme MEi:CogSci Cognitive Science
Betreut von / Supervisor:	Dr. Tristan Miller

Acknowledgements

I would like to thank my supervisor Tristan Miller for his invaluable advice, the skills he taught me, and for bearing with me even though it took much longer than planned to write this thesis. Further, thanks to Lisl, Markus, and all the other teachers in the CogSci programme in Vienna and Ljubljana for providing me with the necessary skills to become a researcher.

Of course also a big thank you to my fellow CogSci students for two and a half (very remote) years of which we tried to make the best. Especially also to the ones who went to Ljubljana: thank you for a nice, yet short experience abroad with a lot of coffee meetings and the most amazing autumn hikes I've ever had. In that context, also thanks to Consti and Manu for their help with the statistical analysis for this thesis before the conference. And to Flávia for uncountable phone calls, for talking deep philosophy and exchanging Christmas cookie recipes, and for embracing the weirdness in the world and its inhabitants.

Much love to my friends in Maastricht and Vienna for their company during the writing process, and for sending me puns all the time - when I needed them and also when I didn't.

To my Kleeblatt, of course and always.

Many thanks to my two DnD groups for keeping me distracted, providing me with a nice balance to the academic work, and for making my life interesting during a pandemic and beyond.

And of course, thank you Nitzan, for mildly smiling at every stupid pun I make, for sharing this very particular type of humour and the conviction that talking about random things is often better than talking about serious ones. For endless conversations about language, cognition, and culture - ideally while cooking a way too complicated meal together.

I would furthermore like to express my thanks to the countless cups of tea that I drank while writing this thesis. You sometimes were the only thing I had to hold on to.

And lastly, I would like to thank my parents for exposing me to the very peculiar type of humour we share in our family. For showing me Loriot, Monty Python, and all the others when it was probably way too early. For sending me articles about AI, psychology, and cognitive science and for uncountable heated discussions during meals and walks. Even though sometimes exhausting, I am grateful for every single one of them.

Table of Contents

Chapter 1

Introduction

Punning is a form of wordplay based on the semantic opposition of two phonologically similar words. Of those two words, the one explicitly expressed in the punning joke is called pun while the target is the one whose meaning is invoked implicitly. The humorous nature of a punning joke is assumed to arise from the presence of semantic ambiguity between pun and target in a syntactic context, in which the meanings of both words are more or less acceptable (Kao et al., 2016). Given that humour in punning jokes relies strongly on phonological, i.e. sound-related, features, several attempts have been made to quantify those and investigate possible correlations with acceptability and comprehensibility (for an overview see Hempelmann & Miller, 2017). However, phonology is only one aspect in the description and analysis of linguistic information in general and semantic, i.e. meaning-related features, are just as important. Therefore, the success of a punning joke is assumed to be determined by both phonological and semantic aspects, alone and in interaction (Hempelmann, 2003).

1.1 Objectives

Even though the success of a punning joke depends on a number of different interacting factors, it seems worthwhile to take a closer look at more fine-grained linguistic aspects located in the fields of phonology and semantics. To date there has been no empirical investigation of the relationship of phonological as well as semantic features of pun and target word and the perceived funniness of punning jokes.

Thus, it is the aim of this study to investigate the relationship between phonological as well as semantic aspects within a punning joke and its perceived funniness. In more detail, this study aims to investigate possible correlations between phonological distance and semantic similarity of pun and target word on the one hand, and the perceived funniness of a punning joke on the

other hand. The goal of the present empirical investigation is to answer the following research questions:

What is the relationship between the phonological distance between pun and target word and the perceived funniness of a punning joke?

What is the relationship between the semantic distance between pun and target word and the perceived funniness of a punning joke?

Phonological distance will be calculated using several distance measures within the Python library *abydos* (Little, 2018). Semantic similarity between pun and target will be calculated using several different measures (for an overview of different approaches see Budanitsky & Hirst, 2006). These measures are based on sense identifiers from the semantic network WordNet (Fellbaum, 1998). Furthermore, another approach to obtain semantic distance measures based on word embeddings in the Word2Vec model (Mikolov et al., 2013) will be applied. Afterwards, possible relationships between semantic distance of pun and target and funniness ratings will be assessed. Funniness ratings were obtained in a crowdsourcing study using Amazon's Mechanical Turk. While initially collected in the form of pairwise judgements, scores will be transformed into ranks using the methods of best–worst scaling and Gaussian process preference learning.

1.2 Hypotheses

Based on previous analyses of punning jokes (e.g. Lagerquist, 1980), it is hypothesised that punning jokes with lower phonological distance between pun and target word will be associated with higher funniness ratings. More simply put, the more pun and target sound alike, the funnier the punning joke is perceived as. When comparing types of punning jokes, this leads to the hypothesis that homographic puns – which are in most cases also homophonic – result in higher funniness ratings when compared to heterographic puns – which are more likely to also be heterophonic.

Regarding semantic distance, there are no directed hypotheses. One possible outcome could be that the semantic distance associated with highest funniness ratings is located in the middle range. That might be because it can be assumed that too closely related meanings of pun and target do not evoke script opposition, which is necessary for a humorous effect (Attardo & Raskin, 1991). On the other hand, too distant meanings would result in failure to make sense of the joke or detect its humorous intention in the first place, which is necessary for humour to arise (Ritchie, 2018).

1.3 Outlook

In this thesis, we will first give an introduction to humour studies in general. This is followed by an introduction to the field of linguistic humour with specific focus on the phenomenon of punning jokes. After that, the data collection and dataset, as well as the various methods for the calculation of phonological and semantic distance will be described. This is followed by a presentation of the results obtained in the correlation analyses. These results will then be interpreted and critically discussed. Finally, they will be put into context within a larger theoretical framework followed by suggestions for future research on the topic.

Chapter 2

Theoretical background

2.1 The phenomenon of humour

The term *humour* denotes a number of different concepts. On an individual level, it can be a psychological state, a personality trait, or an emotion. At the same time, the term can also stand for sense of humour – the ability to express and appreciate what is humorous. On a larger scale, the term humour is also used to describe a complex social and cognitive construct (Hempelmann, 2017).

2.1.1 Humour and cognition

Several cognitive mechanisms have been found to be associated with humour. These are for example general and attitudinal knowledge, including the understanding of common cultural and social practices, as well as the detection and correct placement of attitudes (Uekermann et al., 2007). But also working memory and other executive functions such as switching and inhibition are crucial in order to process humorous utterances. While *switching* or cognitive shifting denotes the ability to redirect attention, *inhibition* is the ability to ignore irrelevant stimuli or information. Both these mechanisms are important in order to apply frame shifting between two opposite meanings in ambiguous situations or utterances, giving rise to humour. Additionally, the ability to draw inferences as well as to employ theory of mind and empathy are crucial for humour processing – especially in social situations (Uekermann et al., 2007). Furthermore, humour involves both expressive abilities during the generation of humorous stimuli and receptive abilities during comprehension and appreciation of such (Uekermann et al., 2007). Regarding the latter, it needs to be pointed out that even though smiling or laughter have been found to be the most common reactions to humourous utterances or situations, they do not necessarily correspond directly to enjoyment or appreciation (Guidi, 2017). The exact way, amount, and

targets of humorous expression have been found to vary greatly across cultures, languages, social classes, or ethnic groups. Nevertheless, humour as well as laughter or smiling are considered universal innate expression patterns present across cultures (Guidi, 2017). According to Guidi (2017), the universal character of humour can be assessed on three different dimensions: conceptual features, phenomena, and aspects of these phenomena. *Conceptual features* are general working mechanisms of humour that are present cross-culturally, such as ambiguity, the violation of social norms, or unexpected turns of events. *Phenomena* denote the ways of expressing humour, such as jokes, puns, or irony. These vary greatly across cultures. *Aspects* of such phenomena denote the way in which they are arranged and how joke mechanisms are expressed. These too vary depending on culture with the exception of verbal expression, which is present universally (Guidi, 2017).

2.1.2 The social function of humour

Humour can be seen as a form of communication present in all historical periods and cultures (Larkin-Galiñanes, 2017). It is considered an important tool in human social interaction (Uekermann et al., 2007) with the purpose of establishing bonds and relationships during complex communicative situations that require negotiations of meaning on different levels (Brône et al., 2006). Therefore, mutual understanding in humour depends strongly on the shared background of speaker and listener (Brône et al., 2006). According to Maraev et al. (2020), humour in interaction is dependent on general and domain-specific knowledge about language and the world. A similar degree of knowledge is necessary on both sides, the one of the speaker who needs sufficient cognitive and linguistic performance, and the listener who needs sufficient cognitive and linguistic competence for a joke to work (Lagerquist, 1980).

In terms of social functions of humour, two main mechanisms can be described from an anthropological perspective (Larkin-Galiñanes, 2017). One is *social management*, which serves the purpose of pointing out the violation of rules in a society, reinforcing social relationships or gaining attention as an individual. In line with this, Henri Bergson in his famous essay *Le rire* describes the social function of humour as a way of keeping people to adhere to social norms, so that they would act in a certain way in order not to be laughed at (Larkin-Galiñanes, 2017). Humour has further been found to be a tool for resolving inter-group conflicts (Kao et al., 2016). Another function of humour is *defunctionalisation*, which relates to playful social situations where language is used for humorous purposes. This involves the establishment of a joking relationship between actors in a communicative situation, but also the assessment of social circumstances in order to understand whether they support or ban humour (Larkin-Galiñanes,

2017). Logically, humour emerges more in socially playful or paratelic (as opposed to goal-oriented or telic) situations (Smith et al., 2020). In this state, social clues indicate a humourous mindset and incongruities are more accepted or even perceived as stimulating.

2.1.3 Humour theories – an overview

The oldest theories on humour stem from ancient Greek philosophy. Both Plato and Aristotle considered humour to be vulgar and not suited for virtuous and free people (Larkin-Galiñanes, 2017). According to Plato, laughter as form of emotion needed to be restricted by reason. During the European Middle Ages – for the most part influenced by a Christian world view – this negative attitude towards humour persisted. Only starting in the 18th and 19th century did humour become subject of philosophical discussions and looked at in a more objective manner (Larkin-Galiñanes, 2017). A number of different theories on humour have been developed throughout the ages and by scholars from various disciplines, such as philosophy, psychology, biology, or anthropology. Humour theories can be roughly subdivided into three categories: incongruity, superiority, and release theories (Raskin, 1985). While *incongruity theories* are concerned with the nature of humorous stimulus and are thus necessarily connected to the speaker, *superiority theories* focus on the relationship between speaker and listener. *Release theories* on the other hand are based on features of the psychology of the listener (Attardo & Raskin, 2017). In the following, a selection of representative humour theories from those three categories will be described. However, it needs to be pointed out that there exist several more theories that will not be addressed in detail.

Incongruity theories

Starting with Aristotle and Cicero, authors underlined the importance of surprise and the deception of expectations in order for humour to arise. This view is based on the idea that the listener's initial expectations are broken in an unexpected manner leading to a sensation of incongruity and surprise. In these so-called incongruity-resolution theories, humour is viewed as two-step process (Uekermann et al., 2007). In the first step, an incongruous element is detected among several incompatible elements, while in the second step the incongruent element is linked in a sense-making manner to the context – leading to resolution of the incongruity and a subsequent humorous sensation (Uekermann et al., 2007). According to incongruity-resolution models, incongruity is a necessary feature and its resolution a sufficient feature of humour (Hempelmann, 2003).

In the 1970s, more structured information-based approaches described the process of de-

tecting incongruity followed by its resolution. According to those approaches, this process is similar to a problem solving task, with the goal to identify the punchline of a joke (Koestler, 1964). In this context, the term *bisociation* (Koestler, 1964) is used to describe the action of simultaneously viewing a situation or an object from two different angles that are normally incompatible. Through this clash or combination of two different conceptual levels, mental routines are disrupted and humour but also creativity can emerge (Koestler, 1964).

Kao et al. (2016) point out that one issue regarding incongruity theories is the lack of a uniform definition of the term incongruity. Subsequently, it remains unclear how to observe its presence in a humorous situation or utterance. This makes it complicated to empirically test the role that incongruity really plays in the creation and processing of jokes. Additionally, according to Hempelmann (2003), the questions of whether an element is incongruent in a certain context and whether there is script opposition present is a matter of situation, cultural context, and individual knowledge resources. All of those can vary greatly and influence incongruity to different extents and in interaction, which makes it difficult to grasp and describe the emergence of incongruity in jokes.

Superiority theories

In the 18th century, superiority theories became the most popular view on humour. They describe the act of laughter and generally the expression of humour as indicative of the relationship between speaker and listener. Superiority theories argue that humour requires either elements of violation, surprise, social distance, or temporal distance. Further, they state that humour is generally based on a negative attitude towards the subject of the humorous utterance or situation (Attardo, 2014).

According to the *disparagement theory* – a form of superiority theory – humour results from a sense of superiority caused by disparagement of another person's or one's own foolish behaviour (Uekermann et al., 2007). This highlights the importance of social knowledge – especially also since there seems to be a correlation between the subjective perception of the funniness of a situation or joke and the extent to which the disparaged person is liked. Accordingly, a joke is perceived as funnier when the aggressor (and not the victim) is closer in social status to and liked more by the rater (Uekermann et al., 2007). Combining incongruity resolution and disparagement theories leads to the notion that unexpected disparagement makes more sense when the victim is disliked, which in turn leads to incongruity reduction (Uekermann et al., 2007). Contrarily, *benign violation theories* argue that the listener needs to notice a violation and rate it as benign in order for humour to occur (Attardo, 2014).

Release or relief theories

Release or relief theories adopt a slightly different view on humour, focusing on the release of excess energy or tension that built up in a situation due to various reasons and in different ways. In line with this view goes the finding that humour improves personal psychological well-being (Kao et al., 2016). Sigmund Freud propagated the role of humour as serving the purpose of liberating primary instincts of aggression that humans usually need to repress when living in a civilised society (Larkin-Galiñanes, 2017). According to this view, humour serves the purpose of saving psychological energy because there is no necessity for inhibition in a humorous situation, which makes it so pleasant for the involved actors. According to Freud (1961), the main purpose of humour is thus stress relaxation on an individual level and helps to temporarily ignore fear of authorities and societal rules. Similarly, the *arousal-jag theory* (Berlyne, 1960) focuses on the build-up of arousal through tension created during a comic situation. This tension is subsequently resolved, which leads to a humorous relief reaction (Hempelmann & Attardo, 2011). A slightly different type of relief theory describes the assumption that humorous amusement is essentially the reward for successful mental error detection (Roberts, 2017). This is based on the assumption that humans create quick – and thus fallible – heuristics while perceiving the world, alongside of which mental safeguard processes are executed. The detection of conceptual errors or ambiguities can be seen as such safeguard actions and if they succeed, a feeling of reward and subsequently humour can arise (Roberts, 2017).

2.2 The linguistics of humour

Frequently, humour is part of social communicative situations and transmitted linguistically. Already in classical Greek philosophy, the strong connection of language and humour was acknowledged and a number of linguistic devices to create humour were defined (Larkin-Galiñanes, 2017). These were for example homonyms and synonyms, unconventional use of language, exaggerations, punning, irony, ambiguity, unexpected turns, strange resemblances, metaphors, or comparisons (Larkin-Galiñanes, 2017). The manipulation of linguistic features can furthermore happen on several linguistic levels, such as pronunciation, spelling, morphology, vocabulary, or syntax (Giorgadze, 2014).

2.2.1 Theories on linguistic humour

The *Semantic-script Theory of Verbal Humour* (SSTVH) developed by Raskin (1985) is a form of incongruity theory based on script opposition as universal mechanism of linguistic humour

(Guidi, 2017). Its main assumptions are that a text under investigation is fully or partly compatible with two different overlapping scripts and that these scripts are opposite in a special predefined sense (Raskin, 2017). In this framework, a *script* or *frame* denotes semantic information associated with a certain word including connected concepts and associated knowledge or experiences. While script overlap – the compatibility of the joke with two scripts – is a necessary condition, the opposition of the two scripts is a sufficient condition for incongruity to arise and a joke to succeed (Hempelmann, 2003). The opposition of scripts depends on situational and contextual factors. Further, there necessarily need to be lexical triggers present that allow to switch from one script to another. These triggers correspond to the punchline of the joke and can be based on contradiction or ambiguity in the sense that they require the listener to rethink the initial script they chose (Hempelmann, 2003).

According to the *five-level joke representation model* (Attardo & Raskin, 1991), there are five levels of hierarchical joke representation or abstraction. These are

- **surface**, the actual text of the joke
- **language**, such as specific words or syntactic structure
- **target and situation**
- **template**, the role model for the script opposition and
- **logical mechanism**, the more basic combination of script oppositions.

One of the most prominent theories in the field of linguistic humour is the *General Theory of Verbal Humour* (GTVH), developed by Attardo and Raskin (1991). It can be seen as revision and combination of the previously described SSTVH (Raskin, 1985) and five-level joke representation model (Attardo & Raskin, 1991). The term *general* denotes the fact that unlike the purely semantic SSTVH, the GTVH incorporates information on phonological, morphological, and sociological levels (Attardo, 2017). Even though it aims at describing instances of verbal humour, the GTVH is not a purely linguistic theory, but can be rather used to describe humourous instances in general and in an overarching manner (Attardo & Raskin, 1991). Further, it is important to note that the GTVH is not a model of joke production but rather of joke processing (Attardo & Raskin, 1991). Based on the critique that the SSTVH does not differentiate between referential and verbal humour, and also does not take into account similarities between certain types of jokes, the GTVH introduces four more parameters resulting in six so-called *knowledge resources* (Attardo, 2017). According to the GTHV, these six parameters have to be present in a humorous text in order for it to succeed. They are hierarchically organised and are connected

via binary logical relations. Based on this structure, an indexed taxonomy of joke variance and invariance can be created. In this taxonomy, each type of variance between two jokes is described through one or more knowledge resources indicating differences (Attardo & Raskin, 1991).

The six parameters or knowledge resources of joke difference defined by Attardo and Raskin (1991) are the following:

- language
- narrative strategy
- target
- situation
- logical mechanism
- script opposition

The **language** resource describes differences in choice of words or sentence structure (Attardo, 2017). It contains a full phonological, morphological, syntactic, and lexical description of the text of a joke. Additionally, it contains statistical information about the frequency of occurrence of certain linguistic units (Attardo, 2017). This is especially interesting when taking into consideration that jokes frequently are paraphrases of each other, where not the meaning but only the exact way of presentation changes. Another characteristic of jokes is that they belong to non-casual language – a type of expression containing an additional layer of meaning, that identifies them as joke. This is for example similar to instructions in textbooks. The linguistic features hinting to this additional purpose can be described using the language resource (Attardo & Raskin, 1991).

Narrative strategy describes how the text is organised and where the humorous element is placed (Attardo, 2017). This relates to the the genre, which the joke is ascribed to. That can for example be a riddle, a question–answer situation, or a simple narration (Attardo & Raskin, 1991). The punchline or semantic script-switch trigger (Hempelmann, 2003) – the part of the joke that is responsible for resolving the incongruity that previously emerged through ambiguity or contradiction – can be expressed either in a straightforward or in a parallel-structured manner and is most frequently located at the final or pre-final syntactic position in a joke (Attardo & Raskin, 1991).

Target describes the "victim" of the joke, assuming that jokes are aggressive in the sense of a superiority theory view, and that this aggression has a direct target – be it a person, an institution, or a belief (Attardo, 2017). Targets can be individuals or groups to which very often

the stereotype of dumbness is ascribed to. However, not every joke necessarily needs to have a target which makes this knowledge resource optional (Attardo & Raskin, 1991). Accordingly, Freud distinguished between tendentious (targeted) and non-tendentious humour (Attardo & Raskin, 1991).

Situation describes the background, in which the events of a joke take place (Attardo, 2017). Besides that, it also includes a description of the props (i.e. participants, objects, instruments, etc.) necessary for the narrative of the joke to work (Attardo & Raskin, 1991).

Logical mechanism is the knowledge resource that explains the incongruity and its resolution (Attardo, 2017). One of the most prominent and simple logical mechanisms is the previously mentioned juxtaposition of two situations based on ambiguity (Attardo & Raskin, 1991).

Script opposition denotes the two opposed and overlapping scripts involved in the joke (Attardo, 2017). In the SSTVH, three levels of script opposition in different stages of abstraction are proposed. These are real vs. unreal, actual vs. non-actual, and other simple oppositions such as good vs. bad (Attardo & Raskin, 1991).

2.2.2 Linguistic ambiguity

Ambiguity is an inherent feature of language. By nature, linguistic elements are ambiguous and disambiguation happens only through context, both linguistically and more generally (Aarons, 2017). Listeners are thus prone to use contextual information for disambiguation, and thereby tend to interpret a signal in the way that makes most sense in a given context (Aarons, 2017). However, linguistic ambiguity is of course only a necessary but not a sufficient condition for a joke since not all ambiguity is automatically humorous (Aarons, 2017). In most communicative situations with no humour involved, ambiguity resolution is assumed to happen as follows. First, ambiguity is detected and its source – either deliberate or unintended – is identified. After that, the ambiguity is resolved by using contextual information and communication can continue in an unhindered manner (Aarons, 2017). However, the linguistic ambiguity present in a joke falls under the category of so-called *non–bona fide communication* (Raskin, 1985). This describes mutual engagement of speaker and listener in humorous play, where language is intentionally defunctionalised and serves not only the purpose of transmitting information but also humorous engagement of speaker and listener. Non–bona fide communication is usually initialised in discourse (Aarons, 2017) using conventional markers such as standard beginnings of jokes, a change in prosody, or non-linguistic communicative signals.

In this context, also the concept of *framing* plays an important role. Framing describes the process of viewing a linguistic phenomenon against the background of a frame of reference

(Brône et al., 2006). Those frames are structured categories based on experience (Brône et al., 2006). In line with the previously described approaches, jokes are based on opposition, overlap, and switch of two contextually different scripts or frames. In the first part of the text, only one script is activated, but the punchline is incompatible with the first interpretation. Only the script-switch trigger, a lexical cue, enables the switch from one script to the other one.

According to the *Graded Salience Hypothesis* (Giora, 1997), salient information is always accessed before less salient information, and this principle can be exploited in humorous expressions relying on linguistic ambiguity (Brône et al., 2006). This happens in a way that in jokes the more contextually salient meaning is accessed first. After reading the punchline, this meaning is discarded in favour of a more marked reading (Brône et al., 2006). Further, it is assumed that if a certain topic is made salient to a listener, this leads to increased appreciation of humorous stimuli connected to this topic (Uekermann et al., 2007).

2.3 The phenomenon of punning

Punning or paronomasia (Lagerquist, 1980) is a form of humorous wordplay based on the semantic ambiguity between two phonologically similar words occurring in a sentence context, in which both meanings are more or less acceptable (Hempelmann & Miller, 2017). A punning joke can therefore be defined as figure of speech in which two similar words or phrases are deliberately confused for a rhetoric effect (Giorgadze, 2014). A punning joke contains a pun word and a target word. The pun word is actually present in the sentence, while the target word is evoked through the similarity in sound with the pun word. Punning jokes can appear in a variety of textual categories, such as slogans, titles, or canned jokes, but are also created spontaneously in conversation (Dynel, 2010).

In common parlance, both punning jokes and pun words can be referred to as *puns*. For further distinction, in this thesis the term *punning joke* is used only to refer to a complete joke – consisting of one or more sentences – containing a pun and target word. The term *pun word* on the other hand, refers to a specific (ambiguous) part of a punning joke. In this thesis, the term *pun* will be used in some contexts when referring to a punning joke and in other contexts when referring to a pun word. However, when a clearer distinction is necessary for the comprehension of certain concepts or approaches, the more precise terms be *punning joke* and *pun word* will be used in order to avoid misunderstandings.

2.3.1 The taxonomy of puns

There are several partly overlapping approaches to the classification and taxonomy of puns depending on differences in the assessment of the phenomenon (Giorgadze, 2014). These can be based on linguistic aspects but also situational context or differences in interpretation of the punning joke.

In this context, it needs to be mentioned that one problem with systematic taxonomies of puns based on linguistic features is that they are mainly centered on Indo-European languages and for example disregard features such as tonality as influencing factors (Guidi, 2017). While problematic, this cannot be taken further into consideration in the empirical part of this thesis, which exclusively focuses on the investigation of English puns.

Punning jokes are a form of humorous wordplay based on the presence of two signs or words indicating a double meaning (Hempelmann & Miller, 2017). The taxonomy of wordplay has its beginnings in classical rhetoric and was later revived in medieval rhetoric. According to Hempelmann and Miller (2017), the following types of punning jokes are originally described in classical and medieval rhetoric:

- **traductio**, where the same words is used twice;
- **adnominatio**, where slightly different words are used twice; and
- **significatio**, where the same word is used once.

More recent literature commonly uses a different classification based on homonymy and heteronymy (Hempelmann & Miller, 2017). *Homonymous* or perfect puns are those that are either homophonic (i.e. pun and target sound exactly the same) or homographic (i.e. pun and target are spelled in the exact same way) or both. However, it needs to be pointed out that the exact type of homonymy depends on the initially chosen definition, which makes the term ambiguous. Lagerquist (1980) adds that in homonymous puns, pun and target word also belong to the same syntactic category. *Heteronymous* puns on the other hand are either heterophonic (i.e. pun and target sound different) or heterographic (i.e. pun and target are spelled in a different manner) or both – again depending on the exact definition. Such punning jokes are also referred to as paronomasic, paronymic, or imperfect puns (Lagerquist, 1980). The difference in both sound and orthography between pun and target word can be seen as a continuum with either slight or large differences. The rare case of a pun that is homographic but heterophonic (i.e. pun and target have identical spelling but do not sound the same) is called an *eye pun*.

In both the homonymous and the heteronymous case the target word can either be explicitly expressed in the punning joke indicating a syntagmatic relationship between pun and target, or

just be implicitly evoked through the pun, which results in a paradigmatic relationship. In both cases, it is possible that only the pun word (single sign) or pun and target word (double sign) are present. Table 1 (taken from Hempelmann & Miller, 2017) depicts this subdivision.

Table 1

One approach to subdivide puns based on the presence of single and double signs

A. heteronymy	*classical term*	*[horological instrument vs. male genitalia]*
1. single sign/ paradigmatic	(pure)	your clock is very big
2. double sign/ syntagmatic	*adnominatio*	his clock was bigger than his cock
B. homonymy		*[domesticated male bird vs. male genitalia]*
1. single sign/ paradigmatic	*significatio*/syllepsis	the farmer has a big cock
2. double sign/ syntagmatic	*traductio*/antanaclasis	the cock has a big cock

Using the aforementioned categorisations, we can differentiate other special types of punning jokes. *Malapropisms* for example, are imperfect paradigmatic puns, where one word is not intentionally but accidentally used in place of another, thereby creating an unintentional humorous effect (Attardo, 2014). *Spoonerisms* are another type of imperfect paradigmatic or syntagmatic pun, where the beginning sounds of two words within a phrase are exchanged (Attardo, 2014).

However, not only the direct relationship between pun and target can be used for the classification of punning jokes, but also their interaction with the sentence context can lead to the creation different subcategories. According to Giorgadze (2014), another common differentiation is the one between

- **lexical homonymy**, where the focus lies on single-word ambiguity;
- **collocotional homonymy**, where word-in-context ambiguity is most important; and
- **phrasal homonymy**, where ambiguity between clauses is the process in play.

Another approach is a differentiation of punning jokes according to the source of ambiguity. Giorgadze (2014) assumes lexical, syntactical, structural, and semantic elements to play a role in this and creates the following classification of punning jokes:

- **lexical-semantic puns**: the classical form described above based on polysemantic homonyms, i.e. words with identical or similar sound and different meaning

- **structural-syntactic puns**: a phrase or a sentence that can be parsed in several ways, where changing grammatical categories of pun and target word lead to a difference in meaning

- **structural-semantic puns**: one of the meanings of an ambiguous word or phrase is based on its widespread use, e.g. in form of idiomatic expressions

Aarons (2017) differentiates similar categories as Giorgadze (2014) but applies a different nomenclature. According to the author, the most identifiable pun type – the so-called *phonological-semantic pun* – is based on complete homophony. A different pun type is the *phonological-morphological-syntactic pun*, where sounds stretch over more than one word (e.g. over several syllables) resulting in a change in the syntactic structure (Aarons, 2017). Compared to the phonological-semantic pun this type requires more cognitive effort in order to identify both meanings. Yet another pun type is the *phonological-semantic-syntactic pun*, where changes in the sentence structure are interconnected with semantic changes, and phonology and semantics are in strong interplay.

In connection to this, another way of categorising puns is based on the way that ambiguity emerges (Dynel, 2010). Attardo et al. (1994) differentiate between the connector, the ambiguous element in a punning joke (which according to the previously used terminology would be called pun), and the disjunctor, the element in the sentence that causes the passing from one meaning to another (i.e. the script-switch trigger). In the *distinct connector configuration*, the connector precedes the disjunctor, while in the *non-distinct connector configuration* the two coincide in one part of the punning joke.

Yus (2003) bases his categorisation of puns on relevance theoretical ideas. The first type describes punning jokes, in which both meanings are balanced and the listener goes back and forth between the two and is not able to decide which one is more relevant. In the second type, the listener identifies one interpretation but keeps looking for another one. The third type describes punning jokes, where the listener does not reach an interpretation at first and even after resolution of ambiguity the joke stays nonsensical. The fourth type describes punning jokes, where the first part of the joke contains an ambiguous term, which is exchanged for a less likely but still possible interpretation.

Dynel (2010) proposes another categorisation of punning jokes based on the investment of cognitive effort and the gain of cognitive benefits during processing. The author differentiates between double- and single-retention puns, depending on whether both (double-retention) or only one (single-retention) meaning of the ambiguous element are kept in the final interpretation. Double-retention puns can either be syntagmatic or paradigmatic. Syntagmatic puns can be

divided further depending on the type of activation of the second, not overtly expressed meaning. Single-retention puns are also further differentiated depending on the position of the pun word and the order of meaning activation. As a third and final category, Dynel (2010) mentions *zero-meaning puns* (or groaners), which only carry absurd or nonsensical information.

Yet another approach to classify punning jokes is based on their emergence in conversational context. According to Ritchie (2005), *self-contained puns* can be used in various situations since their humour does not rely on the direct context but more on general knowledge. The semantic context is therefore arbitrary and not essential to the factors that make an utterance a pun. *Contextually integrated puns* on the other hand, are embedded in a certain discourse context, referring for example to things in the field of view of speaker and listener or a recently happened event or discussion, which both attended. The context is thus not arbitrary and the features of the pun have to be compatible with it (Ritchie, 2005). Contextually integrated puns lose their humorousness when taken out of context, even though target recovery may still be possible on a linguistic level (Jaech et al., 2016).

2.3.2 Why we think puns are funny

In general, many of the theories on linguistic humour mentioned in section 2.2.1 also cover the phenomenon of punning jokes. Yet, re-visiting the underlying theories in more detail and thereby investigating the specific working mechanisms of punning jokes, seems worthwhile. This is because puns as instances of linguistic humour allow for an investigation of humorous processes in social communicative situations as well as in isolated form.

Aarons (2017) describes the general processing of a punning joke as follows: the pun word is detected and identified as homophonous by retrieving the target word, which leads to the notion of ambiguity. This is followed by the understanding that both ways of reading can be equally meaningful in the given sentence context, even though one was initially prioritised. The incongruity that arises due to the ambiguity results in surprise and is resolved through the notion that the word in focus has more than one meaning despite a similarity in sound. This resolution evokes a humorous sensation.

The understanding of a punning joke is based first and foremost on the ability of the listener to recover the target word. In order to do so, the listener relies on phonetic information and language context (Jaech et al., 2016). Therefore, the production and comprehension of punning jokes – just as joke comprehension in general – requires linguistic knowledge, be it tacit or conscious (Aarons, 2017). Additionally, context and general knowledge are essential for target recovery. This is essentially similar to general speech comprehension, where acoustic information

in combination with conversational context are used to make sense of spoken words (Hempelmann, 2003). Hempelmann (2003) brings up the concept of *torso*, which denotes the syntactic context of a punning joke. This torso very often comes in form of formulaic expressions such as phrases, collocations, idioms, proverbs, or titles. The fact that the pun is embedded in those well-known and established syntactic structures, makes it easier to recover the target because of the salience of such expressions (Guidi, 2017).

Dynel (2010) views the phenomenon of punning from the perspective of relevance theory (Wilson & Sperber, 2002), where relevance can be seen as necessary presumption for investing cognitive resources. According to this theory, the listener tends to invest as little cognitive effort as possible and as much as necessary to understand a given stimulus. This is of interest in the case of punning jokes since they require the listener to invest more effort in order to be able to understand the ambiguity. It is assumed that the more relevant and less absurd meaning is automatically activated and only through investment of more cognitive effort does the second meaning become obvious. After the first interpretation has been made, the sentence context and more precisely the script-switch trigger make it necessary for the listener to "think twice" in order to grasp both meanings. This willingness to invest can be explained by the rewarding experience that follows the understanding of the pun in form of intellectual satisfaction and an emergence of humorousness in a social situation (Dynel, 2010).

Roberts (2017) argues that not every pun is automatically humorous. According to the author, there is an underlying logical structure for puns and the components of this structure need to be arranged in a certain way in order to evoke a sensation of funniness. This notion can be placed in the framework of the joke analysis theory developed by Ritchie (2004). This theory describes a number of elements explaining the relationship between pun and target word and subsequently the two different ways of interpreting the ambiguous punning joke. These elements (cf. Ritchie, 2004) are obviousness (i.e. the pun word is more likely to be noticed first), compatibility (i.e. the punchline makes sense also in combination with the target word), contrast (i.e. there is a significant difference between pun and target word), and inappropriateness (i.e. the interpretation that is created by integrating the punchline with the target word is inherently odd, eccentric, or taboo and thus violates a norm). Roberts (2017) notes further that norm violation has to always be put in a social context and can also vary according to that. Furthermore, there seem to exist puns in which no inappropriateness is present – resulting in the conclusion that this is not a necessary condition for the success of punning jokes.

Attardo's General Theory of Verbal Humour provides a framework for understanding the nature and mechanisms of punning jokes (Aarons, 2017). According to Attardo et al. (1994),

puns are based on *cratylism*, a folk theory of false logic. Its name derives from the Platonic dialogue "Kratylos" (Hempelmann & Miller, 2017), where a character assumes there to be a motivated relationship between the form of a sign and its meaning. This stands in direct contrast to the Saussurian idea of the arbitrary relationship between signifier (the linguistic symbol) and signified (its meaning). Even though logically not correct, cratylism is often employed in everyday communication, for example when people claim that a word is onomatopoetic, i.e. that it sounds the way it looks or feels like (Aarons, 2017). Creating or understanding punning jokes requires speaker and listener to initially assume that the involved sound sequences are cratylic, meaning that they have a unique reference point in the world and that there is no ambiguity present at all (Aarons, 2017). Only then, the structure of the joke highlights the script opposition, which brings out the incongruity between the cratylic and the non-cratylic use of language. For this, the listener needs to employ meta-linguistic knowledge resulting in resolution of ambiguity and a non-cratylic attitude, i.e. the notion of ever-existing ambiguity in every utterance (Aarons, 2017). In punning jokes, the logical mechanism (one of the knowledge resources in the GTVH) consists of this cratylistic view. Cratylism, like all other logical mechanisms, is based on what has been entitled *willing suspension of disbelief* (Hempelmann, 2003). This describes the willingness of the listener to accept false logic occurring in a communicative situation for the larger scope of it. In the case of puns, this would be the acceptance of a direct relationship between the incongruous concepts involved (Hempelmann, 2003). Thus, puns fall under Raskin's notion of *non–bona fide communication* implying that one of the essential elements for the success of a punning joke is that the communicative situation is perceived as a playful one (Aarons, 2017). If not, there is a risk for the pun not to be recognised as joke but classified as speech error. All the previously mentioned aspects set the base for differentiating puns from phonological-lexical errors, slips of the tongue, or spontaneous word blends (Guidi, 2017).

Ambiguity is especially strong in a punning joke when both senses have a high probability of making sense in the given context (Hempelmann & Miller, 2017). Ideally, both meanings are supported by different words in a context, indicating a high distinctiveness between the two. This distinctiveness can be calculated as the probability of the word meaning given the context. This calculation is based on how often a certain word has previously been observed in the vicinity of these words (Hempelmann & Miller, 2017). According to Hempelmann and Miller (2017), there seems to be a correlation between distinctiveness of pun and target word and funniness ratings of punning jokes in a way that higher distinctiveness results in higher funniness ratings. Similarly, Kao et al. (2016) found significant correlations between funniness ratings of punning jokes and both ambiguity as well as distinctiveness scores. For their analysis, they used a computational

model of language understanding to predict humour. The authors defined ambiguity as presence of two meanings that are both compatible with the sentence context, and argued that this alone is not sufficient for a pun to be perceived as funny but that additionally distinctiveness plays a crucial role. They defined distinctiveness as support coming from other words in the sentence context, which can be measured in form of the degree to which each meaning is supported by the different parts of a sentence. In order to assess the relationship of the meaning of a sentence and the words it is made out of, Kao et al. (2016) created a simple probabilistic generative model. They then inferred the joint probability distribution over sentence meanings and semantically relevant words using Bayes' theorem. Subsequently, they used this model to predict ambiguity and distinctiveness in a sample of punning jokes taken from the website "Pun of the day" and looked at their ability to predict humorousness. The jokes were rated in funniness on a scale from 1 to 7 by human raters taking part in the assessment through the website Amazon's Mechanical Turk. Kao et al. (2016) found that both ambiguity and distinctiveness were higher in puns when compared to non-puns. Further, puns with higher distinctiveness were perceived as funnier when compared to those with lower distinctiveness. The authors therefore conclude that ambiguity (in the sense of incongruity) is used to distinguish humorous from non-humorous sentences, while distinctiveness and the resolution of incongruity may contribute to differentiating between degrees of humorousness.

As for any type of joke, one crucial element of puns seems to be the evoking of tacit linguistic knowledge (Aarons, 2017). This is based on the assumption that most of linguistic knowledge is stored subconsciously on different levels and is only made conscious when the listener's attention is drawn to it by disruption of normal language processing, for example due to an unusual word. The listener then makes use of this tacit knowledge to solve the ambiguity that arises (Aarons, 2017). In terms of incongruity, some puns may be perceived as funnier because they are more novel or provide a more satisfying resolution of incongruity (Smith et al., 2020). Funnier puns are thus assumed to be richer in connections that enhance the aptness and the noticing of the incongruity (Smith et al., 2020). According to Simpson et al. (2019), the perceived funniness of a joke can vary depending on content and structural features but also on the cognitive effort necessary to recover the target word. In that context, Binsted et al. (2003) argue that comprehensibility is crucial, so that a pun must not be too easily understandable (lest it becomes facile and obvious) nor too difficult (which would result in it being more a riddle than a joke). Binsted et al. (2003) further note that prosody, in the sense of pronunciation and timing of a spoken sentence, is a relevant tool to differentiate between the two possible ways of reading a pun. Assuming that canned jokes in general – and also canned punning jokes as

subcategory of such – usually consist of two clauses, another way of differentiation of the two readings can be based on the relation between the first and the second clause. In more detail, this relates to whether the expectations that have been built up in the first clause are confirmed or broken in the second clause (Binsted et al., 2003).

2.3.3 Phonological features in punning jokes

As previously established, the success of a punning joke depends on the opposition of semantically different but at the same time phonologically similar words. The underlying theoretical framework can be thus denoted as phonosemantics and a neat division between phonological and semantic aspects seems challenging (Hempelmann & Miller, 2017). Nevertheless, phonology and semantics of punning jokes will be addressed separately here, in order to allow for a more systematic approach.

Like many others in the field, Attardo et al. (1994) consider phonetic similarity as crucial for pun processing and appreciation. Since the 1980s, phonological aspects of puns have been systematically investigated. By doing so, the focus of investigation shifted from lexical and semantic aspects towards the more basic abstract phonological building blocks of a word and their manipulation (Guidi, 2017). In that context, phonological distance emerged as a core measure for the sound-related features in punning jokes. The pun and target word in a punning joke can be similar in sound to varying degrees. Attardo et al. (1994) assume the existence of a threshold in phonological distance that, when reached, makes it impossible to understand a joke because pun and target word are too distant in sound. Hausmann (1974) was one of the first to quantitatively analyse the phonology of puns. He investigated French punning jokes based on the number of differing phonemes between pun and target word and found the largest possible distance to be four phonemes (Hempelmann & Miller, 2017).

Vitz and Winkler (1973) developed the concept of *predicted phonetic distance*, a measure of sound similarity between words. In their approach, two words are first aligned by pairing phonemes of one word with either a corresponding phoneme of the other word or a null segment, and then calculating the proportion of phoneme positions that do not match. This results in the so-called Hamming distance as output measure (Hempelmann & Miller, 2017). However, the authors point out that besides phonetic aspects, it is also syllable structure that plays a crucial role for sound similarity judgements.

Lagerquist (1980) investigated the types of change between the phonemes of pun and target words in English puns. Among these changes were transpositions, insertions, deletions, and mutations. Most of the changes were found to involve consonants and most mutations were

focused on a single feature (Lagerquist, 1980). Since only a quarter of the data showed alterations in the initial segment of the word, the author hypothesised that there is a special role to the preservation of the beginning of the sound sequence for the perception of homophony (Hempelmann & Miller, 2017). However, other studies found the opposite result, indicating that in punning jokes frequently the first segment is modified (Hempelmann & Miller, 2017). In any case, the findings by Lagerquist (1980) indicate that during the production of a punning joke, the speaker aims at preserving homophony. This is also supported by the facts that there is only a low percentage of changes in stress pattern between pun and target word in the dataset, and that in most punning jokes the syllable number is maintained. Keeping in mind the goal of making target recovery easy for the listener, it seems reasonable that the applied changes are not major (Lagerquist, 1980).

Zwicky and Zwicky (1986) investigated contrasting sound patterns in English puns. They described a phenomenon called *ousting*, which includes the notion that some segments do not appear equally often in the pun as they do in the target. In more structural terms, X ousts Y when X appears as pun substitute for the latent target Y significantly more often than the other way round (Hempelmann & Miller, 2017). According to this approach, marked features show a tendency to oust unmarked ones. In his book on the phonology of puns, Sobkowiak (1991) quantified sound similarity based on distinctive features. He was able to confirm the notion made by Zwicky and Zwicky (1986) regarding ousting for stops, but not for other features. Sobkowiak (1991) further confirmed the hypothesis made by Lagerquist (1980) that consonants are more likely than vowels to undergo changes or deletions. He further noted that the understandability of a punning joke increases when the consonant structure is kept intact, since vowels are more mutable and carry less information (Hempelmann & Miller, 2017).

Also in line with previous research, Fleischhacker (2005) stated that the pun word needs to be sufficiently similar to the target for the latter to be recovered. Further, the degree of common representation within a corpus correlates positively with goodness of a punning joke, so that more similar pun–target pairs also occur more frequently. However, according to Hempelmann (2003), this can explain a higher rate of target recovery but not greater funniness. Fleischhacker (2005) on the other hand, assumes that punning jokes with a subtle but quickly recognisable phonological relationship between pun and target are also those that are perceived as funnier.

Hempelmann (2003) used a framework grounded in optimality theory in order to assess how much phonological, semantic, and syntactic contrast is possible between pun and target in imperfect punning jokes. The author stresses the point that measuring sound similarity is not enough in order to describe the funniness of a pun, and neither does it directly equal

or relate to semantic similarity. Hempelmann (2003) defines perceived similarity as the least necessary difference between a pair of output forms and points out that it is also an acoustic and not exclusively phonological issue. Regarding the degree of ambiguity, he argues that it should neither be too low nor too high in order for a pun to be perceived as maximally funny. Kawahara and Shinohara (2009) executed a corpus-based analysis of Japanese imperfect puns and found in line with Hempelmann (2003) that besides phonetic features, psychoacoustic features play a role for the perceived funniness of punning jokes. Thus, the perceived similarity between pun and target is based on acoustic information and correlates positively with funniness ratings. This draws attention to the fact that not only objective sound features but also their subjective perception through the listener seems to play a role (Kawahara & Shinohara, 2009).

Jaech et al. (2016) developed a computational model for the recovery of the target word in a punning joke based on the position of the pun word within a sentence context. This model is divided into several stages, beginning with a calculation of the probability of a pun phoneme sequence given a certain target phoneme sequence. This is followed by modelling the pronunciation of each word in the lexicon and finally applying a language model to recover the target word. Rather than calculating phonological similarity between pun and target, Jaech et al. (2016) aimed to model the transformation probability of the two, resulting in a phonetic-edit score for each pun–target pair. They concluded that lower phonetic edit costs correlate with increased goodness of a pun and suggest to use this finding for the creation of pun generation and humour classification programs. Many of the existing systems are limited to perfect puns, while Jaech et al. (2016) trained their model on imperfect puns and tested it on both perfect and imperfect puns in order to create a more varied sample of puns in real-world texts.

2.3.4 Semantic features in punning jokes

Besides phonological information, semantic features of pun and target play a crucial role for the processing of punning jokes, given that they essentially draw from an opposition in meaning between the two.

Several marginally different terms are associated with semantic differences between pun and target. Semantics focuses on concepts, the underlying senses of given words. The term *semantic relatedness* describes any type of connection between two words and includes all types of lexical relationships such as antonymy or association (Budanitsky & Hirst, 2006). *Semantic similarity* denotes a special case of semantic relatedness, where two words are compared in terms of their closeness in meaning. Its inverse is *semantic distance* (Budanitsky & Hirst, 2006).

In computerised natural language processing, semantic concepts are encoded in the form

of nodes in hierarchical networks linked to each other by so-called edges or links, which stand for their semantic connection. In this framework, measures of semantic relatedness are based on properties of the path between two nodes within a network (Budanitsky & Hirst, 2006). There are different approaches on how to create such semantic networks. In dictionary-based approaches, a separate node is created for each word and linked to all the nodes used in its definition. Another approach is based on Roget-structured thesauri, which are a way of structuring words in categories that are further subdivided into loosely defined classes. In this view, words are defined as semantically close when they have a category in common, when their definition contains a pointer to another category, when one is part of the other's category, when both are part of the same subcategory, or when they both have categories pointing to a common category (Budanitsky & Hirst, 2006). Another approach is the use of semantic networks, such as WordNet (Fellbaum, 1998). In WordNet, semantic similarity between two concepts can be calculated through the length of the shortest path between the two. Information-based and integrated approaches on the other hand, use corpora to describe semantic relations between concepts. In such approaches, semantic similarity is described as the extent to which two concepts share common underlying information.

All these measures can be used to investigate semantic relationships between two concepts; however meaning is always to be seen in relation to the surrounding sentence context. In line with this, distributional similarity or co-occurrence similarity describe the degree to which two words tend to occur in a similar context. This context can range from a window of a few words around the word under investigation, to an entire document. Distributional similarity has been found to be a good proxy for semantic relatedness (Budanitsky & Hirst, 2006). As Budanitsky and Hirst (2006) state, "if two concepts are similar or related, it is likely that their role in the world will be similar, so similar things will be said about them, and so the contexts of occurrence of the corresponding words will be similar". However, it needs to be kept in mind that semantic relatedness acts on the concept level and describes symmetric relationships, while distributional similarity focuses on the word level and is potentially asymmetrical. Furthermore, distributional similarity is only relative to a certain corpus, while semantic similarity is more generally dependent on predefined knowledge resources. One problem in this context is also the emergence of so-called ad hoc categories, where lexical semantic relatedness is constructed in context and can therefore not be captured by a priori defined knowledge resources.

The *Graded Salience Hypothesis* states that the meaning of an ambiguous word depends on the word's salience (Giora, 1997). In this framework, salience is defined as function of conventionality, frequency, familiarity, and prototypicality – regardless of context. McHugh and

Buchanan (2016) tested the Graded Salience Hypothesis in the light of lexical co-occurrence semantics as measure of salience in punning jokes. They assumed that punning jokes with high lexical co-occurrence between pun and target word would lead to faster access as compared to low lexical co-occurrence. In their paper, the authors deliberately avoid the use of the words "dominant" and "subordinate" when describing the two ways in which a punning joke can be interpreted, putting emphasis on the fact that both ways are equally probable. The dataset used by McHugh and Buchanan (2016) consisted only of homographic puns. Pun and target word were initially identified and their semantic similarity was measured based on the Windsor Improved Norms of Distance and Similarity of Representations of Semantics (WINDSORS) model (Durda & Buchanan, 2008). WINDSORS is a graph theory–based model of semantic co-occurrence including semantic neighbourhood size and semantic density of word meanings. It is based on calculations of the frequency in which two words occur in the same sentence or a window of ten words, and positively correlates with semantic similarity. In their first behavioural experiment, McHugh and Buchanan (2016) tested whether co-occurrence of context in puns and targets has an influence on successful priming. They found a priming effect for high- and low-occurrence meanings, indicating that meanings are indeed simultaneously activated. In line with the Graded Salience Hypothesis, the authors concluded that first the most salient meaning and then, after a process of reinterpretation, the implied meaning is accessed. For ambiguous words, there is also the possibility of there being a third context, which enables the simultaneous resolution of both meanings.

Mihalcea et al. (2010) aimed to develop and evaluate models for the automatic detection of incongruity in humour using simple one-liners as stimuli. They based their model on joke-related features, as well as knowledge-based and corpus-based semantic relatedness. They hypothesised that the smaller semantic relatedness between set-up and punchline, the funnier the joke because there is a higher level of surprise. Mihalcea et al. (2010) calculated semantic relatedness within the WordNet framework applying six different measures. Additionally, they used domain fitness as measure of semantic relatedness, assuming that the membership of a certain domain gives information about semantic relatedness. However, they found joke-specific features to be a more meaningful indicator of funniness than semantic relatedness. The combination of semantic relatedness and joke-specific features as influencing factors yielded the most meaningful results.

Chapter 3

Data

The dataset for this study is the same as used by Miller et al. (2017) in their SemEval-2017 paper on the detection and interpretation of puns and by Simpson et al. (2019) in their research on humorousness predictions. In total, the dataset consisted of 4030 short texts with an average length of 11 words each. 3398 of them contained humour (520 not in form of puns, but as conventional jokes) and 632 did not contain humour, but consisted of proverbs or aphorisms. Unlike in the original paper, where Miller et al. (2017) used two separate datasets for homographic and heterographic puns, in this study they were merged into a single dataset also including non-pun jokes and non-jokes. All short texts were retrieved from professional humorists and online collections (Miller et al., 2017). The inclusion criteria for the punning jokes were the following:

- maximal one pun per text
- the pun consists of exactly one content word and zero or more non-content words (where content words are defined as nouns, verbs, adjectives, and adverbs)

Homographic puns were defined by Miller et al. (2017) as those texts where pun and target are spelled in the exactly same way – however disregarding inflections and particles, which is a rather liberal definition. Heterographic puns were defined as those where pun and target word are spelled differently – again not taking into consideration inflections and particles. Semantic word sense annotations were carried out independently by three human annotators who chose the corresponding WordNet 3.1, key for pun and target words, respectively. In some cases, more than one sense key was indicated because the meaning of the word was ambiguous or there was more than one option for interpreting it in the given sentence context. In other cases there was no corresponding sense key in WordNet 3.1 available, which was then also noted down.

Funniness ratings were collected using the crowdsourcing platform "Amazon's Mechanical Turk" (https://www.mturk.com/), where so-called click workers are payed for their participation

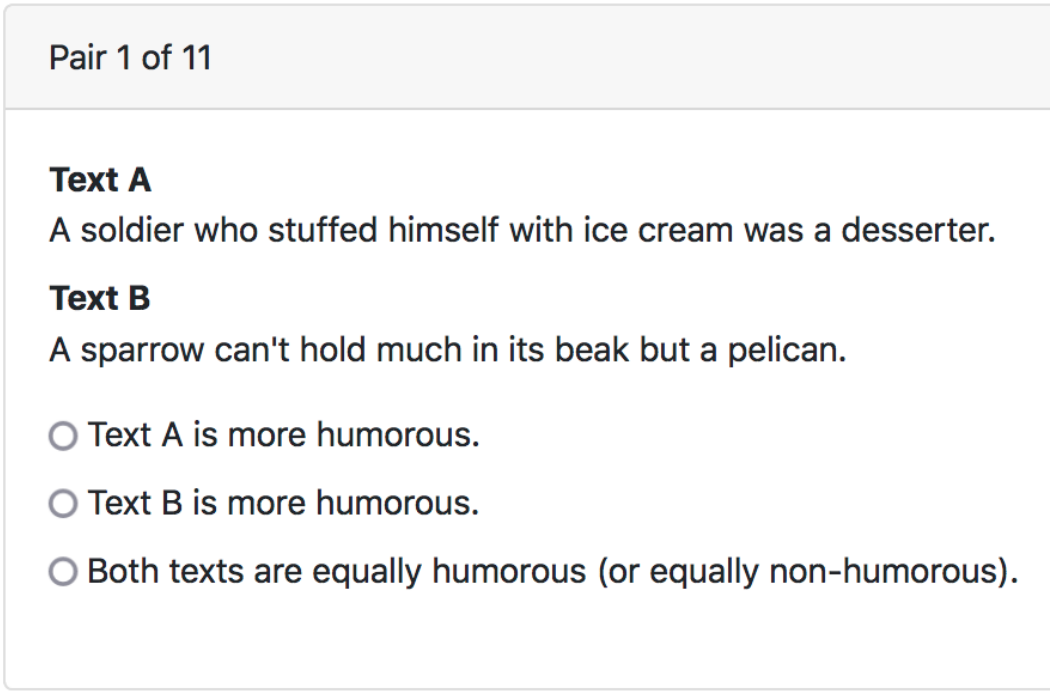
in surveys, content moderation, and other online tasks. In total, 1063 click workers participated in the study and received compensation in line with the US federal minimum wage for their participation (Simpson et al., 2019). Each participant was free to choose how many pairs of jokes they wanted to annotate, ranging from a minimum of 10 to a maximum of 2200 pairs of jokes that were annotated by a single person. According to Buhrmester et al. (2016), data obtained through Amazon's Mechanical Turk can be considered at least as reliable as data gathered using more traditional data collection approaches, such as online surveys or lab studies. One drawback however is the lack of demographic information about the participants, even though Buhrmester et al. (2016) could show that participants recruited through "Amazon's Mechanical Turk" are slightly more demographically diverse than other internet samples. In our case, the only information about the participants that was made available is that they were self-indicated native speakers of English. This lack of information on cultural, linguistic, or socioeconomic background of participants will be addressed in more detail in Section 6.6. Simpson et al. (2019) showed that the funniness ratings for individual pun–target pairs were rather similar across participants, indicated by a rather high inter-annotator agreement (Krippendorff's α of 0.80). Thus, five annotators per pair seems to be an adequate number to reach a consensus ranking between participants.

3.1 Pairwise judgements

During the so-called *Humor Identification Task*, annotators were presented with two puns at a time and were asked to indicate which one of the two they considered funnier. Further, they had also the option to indicate that the presented puns were equally funny or that neither of them was funny. Thus, the raw data that resulted from the data collection for each pair consisted of the indication that either text A or text B or both (or neither) was funnier. For this, the puns, jokes, and non-humorous texts were paired randomly and also across categories, so that each one appeared in 14 unique pairs, resulting in a total of 28,2010 pairs (Simpson et al., 2019). The minimum amount of work a click-worker could submit at once was one experimental run consisting of 11 items. An example item is shown in figure 1. The number of runs varied from participant to participant, as they could freely decide when to stop.

Figure 1

Example item from the Humor Identification Task



3.2 Converting pairwise judgements into ranks

The method of pairwise judgements was used to obtain funniness ratings because it is known to place less cognitive burden on participants and is not affected by biases towards high, low, or middle values and changes in individual rating behaviour over time (Simpson & Gurevych, 2018). Not only does this approach speed up the data collection and prevent biases, but these pairwise labels also allow a total sorting of the text without asking the participants to value them on an overall scale (Simpson et al., 2019). Shahaf et al. (2015) argue as well that when dealing with subjective evaluations, ordinal rating data should not be treated as interval data but rather direct comparisons should be used. In order to be able to draw meaningful conclusions not only about the relative but also the absolute funniness of a certain punning joke, these pairwise judgements need to be transformed into ranks or scores. That is necessary because the interpretation of untransformed rating scales is complicated given that participants are not required to discriminate among all items. Further, the reliability and validity of rating scales are frequently unknown (Flynn & Marley, 2014). Ranks from pairwise comparisons can be calculated assuming a random utility model, where the annotator is assumed to choose an instance (in our case text A or text B or option C) with a certain probability, which is defined as the function

of the utility of that instance (Simpson et al., 2019). When two instances have similar utilities, the probability to chose one of them will be closer to 0.5. When they have different utilities on the other hand, the one with the higher utility is more likely to be chosen. Two approaches to transform pairwise comparisons into ranks based on this random utility model are best–worst scaling (BWS) and Gaussian process preference learning (GPPL). While BWS uses MaxDiff as random utility model, GPPL relies on the Thurstone–Mosteller model (Simpson et al., 2019).

3.3 Best–worst scaling

One approach to make data from rating scales more accessible by transforming it into ranks is best–worst scaling. BWS is used to obtain more information from a respondent looking at the least and the most preferred item of a list. It is based on the maximum difference (MaxDiff) model for best–worst choice. According to Flynn and Marley (2014) there are three use cases for BWS depending on the topic under investigation. For our data, the so-called object case was most appropriate. This approach results in the relative values associated with objects in a list based on their choice frequencies. The score of an instance is thus defined as the fraction of times it was chosen as best minus the fraction of times it was chosen as worst (Simpson et al., 2019). Assuming that the data come from an underlying utility function and have the same variance scale factor, methods based on maximum likelihood or simple counting procedures can be used to create instance utilities in form of MaxDiff scores (Flynn & Marley, 2014).

3.4 Gaussian process preference learning

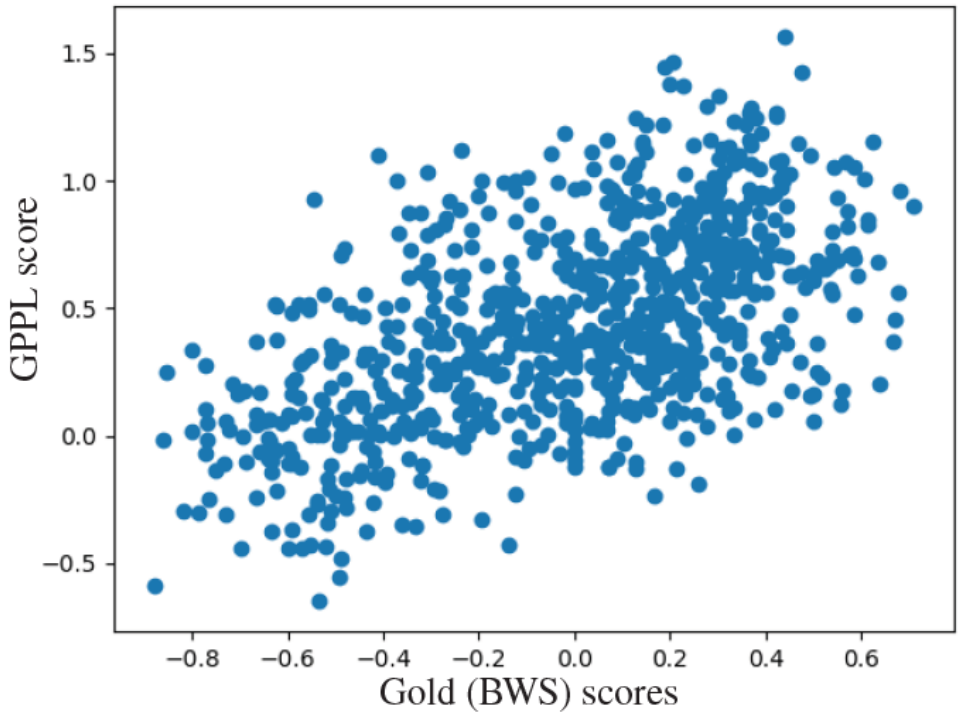
A different method to gain insights from pairwise comparisons is Gaussian process preference learning (GPPL), which is a Bayesian approach employing stochastic variational inference (Simpson & Gurevych, 2018). Bayesian inference methods combine observed data with prior information and subsequently use this information to make predictions about unseen data or latent variables within it. Gaussian processes are distributions over functions of input features (Simpson & Gurevych, 2018). That means that the posterior distribution is estimated over the utilities of an instance given their features (Simpson et al., 2019). At the same time, relationships between instances can be computed based on the covariance between instance utilities as a function of their features (Simpson et al., 2019). GPPL is particularly useful for noisy or small datasets and cold-start scenarios with small amounts of labelled data, and is trained using unsupervised or semi-supervised learning (Simpson & Gurevych, 2018).

In a direct comparison, Simpson et al. (2019) showed that GPPL and BWS yield similar

results when predicting humorousness in the dataset that was also used in our study (see figure 2). However, Simpson et al. (2019) found GPPL to outperform BWS when predicting humorousness and metaphor novelty in a sparse dataset of pairwise comparisons. The GPPL approach achieved a Spearman's r of 0.56 against the BWS gold standard, where word embeddings and linguistic features were used as indicators of humorousness. Given these findings, GPPL seems most appropriate for the correlation analysis of our dataset because of the varying number of responses per annotator as well as the noisiness due to the annotators being able to indicate that neither or both of the jokes are perceived as funny. However, the correlation analysis will be run with ranks resulting from BWS as well as with the ones from GPPL in order to investigate whether there is any difference in results.

Figure 2

BWS vs. GPPL scores for humorousness taken from Simpson et al. (2019)



3.5 Preprocessing

For the analysis, only the data from ratings of the homographic and heterographic puns but not for the non-puns and non-jokes was used. In total, our dataset consisted of 2772 punning jokes, of which 1185 were heterographic and 1587 homographic. The final number of heterographic puns used for the correlation analyses was affected by exclusions during the preparation of the dataset. For the phonological distance analysis, funniness ratings were available for 1131 items,

which were used to calculate correlations. After excluding all proper names and cases where there was no entry in WordNet 3.1 scores of semantic similarity were calculated for a total number of 987 items. These were again used to calculate correlations with funniness ratings.

The transformation values from pairwise judgements to ranks were taken directly from Simpson et al. (2019) and therefore the dataset already included both sets of transformed ranks, using best–worst scaling as well as Gaussian process preference learning. In order for the ranks to be comparable, they had to be normalised and after that were located on a scale from 0 to 1. For the phonological analysis, the pun and target words in their original form as well as in their uninflected form (lemma) were transcribed into the International Phonetic Alphabet (IPA).

Since the funniness ratings were not provided together with the phonological and semantic information, a new spreadsheet was created, connecting funniness ratings with all remaining information. Using this spreadsheet, phonological and semantic distance between pun and target word were calculated and statistical analyses were executed with the goal of finding possible correlations with funniness ratings. The spreadsheet contained the following columns:

- Number
- Text of the punning joke
- Pun word
- Pun word in IPA transcription
- Pun lemma
- Pun lemma in IPA transcription
- WordNet sense key for pun word
- Target word
- Target word in IPA transcription
- Target lemma
- Target lemma in IPA transcription
- WordNet sense key for target word
- Type of pun (homographic or heterographic)
- BWS rank
- GPPL rank

Chapter 4

Methods

4.1 Phonological distance

Linguistic distance is defined as the amount of distinctness between languages, corpora, or individual words. An important role in that context is played by the difference in sound, described by phonology. In order to measure these acoustic differences in a quantitative way, several computerised phonological distance measures have been developed in the past decades (Sanders & Chin, 2009).

4.1.1 Abydos

To calculate phonological distance in the present study, the Python library *abydos* (Little, 2018) was used. Amongst other applications, it is comprised of an exhaustive number of phonetic algorithm and string distance measures. For our calculations, we used the *abydos.distance* package, which implements string distance measures and metric classes. Within this package, we focused on algorithms based on Levenshtein edit distance, which will be described in more detail in the next section. For every character, there are certain phonetic features defined in the source code of abydos which can be used for the distance calculations between single characters or between the strings they make up. These features are based on the ones defined in the International Phonetic Alphabet (IPA). For consonants, these are, for example, place, manner, syllabicness, voicedness, nasality, laterality, retroflex quality, and aspiration. However, abydos does not recognise the whole range of IPA characters; one phoneme (the open-mid back unrounded vowel 'ʌ') was not initially defined. In this case, the source code needed to be adapted manually in order to include this phoneme by defining the corresponding features.

In order to compare strings of characters, the algorithm employed by abydos first aligns the characters of two words, and then calculates the distance between them using a number of

different calculations. Out of the distance measures available in abydos, four were chosen for our analysis. These were the following:

- Levenshtein distance
- Covington's distance
- ALINE distance
- Phonetic edit distance

4.1.2 Levenshtein distance

Levenshtein distance (Levenshtein, 1966) is a measure of edit distance employing the Wagner–Fischer dynamic programming algorithm. It assigns costs for unit insertion, deletion, and substitution. Accordingly, the algorithm calculates the number of each operation needed to convert one string into another string. The sum of the costs of all these operations results in the total distance (Sanders & Chin, 2009). Levenshtein distance is solely based on characters and their relationship to each other, and does not take their specific phonological features into consideration. This makes the measure widely applicable but also less informative in terms of phonology. For this, feature-based substitution methods may prove more useful.

4.1.3 Covington's distance

Covington's distance (Covington, 1996) was initially developed to align words for historical comparisons. It uses a special distance function not based on phonological features, but a categorisation of characters based on a differentiation between consonants, vowels, and glides (Kondrak, 2000). Its base is an evaluation metric consisting of an 8-tuple of costs for each kind of match or mismatch. This means that there is a binary output for each of the following cases depending on whether they are true or not during the comparison of two characters:

- exact consonant or glide match
- exact vowel match
- vowel–vowel length mismatch or i and y or u and w
- vowel–vowel mismatch
- consonant–consonant mismatch
- consonant–vowel mismatch

- skip preceded by a skip
- skip not preceded by a skip

The algorithm aims to find the alignment with the lowest costs associated based on substitution costs and context-dependent indel (insertion or deletion) costs. The values of the distance between two characters range from 0 to 100, where 0 means identical and 100 maximally different. The total alignment cost of two strings is then calculated by summing the costs of all substitutions and indels (Kondrak, 2000). However, since Covington (1996) does not specify phonetic features in depth and only distinguishes between consonants, vowels, and glides, also this method does not fully exploit phonological aspects for measuring linguistic distance.

4.1.4 ALINE distance

The *ALINE alignment and distance algorithm* was developed by Kondrak (2002) and is based on phonetic features and feature salience weights. Here, the relationship of characters is calculated based on their phonetic features as defined in the IPA. The author argues that binary features as used by Covington (1996) and others are not optimal to calculate phonetic alignment and therefore he uses multi-valued features in order to depict relationships between characters in a more naturalistic way (Kondrak, 2002). This results in around twenty distinct features with different amounts of possible values, for example place of articulation, voice, roundness, and others. Kondrak (2002) draws from research by Connolly (1997) and Somers (1998); however, these authors base their works on the assumption that all features are of similar importance. In order to account for differences in feature importance, Kondrak (2002) proposes feature salience weights that vary for individual features depending on their importance for the distance measurements. Taking into account these multi-valued features differentiated by salience weights, the cost function is then calculated based on the costs of an insertion or deletion, substitution, expansion or contraction, and the additional costs resulting from a vowel substitution, expansion, or contraction. Unlike in the previously described algorithms, here the similarity score instead of the distance score is the final result. In his original code, Kondrak did not use the same characters as the ones used in the IPA. In the corresponding abydos package, both the IPA symbols as well as the specific Kondrak symbols are defined in terms of their features in order to enable alignment for both input types.

4.1.5 Phonetic edit distance

As a final distance measure, the so-called Phonetic edit distance was applied. It is a custom calculation included in abydos, created by Little (2018). The algorithm is a variation of Leven-

shtein distance adapted for strings in IPA with the goal to compare individual phonemes based on their feature similarity. Similar to the aforementioned approaches, its cost function is based on insertions, deletions, substitutions, and transpositions. Within the framework of the Phonetic edit distance, the latter have a lower cost value than indels and substitutions. Further, the approach also allows for feature weighting, but unlike ALINE does not use multiple values for one feature.

4.1.6 Comparison of distance measures

The measures described above were chosen for different reasons. Levenshtein distance as the most standard edit distance measure was included in order to have a general view on the data based on similarity between characters only. Covington's distance goes more in depth by looking at the difference between character types. ALINE as well as the Phonetic edit distance are even more detailed since they take into account phonetic features, and were therefore chosen as additional distance metrics. Table 2 summarises noteworthy differences between the four phonological distance measures that were applied in the analysis. Of these, ALINE seems most promising for our purpose. This is due to its use of weighted multi-valued features that allow for the investigation of more fine-grained phonological differences – which we believe to be crucial when analysing punning jokes.

Table 2

Phonological distance measures used in this study

Type	**Costs based on**	**Phonetic features used**
Levenshtein	indel, substitution	none (characters)
Covington	indel, substitution	none (consonants, vowels, glides)
ALINE	indel, expansion, contraction	weighted, multivalued
Phonetic edit distance	indel, substitution, transposition	weighted, 3-valued (absent, present, neutral)

4.1.7 Preprocessing

In order to analyse the phonological features of pun and target word, it was necessary to transcribe them into the International Phonetic Alphabet (IPA), which is the standardised written representation of speech sounds. For the pun and target words and their respective lemmata (the uninflected forms of the word) used by Miller et al. (2017), IPA transcriptions were made avail-

able for the present study by the first author of that paper. The transcriptions were provided in ARPABET format. ARPABET is a set of phonetic transcription codes developed in the 1970s and widely used in speech synthesisers. In ARPABET, phonemes are represented by distinct sequences of ASCII characters. ASCII, the American Standard Code for Information Exchange, is the standard way of character encoding in electronic communication. In this framework, stress is indicated by a digit following vowels: 0 means no stress, 1 means primary stress, 2 secondary stress, and 3 tertiary and further stress. For every ARPABET character, there is a direct counterpart in the IPA. In order to execute phonetic distance calculations, the words in the dataset needed to be converted from ARPABET into standard IPA since abydos works with the latter. Therefore, as a first step, the digits indicating stress were deleted and the individual characters transcribed. For some words, several possible phonetic transcriptions were provided by the original annotator in case of doubt about the correct pronunciation. However, after a number of random inspections it was decided in order to facilitate the analysis to consistently only use the first one of these, which seemed to be correct in all cases.

4.2 Semantic similarity

Besides phonological distance, another way to quantify linguistic distance is by looking at differences or similarities in meaning between words. Semantics is considered a rather complicated field for empirical investigations since the meaning of a word and its underlying concept is in many cases not clearly defined and subject to individual interpretation. Further, single words may also have more than one meaning, which naturally can lead to ambiguities in the analysis process. This imposes big challenges on computerised tools that were developed with the aim to calculate semantic similarity.

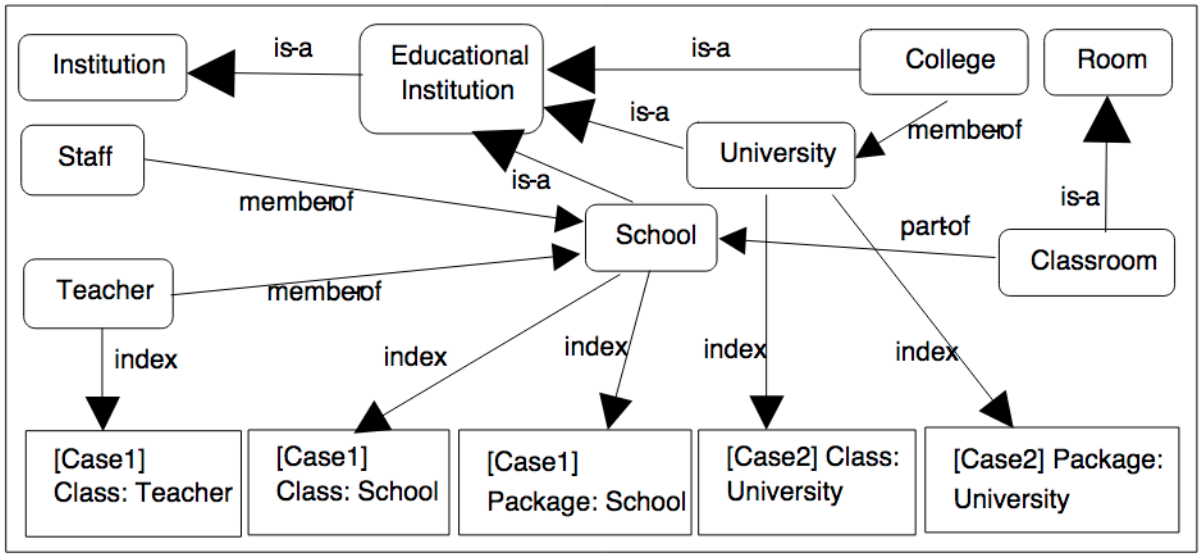
4.2.1 WordNet

WordNet (Fellbaum, 1998) is a lexical database of semantic relations where words of certain parts of speech (in particular nouns, verbs, adjectives, and adverbs) are organised in so-called *synsets* in the form of a network. Each synset represents one underlying lexical concept interlinked with a number of non-hierarchical relations (Budanitsky & Hirst, 2006). Synsets are associated with one or more word senses, which are indexed and include definitions and usage examples of this word. Synsets are connected via different relations in the network, depending also on their part of speech. For example, the noun network consists of eleven unique beginners (e.g. entity, psychological concept) and nine types of relations. These nine relations are the following: hyponymy (IS-A), hypernymy (the inverse of hyponymy), six meronymic (PART-OF)

relations – COMPONENT-OF, MEMBER-OF, SUBSTANCE-OF and their respective inverses – and antonymy (COMPLEMENT-OF). Synonymy is naturally implied in every node. Initially developed for English language, the WordNet framework is now available for a number of different languages. WordNet's current version is WordNet 3.1, the one also used in this research. Figure 3 (taken from Gomes et al., 2003) depicts an example of WordNet's network structure.

Figure 3

WordNet example for the word "school"



In WordNet, sense keys are a form of encoding a word sense. Sense numbers on the other hand, are decimal integers indicating the sense number of a word within the part of speech encoded in the sense key (Fellbaum, 1998). Synset types encode the type of word, which can be noun, verb, adjective, adverb, or adjective satellite. Sense keys operate independently of sense numbers and WordNet versions, and are therefore more suitable for large data analysis projects. According to the WordNet documentation (Fellbaum, 1998), a sense key is represented as *lemma%ss_type:lex_filenum:lex_id:head_word:head_id*. The different components of this general formula can be defined as follows:

lemma is the ASCII text of the word as found in the WordNet database index file

ss_type is a one-digit decimal integer representing the synset type

lex_filenum is a two-digit decimal integer representing the name of the lexicographer file containing the synset

lex_id is a two-digit decimal integer that, when appended onto the lemma, uniquely identifies a sense within a lexicographer file

head_word is only present if the sense is in an adjective satellite synset and if so, consists of the lemma of the first word of the satellite's head synset

head_id is a two-digit decimal integer that, when appended onto head_word, uniquely identifies the sense of head_word within a lexicographer file

As an example, the WordNet entries for the words "die" and "dye" are depicted in figure 4 alongside a calculation of their semantic similarity using a method developed by Wu and Palmer (1994).

Figure 4

Example for WordNet synsets, sense keys, definitions, and semantic similarity

```
# Yesterday I accidentally swallowed some food coloring.
# The doctor says I'm OK, but I feel like I've dyed a little inside.

dye_key="dye%2:30:00::"
die_key="die%2:37:03::"
print("Sense key:",dye_key)
print("Sense key:",die_key)

dye = wn.lemma_from_key(dye_key).synset()
print(dye)

die = wn.lemma_from_key(die_key).synset()
print(die)

print('Definition "dye" :', wn.synset('dye.v.01').definition())
print('Definition "die" :', wn.synset('die.v.05').definition())

print('Semantic similarity:', dye.wup_similarity(die))
```

```
Sense key: dye%2:30:00::
Sense key: die%2:37:03::
Synset('dye.v.01')
Synset('die.v.05')
Definition "dye" : color with dye
Definition "die" : feel indifferent towards
Semantic similarity: 0.2857142857142857
```

4.2.2 NLTK

Several algorithms have been developed to determine semantic similarity based on WordNet entries. One major implementation is the WordNet function of the Python package *Natural Language Toolkit* or NLTK (Bird et al., 2009), an exhaustive package developed for natural language processing in English. Within the *nltk.wordnet* framework, it is possible to choose

from various measures of semantic similarity. Out of these, six were chosen for our analysis. These are the following:

- Path similarity
- Leacock–Chodorow similarity
- Wu–Palmer similarity
- Resnik similarity
- Jiang–Conrath similarity
- Lin similarity

4.2.3 Path similarity

A straightforward way of measuring similarity between concepts is looking at the path length between two nodes in a semantic network graph. Semantic similarity is then defined as the shortest path between the nodes that represent the concepts. Accordingly, path similarity calculates the shortest path that connects senses in an IS-A (hypernym/hyponym) relationship in WordNet (Goodger, 2013). The score for the path similarity is equal to the inverse of the shortest path length (Pedersen et al., 2004) and ranges from 0 to 1, with higher numbers indicating higher similarity. A path similarity of 1 therefore describes identical words.

4.2.4 Leacock–Chodorow similarity

The semantic similarity measure proposed by Leacock and Chodorow (1998) is based on the same shortest path calculation as described above and additionally combines it with the maximal depth of the taxonomy in which the senses occur. This relationship is defined as $-\log(p/2d)$, where p is the shortest path length between two synsets and d denotes the depth of the taxonomy (Goodger, 2013). Scores range from 0 to infinity and the maximum score varies depending on the taxonomy depth.

4.2.5 Wu–Palmer similarity

The Wu–Palmer approach (Wu & Palmer, 1994) measures semantic similarity based on the depth in taxonomy of the two senses plus the depth of their least common subsumer, i.e. their most specific ancestor node (Goodger, 2013). For the Wu–Palmer similarity, scores are necessarily greater than 0 and smaller or equal to 1. This is because the depth of the least common subsumer can never be zero. The score is 1 if the two input concepts are exactly the same.

4.2.6 Resnik similarity

While the previously presented measures are based on distances between nodes in a network, the similarity measure proposed by Resnik (1995) is information-based. It relies on the amount of shared information between concepts, i.e. the information content (IC) of the least common subsumer of two word senses (Budanitsky & Hirst, 2006). For this, it is necessary to implement an information content file that was created using a corpus. Similarity is therefore additionally based on the probability of encountering an instance of a concept in a corpus (Mihalcea et al., 2010). Naturally, results may vary depending on the corpus and corresponding IC file. In this analysis, the Brown Corpus (Francis & Kucera, 1979), a large collection of American English texts from different genres, was used. The scores obtained when calculating Resnik similarity are necessarily greater than or equal to 0. The upper bound depends on the size of the corpus used as IC file.

4.2.7 Jiang–Conrath similarity

Jiang and Conrath (1997) combine network- and corpus-based approaches by primarily relying on an edge- and node-based approach. At the same time, they use corpus statistics for possible corrections (Budanitsky & Hirst, 2006). The Brown Corpus was used for the IC file for this approach as well. Similarly to Resnik similarity measures, scores obtained using the Jiang–Conrath approach depend on the size of the corpus used as IC file and range from 0 to infinity.

4.2.8 Lin similarity

Lin et al. (2013), on the other hand, attempted to create a more universally applicable semantic similarity calculation method, not based on particular resources but solely on commonality and difference between objects. In this framework, semantic similarity is defined as the ratio between the amount of information needed to describe their commonality and the information needed to fully describe the separate concepts (Budanitsky & Hirst, 2006). As with the Resnik and Jiang–Conrath similarity, the Brown Corpus is used here for the IC file. Scores for the Lin similarity range from 0 to 1.

4.2.9 Comparison of distance measures

Budanitsky and Hirst (2006) compared five semantic similarity measures based on WordNet. These were measures proposed by Jiang and Conrath (1997); Leacock and Chodorow (1998); Hirst, St-Onge et al. (1998); Lin et al. (2013); and Resnik (1995). In their comparison, the authors looked at the performance of the approaches in automatic detection and correction

of spelling mistakes. They found that best results were obtained using the Jiang–Conrath similarity, in which path-based calculations are combined with information-based calculations. In terms of performance it is followed by Leacock–Chodorow's and then Resnik's approach. Further, the authors also compared all five measures to human judgements of semantic similarity in two datasets and found that the highest correlations with human annotators were with Jiang and Conrath's approach in one dataset and Leacock and Chodorow's approach in the other (Budanitsky & Hirst, 2006). From the measures used in this study, therefore, the Jiang–Conrath and the Leacock–Chodorow similarities seem the most promising ones.

4.2.10 Word2Vec

The semantic similarity measures described so far are thesaurus-based methods relying on WordNet entries in a classical dictionary sense. However, meaning is not necessarily restricted to those definitions and is strongly affected by the context, in which words evoke certain concepts. Therefore, it seems worthwhile to additionally consider a different approach for assessing semantic similarity. This was done using Word2Vec (Mikolov et al., 2013). Word2Vec is a model, in which words are represented as word embeddings, which means in the form of real number vectors in a multi-dimensional vector space (Jatnika et al., 2019). It is a machine-learning method, where large online corpora (e.g. from Twitter or newspaper articles) are used as training data. Semantic similarity can be calculated in Word2Vec using cosine similarity, where closer words yield higher values on a scale from -1 to 1. Two different architectures can be applied by changing both the vector size and the window size, where the window refers to the number of words around the word under investigation. While the CBOW (continuous bag of words) approach predicts a word based on a given context, skip-gram architectures predict the words around a given word. The Python library spaCy (Honnibal & Montani, 2017), which is used for natural language processing, provides a straightforward approach for calculating semantic similarity based on Word2Vec vectors. It includes a number of already trained pipelines, where part-of-speech tagging, lemmatisation, and named entity recognition are already performed. In this study, the *en_core_web_lg* pipeline was used, which includes English words retrieved from written texts on the internet in form of blogs, news, and comments that have been annotated for vocabulary, syntax, entities, and vectors.

4.2.11 Preprocessing

The dataset as taken over from Miller et al. (2017) contained the sense key identifiers for the target and pun word, as well as for their respective lemmata. Since there was sometimes more

than one sense key indicated, the others were deleted in order to facilitate the analysis procedure, assuming that the first sense key was the most suitable. Further, all entries containing the symbol "U" for "unknown sense key" or "P" for "proper name" were deleted. The remaining sense keys had to be transformed into synsets since the chosen similarity calculations are based on those. This was done using the command *wn.lemma_from_key(key).synset()* from the nltk.wordnet package. For the Resnik, Lin, and Jiang–Conrath similarities, the issue occurred that the part-of-speech tags "a" (adjective), "s" (satellite adjective), and "r" (adverb) were not recognised and therefore words with these annotations were also excluded from the analysis.

For the Word2Vec-based calculation of semantic similarity, pun and target words were used in the form in which they appeared in the original punning joke.

Chapter 5

Results

5.1 Descriptive statistics

Funniness ratings were transformed from pairwise judgements into ranks using both BWS and GPPL. For both approaches, the resulting values were normalised so that they ranged from 0 to 1, with the goal to enable a meaningful comparison. Spearman's rank correlation was computed to assess the relationship between BWS and GPPL. As can be seen in figure 5, there was a strong positive correlation (indicated by the red regression line) between the two approaches ($r = .84$, $p = 0.0$).

Figure 5

Correlation between BWS and GPPL values

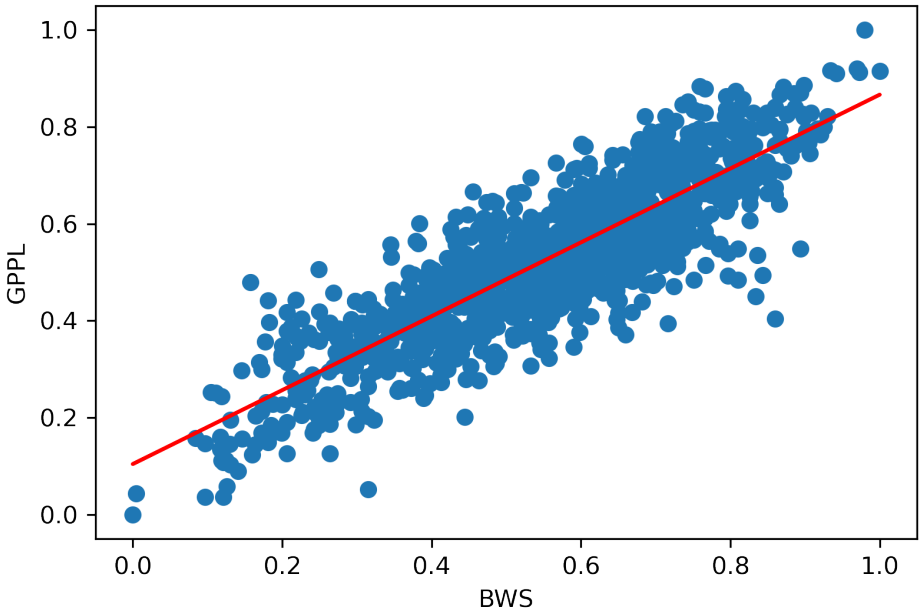
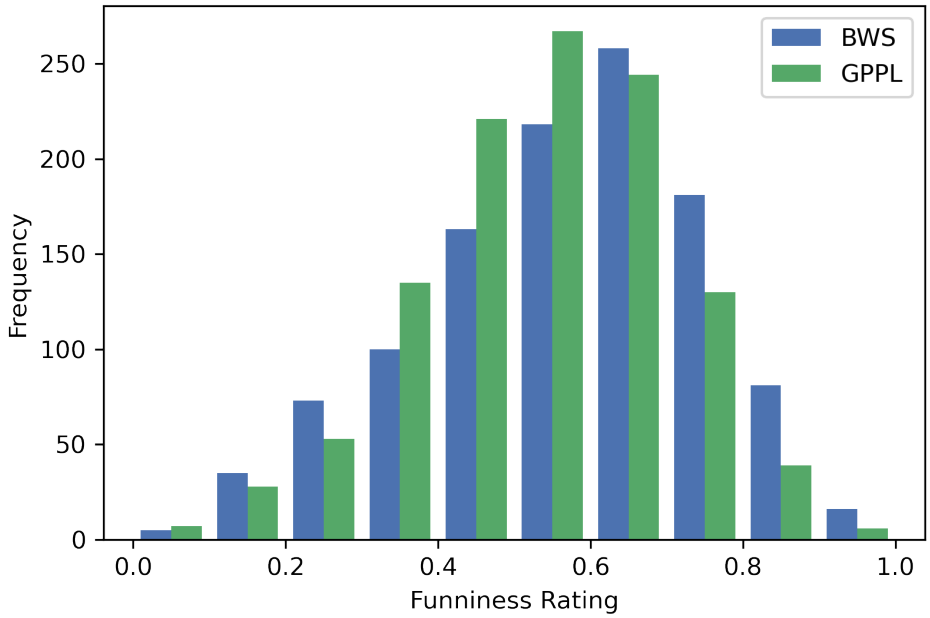


Figure 6 shows the frequency distribution and strength of funniness ratings calculated using BWS and GPPL. Most ratings scored around 0.6, which is slightly above the median. This normal distribution indicates that most punning jokes were rated in a medium range for funniness with only few of them rated as very low or very high in funniness.

Figure 6

Frequency distribution of funniness ratings (normalised and quantised)



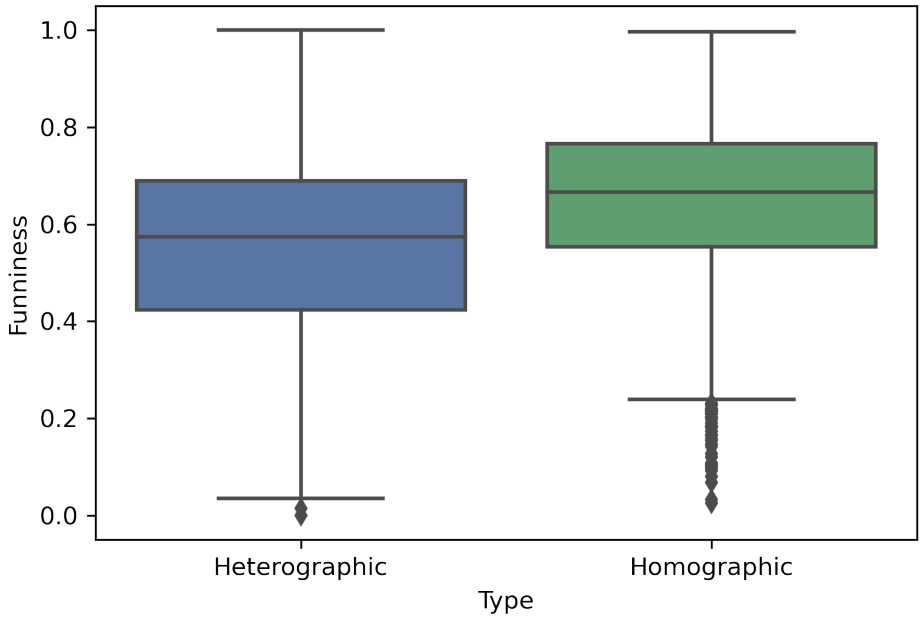
5.2 Phonological distance

In order to assess correlations between phonological distance and perceived funniness, first a group comparison between heterographic and homographic puns was conducted. This comparison followed the assumption that homographic puns are most likely to be also homophonic, while heterographic puns are more likely to be heterophonic. In heterophonic puns, the pun and target words are phonologically further apart. Because of this difference, belonging to one or the other category can be seen as "basic" measure of phonological distance. An independent samples t-test was calculated comparing mean funniness ratings of homographic and heterographic puns. There was a significant difference in the scores for homographic ($M = 0.60$, $SD = 0.14$) and heterographic ($M = 0.52$, $SD = 0.16$) puns ($t(2770) = 13.45$, $p < 0.0001$). The boxplots in figure 7 display a direct comparison between the funniness ratings of the two pun types, indicating that funniness ratings were higher for the homographic than for the heterographic condition. What is striking here is the large variance in funniness ratings in both directions, indicated by the

relatively large whiskers for both pun types. Further, there seems to be a number of outliers in the lower range of funniness ratings, especially for the homographic group.

Figure 7

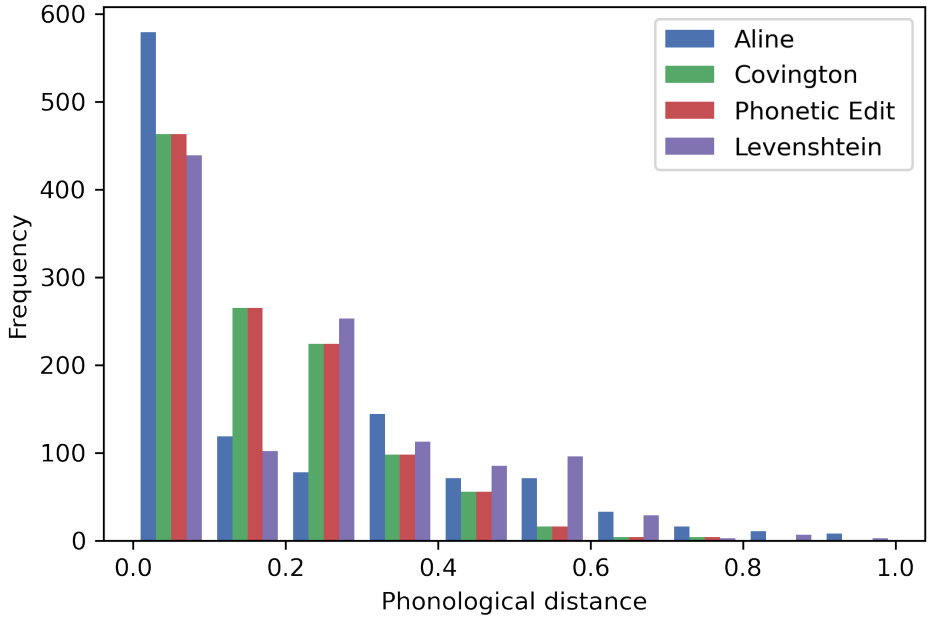
Funniness ratings for heterographic vs. homographic puns



Next, phonological distance measures from the four distance metrics were compared. Further, separate statistical analyses were executed for each of them in order to assess possible correlations with funniness ratings. Figure 8 gives an overview of the frequency and strength of the different phonological distance measures. As can be seen in the plot, the majority of puns under investigation were characterised by a low phonological distance between pun and target word. Even though the results from the different measurements are slightly different, the results are still comparable. In order to assess this, Spearman's rank correlation was computed to assess the relationship between all four phonological distance measures. There was a positive correlation between ALINE and Levenshtein distance ($r = 0.91$, $p = 0.0$), ALINE and Covington distance ($r = 0.83$, $p < 0.0001$), ALINE and Phonetic edit distance ($r = 0.83$, $p < 0.0001$), Levenshtein and Covington distance ($r = 0.94$, p = 0.0), Levenshtein and Phonetic edit distance ($r = 0.94$, $p < 0.0001$), as well as Covington and Phonetic edit distance ($r = 1.0$, $p < 0.0001$).

Figure 8

Comparison of phonological distance measures (normalised and quantised)



After computing and comparing all four distance measures, a more detailed investigation of the correlation between perceived funniness and phonological distance was executed for each of them separately. To do so, Spearman's rank correlation measures were again applied. Correlation analyses resulted in a significant negative correlation for Levenshtein distance and GPPL ($r = -0.089$, $p = 0.021$), Levenshtein distance and BWS ($r = -0.115$, $p = 0.002$), as well as ALINE distance and GPPL ($r = -0.165$, $p = 0.0003$), and ALINE distance and BWS ($r = -0.175$, $p = 0.0001$). The negative correlation indicates that lower phonological distance values were associated with higher funniness ratings. For the other phonological distance measures, there was no significant correlation with funniness ratings.

Figure 9 and figure 10 display the respective scatter plots for the negative correlations (indicated by the red regression line) between ALINE and GPPL funniness ratings, as well as Levenshtein distance and GPPL funniness ratings. Given the high correlation between GPPL and BWS funniness ratings, only the GPPL ones are displayed in the figures.

Figure 9

Negative correlation between ALINE distance and funniness ratings

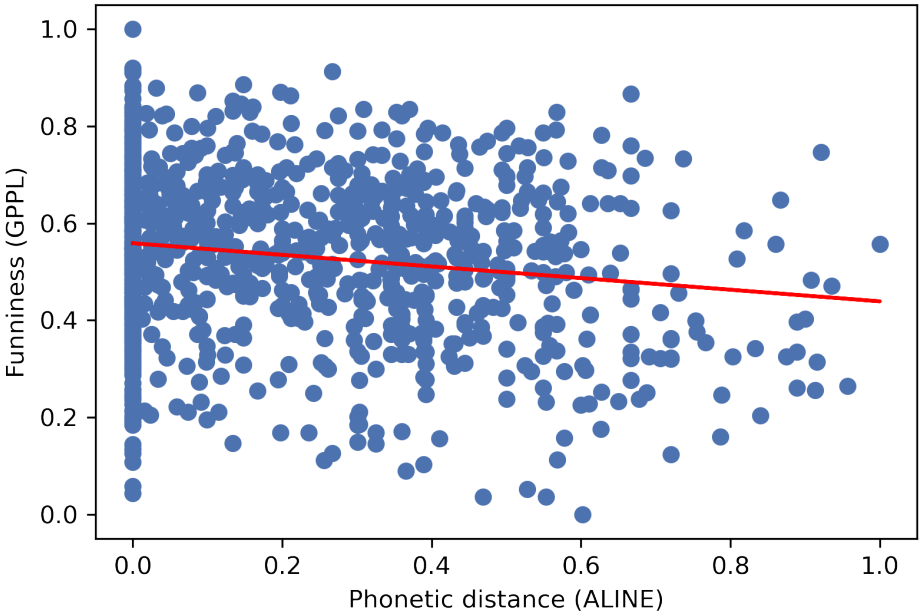
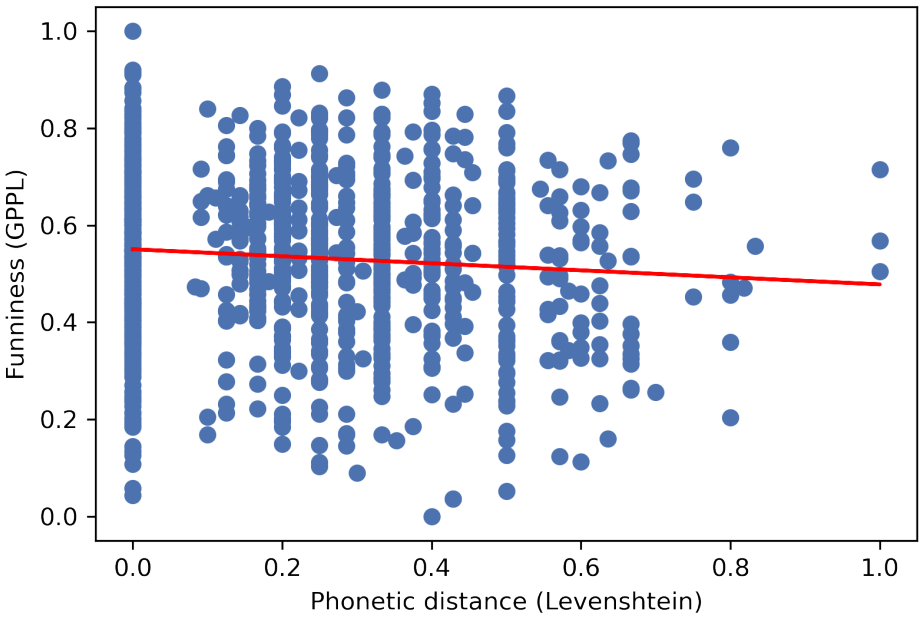


Figure 10

Negative correlation between Levenshtein distance and funniness ratings



5.3 Semantic similarity

With regards to semantic similarity, we first ran Spearman's rank correlations to assess the correlation between different NLTK-based metrics. Both Wu–Palmer similarity and Jiang–

Conrath correlate significantly with all other measures ($p < 0.001$). There was no significant correlation between path similarity and Leacock–Chodorow, Resnik, or Lin similarity. The latter three did however correlate with all other semantic similarity measures ($p < 0.001$). Table 3 displays the correlation between all semantic distance metrics and figure 11 depicts a comparison of the scores derived from the metrics with normalised values in order to allow for a meaningful comparison.

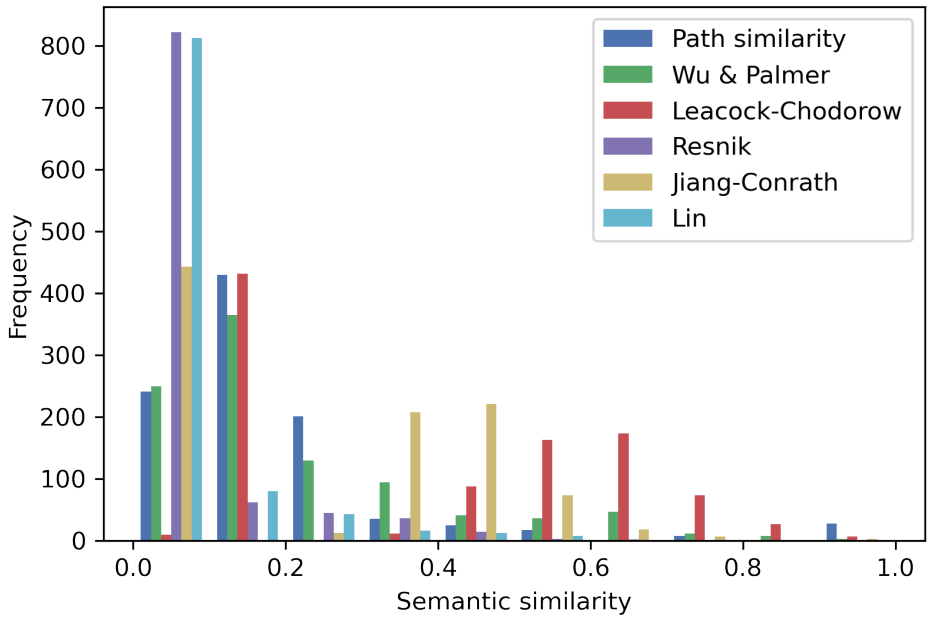
Table 3

Correlation table for semantic distance measures

	Path	Wu & Palmer	Leacock–Chodorow	Resnik	Jiang–Conrath	Lin
Path	1.000	0.614	−0.105	0.006	−0.162	0.007
Wu & Palmer	0.614	1.000	0.352	0.702	0.295	0.095
Leacock–Chodorow	−0.105	0.352	1.000	0.523	0.942	0.158
Resnik	0.005	0.702	0.523	1.000	0.531	0.134
Jiang–Conrath	−0.162	0.295	0.942	0.531	1.000	0.148
Lin	0.007	0.095	0.158	0.134	0.148	1.000

Figure 11

Comparison of semantic similarity measures (normalised and quantised)



As can be seen in figure 11, most NLTK-based semantic similarity measures scored around 0, which indicates low semantic similarity between pun and target word. Furthermore, there is a relatively large variation between the results obtained by the different measures, also indicated through the partly non-significant correlations in table 3. A correlation analysis between all six semantic similarity measures separately and the funniness ratings (GPPL and BWS) yielded no significant results. As for the semantic similarity that was calculated using the cosine distance approach in Word2Vec, there was also no significant correlation with funniness ratings.

Chapter 6

Discussion

6.1 Phonological distance

Statistical analysis resulted in a significant difference in funniness ratings when comparing homographic and heterographic punning jokes. Homographic puns, which naturally showed smaller phonological distance between pun and target word, were associated with higher funniness ratings. The use of homography and heterography as indicative of phonological distance between pun and target word was based on the assumption that homographic puns are in most cases also homophonic, and heterographic puns are mostly also heterophonic (Hempelmann & Miller, 2017). This is of course not the case for all punning jokes – especially when taking into account that the English language is not orthographically transparent. However, in this study the number of punning jokes with overlapping orthographic and phonological features was assumed to be high enough in order to use the pun type as meaningful way of phonological differentiation.

Our findings are in line with previous research and initial hypotheses, which indicated that if pun and target are phonologically closer, then this is associated with higher funniness ratings. In the broader framework of linguistic humour theory, this can be explained with the concept of target recovery. According to theories on the linguistic mechanisms involved in pun processing (e.g. Jaech et al., 2016) the implied target word needs to be recovered by the listener in order to understand the punning joke and perceive it as humorous. Lagerquist (1980) stated that speakers aim to preserve homophony in order to facilitate target recovery. This is because if pun and target are closer in sound, it is naturally easier for the listener to also activate the second meaning, because there is no need for phonological "stretching". What remains is a cognitive exercise on the semantic level in order to evoke the second meaning. Of course, this also holds true for the homographic vs. heterographic case, so that target recovery is easier in homographic rather than heterographic puns – not necessarily because of their phonological

features but simply also because of the identical orthography of pun and target word. In our dataset, homophonic puns were indeed rated as funnier, which can be explained through the fact that target recovery was easier for the raters.

When comparing mean funniness ratings for homographic and heterographic puns, we found a large variance in both pun types. This indicates that there was a broad range of indicated funniness, ranging from almost 0 to almost 1 on a normalised scale. This raises the question of how meaningful this comparison is in the first place given the broad range of funniness ratings. On a more conceptual level, the fact that raters perceived the funniness of the presented punning jokes very differently might indicate that humour cannot be easily measured and quantified in a numerical way – not even when using less straightforward approaches such as pairwise comparisons.

When assessing correlations between phonological distance and perceived funniness for all four metrics separately, we found significant negative correlations for the ALINE distance as well as Levenshtein distance. These findings are in line with our initial hypothesis that lower values for phonological distance are associated with higher values for funniness. A correlation analysis including all four applied metrics indicated a positive correlation between all of them. This raises the question why only the ALINE and Levenshtein distance, but not Covington's and the Phonetic edit distance, were significantly correlated with funniness ratings.

Of all metrics applied, Levenshtein distance is the most basic phonological distance measure. It is based solely on differences between strings of characters and calculates a sum of costs characterised by insertions, deletions, and substitutions needed to convert one string of characters into another. Thus, Levenshtein distance does not take any specific phonological information into account but focuses exclusively on characters. A significant negative correlation of Levenshtein distance with funniness ratings indicates therefore that even without taking into account specific phonological features, the mere orthographic distance between pun and target word is enough to be associated with funniness ratings. This also makes sense when taking into account that our stimuli were presented in a written manner to the participants, and that it was not just phonology but also orthography that played a considerable role. The aforementioned findings from the comparison of homographic and heterographic puns are also in line with the findings regarding Levenshtein distance, since there the differentiation of pun types was also primarily based on orthography.

The ALINE distance, on the other hand, is calculated by taking into account specific phonetic features of sounds as defined by the IPA. It does so in a multi-valued manner and by assigning salience weights to the different features in order to produce more fine-grained distance measures.

Therefore, a significant negative correlation with funniness ratings indicates that the closer pun and target word are in their specific phonological features, the funnier the punning joke is perceived as. This indicates that beyond orthographic similarity, similarity in phonological features of pun and target words does indeed play a role for the perception of funniness of punning jokes.

What still remains an open question regarding phonological distance is why there were no significant correlations for Covington's distance and the Phonetic edit distance. The measurements in Covington's distance are based on a differentiation of characters into consonants, vowels, and glides. Given our non-significant results from the correlation analysis, a distance measure that only takes into account these exact differences between the characters that make up pun and target may not be precise or meaningful enough to produce a differentiation that is correlated with funniness. The Phonetic edit distance, on the other hand, is a variation of Levenshtein distance adapted for IPA symbols and additionally including phonological feature weights. Like Levenshtein distance, it is based on insertions, deletions, and substitutions but includes similar features as the ALINE distance. It is therefore not easily explained how there could be no significant correlation with funniness ratings if both Levenshtein distance, on which the Phonetic edit distance is based, and ALINE distance, which also uses weights for phonological features, resulted in significant correlations. Since the Phonetic edit distance takes its feature weights from the *abydos.phones* package, while the feature weights for the ALINE distance are part of the algorithm itself, this might have led to the difference in results. Further, the ALINE distance but not the Phonetic edit distance uses multi-weighted (and not single-weighted) features, resulting in a more fine-grained phonetic differentiation. That difference might have also played a role for the contrasting results of the two metrics.

6.2 Semantic similarity

Statistical analyses did not yield any significant correlations between semantic similarity measures and funniness ratings. There are several possible explanations for the absence of a significant correlation. The most straightforward explanation is that there is indeed no meaningful connection between semantic similarity of pun and target word in a punning joke and its perceived funniness. Given that there was no strong directed hypothesis in the first place and that this part of the analysis was more exploratory, this is very well possible. One possible hypothesis regarding semantic similarity was that the highest funniness ratings would be found neither for cases where the pun and target are too semantically close nor too semantically distant, but rather where semantic distance is located in the medium range. That was based on the assump-

tion that if pun and target word are too close in meaning, no ambiguity arises and the sentence cannot be classified as a (punning) joke in the first place – and thus also cannot be rated as humorous. If, however, pun and target word are too far apart in meaning, the second possible way of interpreting the sentence might be too unrealistic and would require too much of a semantic "stretch" in order for the joke to be perceived as humorous. Given that a successful punning joke requires both interpretations to be more or less acceptable, a nonsensical interpretation would only result in so called zero-meaning puns or groaners (Dynel, 2010). Those have been found to be perceived as less funny and partly also support the reputation of punning jokes as a lower form of humour. If indeed medium semantic similarity is associated with the highest funniness ratings, this would not result in a linear but a quadratic relationship. However, we did not find any significant relationship between semantic similarity and perceived funniness, so a possible quadratic relationship remains to be investigated by further studies.

Another reason for the absence of significant correlations could be that most of the semantic similarity metrics employed resulted in very low semantic similarity in the first place. As can be seen in figure 11 in the results section, most of the semantic similarity measures scored around 0. This absence of a meaningful distribution of semantic similarity may not be due to a failure of the metrics themselves but rather due to the fact that most words were indeed not semantically similar at all. Given this, it seems almost impossible to calculate correlation analyses since the semantic similarity was not distributed enough in order to be informative or meaningful.

Lastly, the low scores in semantic similarity and the subsequent absence of correlations with funniness ratings might also be explained through the complexity of quantifying semantic relationships in general. The metrics used in this study, are – even though their working mechanisms are slightly different – all based on the assumption that it is possible to denote a word or concept with a string of numbers coding for its meaning. One could argue that this is not a valid approach when assessing and depicting meaning and the semantic relationship between words or concepts. Instead, many different factors could lead to subtle differences in interpretation, such as cultural knowledge, personal experiences, language proficiency, or literacy. Thus, the fact that there was no significant correlation with funniness ratings might simply indicate that the metrics applied in this study were not able to depict semantic similarity in a meaningful way due to the complexity of the task.

This critical approach raises the more general question of how meaningful it is to look at semantic similarity of single words or concepts separately from the context they occur in. While this is in general possible, it could be argued that for the present endeavour it would have been more useful to look at the semantic fit of both the pun and the target word with the syntactic

context of the punning joke. It was initially hypothesised that individual semantic definitions would lead to a similar effect and that there would be no difference in the semantic closeness of pun and target word in relation to all other words in the syntactic context. This is because it was assumed that ultimately the semantic definition of the individual pun and target word would suffice as information since the surrounding words do not change. But when taking into consideration that meaning is created not by assessing words separately but always in context, it might have still been worthwhile to additionally look at the semantic context.

Besides the measures based on WordNet, we used Word2Vec as an alternative approach to calculate semantic similarity. The values obtained from this approach derive from a machine-learning approach which uses information from large corpora and takes into consideration the context of occurrence when denoting the meaning of words. Yet the values obtained from this more elaborate approach were not significantly correlated with funniness ratings either. In line with the previously proposed interpretations, this indicates that it might be a matter of there not being meaningful semantic relations rather than the metrics failing to capture them.

6.3 Results in light of humour theory

In line with the description of universal humour aspects provided by Guidi (2017), the phenomenon of punning is a rather prototypical and therefore more generalisable form of humour expression. That is firstly because like most jokes, puns are based on ambiguity as a conceptual feature. Secondly, they are verbally expressed, which is the only aspect of jokes that is present in all cultures (Guidi, 2017). We found lower phonological distance to be associated with higher funniness ratings. This can be explained in a broader framework of humour theory, which in turn can be used to describe and explain the perceived funniness of punning jokes in more detail.

The humorous connotation of punning jokes can be described best by incongruity resolution theories. These assume a two-step process, where an incongruous element is first detected and then linked in a sense-making way to the context in order to resolve ambiguity. Punning jokes can be described in this framework as instances of bisociation (Koestler, 1964) since they can be read in two different ways. While processing a punning joke, the pun word is first detected and classified as incongruous because a second meaning is evoked through phonological similarity to the target word. Only through a lexical cue in the punning joke can the target word be retrieved and the second meaning of the punning joke made clear. This resolution of incongruity is assumed to lead to the emergence of humour in punning jokes.

In line with the view on release theories proposed by Roberts (2017), the resolution of incongruity can be seen as form of successful error detection leading to a feeling of reward and

humour. Thus, punning jokes can be described using both incongruity resolution as well as humour relief theories. Regarding superiority theories, the punning jokes used in this study do not necessarily fall under this categorisation since most of them do not come at the expense of a person or group.

The Semantic-script Theory of Verbal Humour (Raskin, 1985) as a form of linguistic incongruity theory can be used to describe punning jokes. In this framework, they represent a prototypical example for a sentence with two overlapping scripts, i.e. ways of reading and interpreting. These two scripts are opposed to each other in the sense that only one possible way of interpretation can be activated at a time. There is furthermore a lexical trigger present that requires the reader to rethink the initial script and activates the second possible interpretation. This process can be explained best using an example from the dataset:
When those around King Arthur's table had insomnia, there were a lot of sleepless knights.
In this homophonic punning joke, the pun word is "knights" while the target word is "nights". The target word is activated through the lexical script-switch trigger "insomnia", which activates the additional meaning of night – the time where sleeplessness occurs.

In terms of categorisation, the punning jokes used as stimuli in this study were initially differentiated according to their orthography since the stimuli were presented in written form. The dataset was therefore divided into homographic and heterographic puns. However, there was no indication on whether this differentiation also translated to them being homophonic and heterophonic. Even though there is a possibility for homographic puns being heterophonic and heterographic puns being homophonic, it can be assumed that that was not the case for the majority of stimuli. When characterised according to the source of ambiguity, the punning jokes in this study can be mainly defined as lexical-semantic puns based on polysemantic homonyms (Giorgadze, 2014). Further, structural-semantic puns based on idiomatic expressions were also present. Regarding the situational context, all of the puns used here were so-called self-contained puns, in which humour does not directly relate to the context but more on general knowledge (Ritchie, 2005). Since the sentence context was not further taken into account in this research, no differentiation based on the relationship of pun word and sentence context can be made. Accordingly, no differentiation based on the relationship between pun word and script-switch trigger was made. Further, a categorisation based on the prevalent interpretation and involvement of cognitive effort can also not be made here, since the focus of the data collection and analysis lay on more basic linguistic aspects.

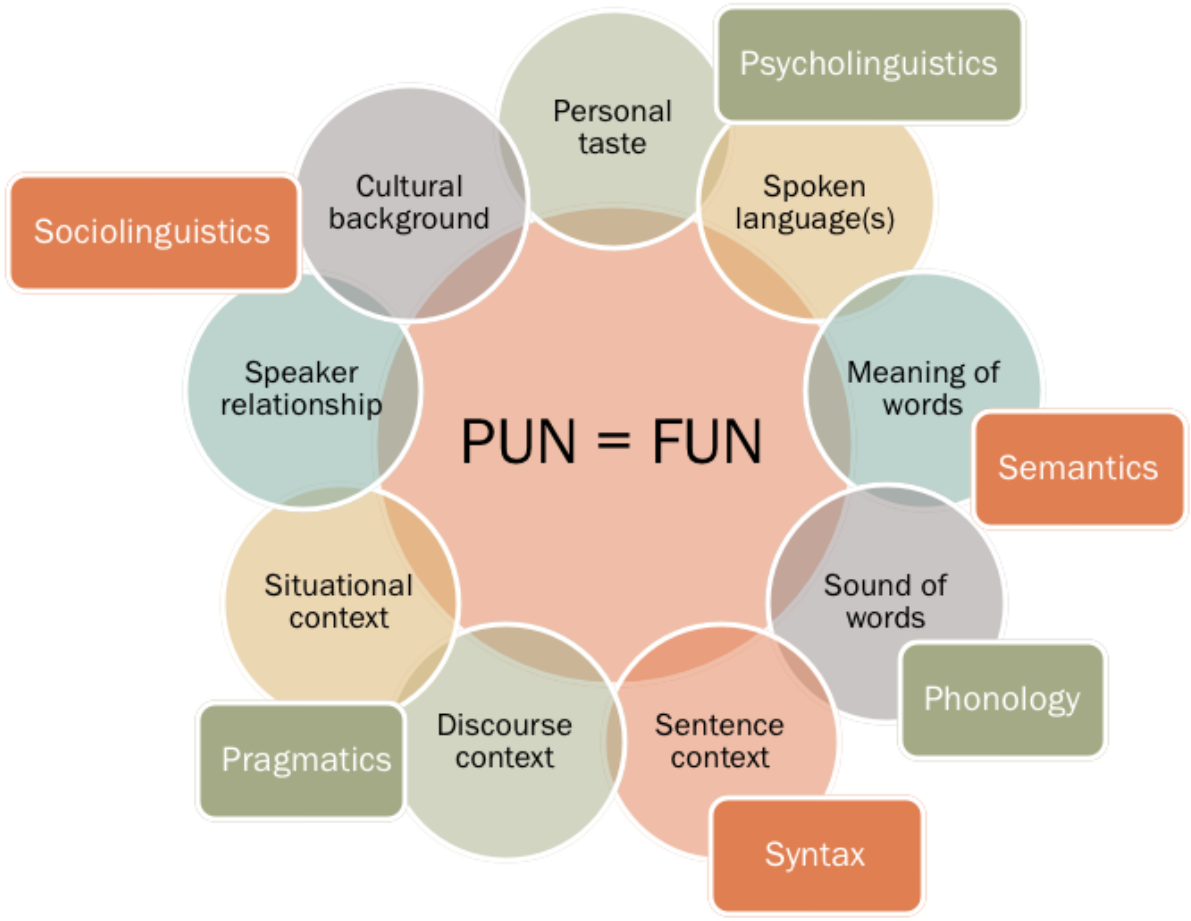
6.4 The bigger picture

Even though this study focused on phonology and semantics, it needs to be kept in mind that many other factors might have played a role for the perception of funniness of the punning jokes used as stimuli in this study (for an overview of other possible influencing factors see Smith et al., 2020). However, it is still a worthwhile approach to focus first and foremost on linguistic aspects when analysing punning jokes, given that their humorousness is based on the manipulation of language.

According to Brône et al. (2006), researching the phenomenon of punning jokes makes it necessary to give up on the artificially created categories in theoretical linguistics, which draw clear boundaries between syntax, semantics, and pragmatics. In their view, meaning itself depends fundamentally on the activation of knowledge shared by speaker and listener in a certain situational, cultural, and linguistic context. On a larger level, meaning creation can be seen as interplay of linguistic, social, and cultural factors. This makes language in general and humour in particular a dynamic and interactive experience. Accordingly, the investigation of phenomena in this field necessarily requires a multidimensional view. Therefore, many different aspects that might play a role in the perception of funniness should ideally be taken into consideration when investigating the emergence of humorousness in association with punning jokes. These are assumed to be related to various sub-disciplines in linguistics and should be taken into account at least from a theoretical perspective. Figure 12 gives an overview on factors that are hypothesised to be of relevance when assessing the funniness of a punning joke, alongside the sub-fields in linguistics they are situated in. These factors include more individual aspects, such as languages spoken by the listener and the level of proficiency in those. But personal taste, which may be influenced by personality factors or cultural background determining exposure to certain types of humour, may also play a role here. Furthermore, the relationship between speaker and listener is crucial in a communicative situation, which in turn is connected to the broader situational and discourse context. Lastly, besides phonology and semantics, other linguistic factors, such as sentence context or morphology, may also play an important role.

Figure 12

Hypothesised factors playing a role for the perception of funniness of punning jokes



Using the framework of the General Theory of Verbal Humour (Attardo & Raskin, 1991), punning jokes can be described in a more exhaustive way by taking into consideration additional factors from a broader field that have not been addressed so far. Based on the differentiation into knowledge resources, mechanisms involved in the processing of punning jokes can be described. Further, it is theoretically possible to use the GTVH to directly compare any two simultaneously presented punning jokes used as stimuli in this study.

The language resource is directly employed through the empirical analysis executed in this study, since it covers phonological and lexical descriptions of words involved in the punning joke. Further, it includes statistical information about the occurrence of certain linguistic units – in this case phonological distance and semantic similarity.

The language structure can be used to describe where in the punning joke the pun and target words as well as the script-switch trigger are located. In this context, it would be interesting to investigate at which position in a sentence the pun word is located. Previous studies found the punchline in a joke to be frequently located at the pre-final or final position

(Attardo & Raskin, 1991). Miller et al. (2017) investigated whether this also holds true for punning jokes. Their results indicate that the last word of the investigated punning jokes was the pun approximately half the time – more precisely in 47% of the cases for homographic puns and 57% of the cases for heterographic puns. Future studies could for example look at the position of the script-switch trigger separately and in relation to the pun word. According to McHugh and Buchanan (2016), the meaning of an ambiguous homograph is generally determined by the surrounding words of the sentence it is embedded in, which serve as semantic cues. For punning jokes, however, there is no single way of interpreting a pun word since it is ambiguous and both meanings could theoretically be appropriate in the given syntactic context. For such cases, a number of models on the processing of words with multiple meanings have been proposed. According to *selective access models*, the context exclusively determines the selection of the appropriate meaning (McHugh & Buchanan, 2016). According to *exhaustive models* on the other hand, all meanings are initially accessed regardless of context and salience. As combination of the two, *ordered-access models* are based on the idea that all meanings are initially accessible but meanings with higher frequency are accessed faster. *Re-ordered access models* view the interpretation of words as function of frequency and context (McHugh & Buchanan, 2016). The Graded Salience Hypothesis proposes that contextually more salient information or meaning are always accessed first (Giora, 1997). In order to investigate under which category the processing of punning jokes falls, it would be most appropriate to use psycho- or neurolinguistic approaches involving eye tracking, functional magnetic resonance imaging (fMRI), or electroencephalography (EEG). These methods allow for conclusions to be drawn about the neural correlates of certain cognitive mechanisms and thereby shed a light on underlying processes.

In terms of genre, the punning jokes used in this study can be classified as so-called canned jokes. Such jokes do not emerge spontaneously in a given context but their way of presentation or the situation they are used in indicate their humorous nature. Regarding this, it would be interesting for future research to compare funniness ratings of canned punning jokes to ratings of jokes embedded in a situational context. Embedded punning jokes could be presented for example in the form of short video clips taken from sitcoms or comedic movies. Further, it would also be interesting to compare funniness ratings for canned punning jokes and canned jokes that are non-puns.

Another knowledge resource is the target – the person, topic, or group the joke is directed to. As previously mentioned, it can be assumed that the majority of the punning jokes used in this study did not have a direct target. Punning jokes in general are very rarely used to make fun of or show superiority towards a certain group. This also underlines the fact that the presence

of this knowledge resource is not obligatory.

The situation resource describes the semantic background and features necessary for a joke to be successful. The calculation of semantic similarity to an extent falls under this knowledge resource; however, further investigations could focus on the situation in more detail.

The knowledge resource of logical mechanism, on the other hand, is not suitable to compare punning jokes presented since it is related to their nature on a more basic level. The main logical mechanism in play in punning jokes is referred to as cratylism and for this the reader needs to apply willing suspension of disbelief. Only when the reader accepts and applies cratylism as logical mechanism can the punning joke be successful. This makes the logical mechanism of cratylism one of the most defining knowledge resources for punning jokes.

Giorgadze (2014) points out that puns always fulfil a pragmatic role too. One aspect of this is the establishment of a speaker–listener relationship using humour as both a way to establish a hierarchy and also to create a bond between all or certain actors in a communicative situation. In this context, social and cultural expectations play a role too, so that humorous utterances in general but also jokes on specific topics may be more or less appropriate in a certain situation and therefore potentially elicit surprise. Smith et al. (2020) investigated how characteristics of a social situation and its actors influence the perception of funniness of puns. Unsurprisingly, they found that puns were perceived as funnier in playful rather than serious situations and perceived as more appropriate when they occurred at the end of a conversation rather than when they were interrupting it. Further, they found that perceived funniness was higher when puns were told by children rather than adults and when age and expertise in the topic of the pun varied across adults (Smith et al., 2020). This study highlights the importance of the situational context and speaker–listener relationship for the success of a punning joke. Further, the success of a joke also depends on whether speaker and listener share the same background. This can refer to contextual or experiential but also cultural background.

Puns are often considered a “low” form of humour and several attempts have been made to explain this. One possible explanation is that puns are relatively easy to produce and do not require the speaker or listener to be particularly witty. Another explanation is that incongruity in punning jokes is based in most cases on phonological and not semantic relations between words, which is connected to the idea that less cognitive effort is needed for phonological word-play (Hempelmann & Miller, 2017). On a contrary note, puns have also been described to be exceptionally clever in the sense that they are appreciated more for their ingenuity and the involved manipulation of language rather than the humour itself (Smith et al., 2020).

Furthermore, it is to be assumed that individual traits, such as personality, autobiographical

history, or language proficiency also play a role in the perception of funniness of punning jokes. These factors can influence personal taste in humour and are for example also associated with upbringing and exposure to certain types of humour. But more basic personality factors as well as literacy or language aptitude may also play a role here. Additionally, the number and types of languages spoken as well as the proficiency in those languages may influence the perception and appreciation of linguistic humour.

Finally, language and culture are necessarily intertwined, with language being an open as well as subtle expression of cultural knowledge and norms. Accordingly, culture determines to a certain extent what type and targets of humorous expression are appropriate. This in turn is again influenced by more basic linguistic features, for example with a language structure that allows better or worse for wordplay.

6.5 Natural language processing and linguistic humour

To date, creative language (e.g. irony, metaphorical speech, or humour) still poses an obstacle for natural language processing. This includes the recognition or detection, the interpretation, and the production of creative language. Accordingly, computerised humour generation is so far mainly based on templates derived from common humour theories (Maraev et al., 2020) and is not very elaborate. More success has been reached in the recognition and classification of linguistic humour.

If successfully managed, the automatic processing of verbal humour can be applied in human–computer interaction scenarios (Simpson et al., 2019). One of the main goals in the field is to make conversational AI more human-like, for example in order to be able to create social robots or chat bots (Maraev et al., 2020). Interactions between humans and computers, such as in chat bots, are suggested to be more realistic when they involve humour as a component (Miller et al., 2017). This is both in the sense that the machine understands when a humorous utterance has been made and is able to reply in an appropriate manner, but also so that it can spontaneously create such utterances. In turn, this is thought to enhance user satisfaction during the interaction with a chat bot, which may have positive effects on user efficiency (Miller et al., 2017). Binsted et al. (2003) argue that the use of puns by conversational second language learning software motivates the learner since the conversational agent is perceived as a worthy dialogue partner and less patronising or intimidating. At the same time, deeper understanding of important cultural concepts – for example in form of idioms or proverbs – and phonological rules are transmitted to the learner. However, the quality of humour plays a determining role for the user's perception of the computer agent (Binsted et al., 2003). Computerised processing of verbal humour can

potentially also be applied in machine translation (Simpson et al., 2019). For example automatic translation of puns could solve the issue that it is sometimes very hard for non-native speakers to detect verbal humour in text (Miller et al., 2017). The creation of ambiguity-preserving translations of verbal humour is a complicated task (Miller et al., 2017) and automatic processes could facilitate translation of sitcoms and other forms of comedic movies. Finally, humour or joke detection algorithms could potentially also be used in the digital humanities to detect jokes in literary texts and catalogue jokes created by a specific author (Binsted et al., 2003).

Simpson et al. (2019) argue that humorous and figurative language, due to their non-literal character, require complex linguistic as well as general knowledge for processing. The first step in the processing of such chunks of text is their recognition. Several attempts have been made to solve this task, resulting in algorithms able to identify humour in text. Computerised humour recognition can happen in the form of the detection of salient linguistic and humour-specific features, but also n-gram patterns, latent semantic structures, or humour anchors (Maraev et al., 2020). Maraev et al. (2020) argue that humour in dialogue can be deconstructed into principles of reasoning, so-called topoi. Conversational analyses together with corpus studies and assumptions based on incongruity theory can shed light on the emergence of humour in different communicative interactions. In their research, the authors focus on humour detection through mining of topoi from data, rather than detecting lexical incongruities by using distributional semantic networks. They argue that reasoning in dialogues is *enthymematic* and defeasible, meaning that the listener needs to draw from contextual or background knowledge – the aforementioned topoi – in order to understand the argument of a conversation. Most importantly, the person telling the joke can guide the listener in a certain direction, revealing a less salient topos and thus transmitting the humorous aspect of an utterance. In their analysis, Maraev et al. (2020) followed several steps starting with extraction of enthymeme candidates in a linguistic joke based on their surface structure. This was followed by an annotation of whether the candidate is an enthymeme and if so, its classification. Finally, a semantic representation was created through enthymeme parsing. By identifying enthymemes and topoi, the authors provide an automated strategy to identify humorous utterances.

The two main mechanisms associated with the computational processing of punning jokes are generation and detection of puns (Hempelmann & Miller, 2017). However, existing pun generators are so far only able to produce homographic (and coincidentally homophonic) puns, without taking into consideration specific phonological features. Similarly, computational pun detection systems rely mainly on syntactic cues or can only detect homographic puns (Hempelmann & Miller, 2017). According to Miller et al. (2017), the underlying mechanism in play

is word sense disambiguation, the identification of the meaning of a word in context. This is however based on the assumption that the word in focus has a single, unambiguous meaning. As previously established, this is not the case in puns, which crucially rely on ambiguity and the activation of phonological and semantic knowledge.

In their research, Miller et al. (2017) evaluated the computerised detection, location, and interpretation of punning jokes. In the pun detection task, the participating systems were required to classify all contexts in a dataset and decide whether they contained a pun or not. During pun location, the task was to decide for all the contexts that had been classified as containing a pun, which exact word the pun was. During pun interpretation, systems were provided with a context containing a pun word and its exact location, and were asked to return the two WordNet senses evoked by it. Ten systems participated in the pun detection, location, and interpretation tasks. While most systems performed well on the pun detection subtask, only few of the participating systems were better than baseline in the pun location subtask and unsurprisingly, pun interpretation was the biggest challenge for all participating systems. Miller et al. (2017) note that even though there is theoretical research on the phonological and semantic features involved in punning jokes, findings from this research have not yet been used in applied systems. This would be a worthwhile development in order to advance the computerised detection, location, and interpretation of puns in given contexts. One step in this direction was done by Hempelmann (2003), who developed a computational model to describe phonological features involved in punning jokes.

6.6 Limitations and further directions

Several factors might have limited this investigation and influenced its results. One problem regarding the data collection process is the lack of demographic information about raters. The only requirement to take part in the study was that the first language of participants was English, which was indicated by them beforehand. This low threshold to take part in the study facilitated the collection of large amounts of data but came at the expense of not gathering basic demographic information, such as age, sex, or country of origin. Besides that, and in line with the previously established additional influencing factors, it would have also been interesting to assess the raters' spoken languages, their respective proficiency in these and their personal preference in humour. Future studies could therefore aim to assess possible correlations between funniness ratings and those rater-specific and demographic factors. In our case, such investigations would have only been worthwhile if the ranks obtained through BWS and GPPL had been calculated for each participant separately. Since this was not done here, it would be impossible to connect

funniness ratings and demographic information, which in some way justifies the lack of the latter in the data collection. Another approach for finding possible alternative factors that play a role in funniness ratings would be to ask participants directly to justify their funniness ratings. In such a more qualitative approach, participants could be asked after trials to indicate why they defined one and not the other punning joke as the funnier one. In the present study, focusing more on individual participant-related factors would have not allowed us to collect large amounts of data in relatively short time and uncomplicated manner.

Another limitation regarding the study design is the way in which phonological features are encoded in the metrics used for the calculation of phonological distance. Hempelmann and Miller (2017) argue that a general problem with feature-based metrics is the fact that sounds perceived as similar by human raters often still differ in a disproportionately large amount of features. Thus, human raters may capture subtle differences between phonological features in a different way than feature-based metrics do. One major reason for that seems to be that standard distinctive features are based on articulation rather than acoustics, and thus focus on the creation rather than the perception of a sound. Hempelmann (2003) stressed the point that it is not only phonological features but also acoustic that ones play a role in the perception of phonetic distance. The perception of sound similarity as a psychoacoustic phenomenon (Hempelmann & Miller, 2017) makes the metrics applied in the present study not fully adequate to model sound similarity since they rely solely on phonetic features. This discrepancy between phonological differences based on acoustic perception and those based on pre-defined features for sound production is an aspect that should be taken into consideration when interpreting the results of this study. However, in the present study design it would have been impossible to account for this in a better way, since written puns were used as stimuli. Even though it may seem counter-intuitive to measure sound differences without actually displaying sounds, the applied metrics have nonetheless proven to be a valid measure of phonological distance.

Further, by presenting the punning jokes in our study in a written and not spoken manner, participants were provided with clearer stimuli and thus it was possible to ensure that they would be in any case able to grasp the punning nature of the sentences. Therefore, differences between homographic and heterographic puns might have been more pronounced and clear-cut as compared to an experimental setting in which punning jokes are presented in the form of an auditory stimulus. The latter allows for more freedom in acoustic interpretation, while differences in orthography necessarily underline the heterographic character of pun and target.

Another aspect to consider is that the punning jokes in this study all fall under the category of so-called canned jokes (Dynel, 2010). These are a type of joke that functions outside of a

conversational context because it provides all the necessary information regarding its humorous nature directly within the sentence. Additionally, the contextual set up – in our case the fact that punning jokes were presented in an experimental study, which required funniness ratings – already prepared the participants for the humorous nature of the stimuli. Therefore, the way in which jokes were presented in our study is not fully translatable to a spontaneously emerging humorous situation. Because of that, participants were supposably not required to actively enter a state of non–bona fide communication (Raskin, 1985) during joke processing but did so in the first place because they were made aware of the humorousness of the situation.

In general, it would be worthwhile to take a more all-encompassing view on the question of what makes a pun funny. In this study, we focused on word-related aspects – more precisely the sound and meaning of pun and target word in a punning joke – and investigated their possible correlation with funniness ratings. To account for a broader range of possible influencing factors in a quantitative or even quantitative manner would have required a much more elaborate experimental design and a more extensive data collection procedure and is something which could be tackled by further studies. Nevertheless, to focus on and manipulate purely linguistic aspects within punning jokes rather than the situation or the rater allows us to investigate a rather stable aspect from the bigger framework of the perceived funniness of punning jokes. Even though no overarching and exhaustive view could be given, this study still contributes to understanding the phenomenon and its components in more detail.

Chapter 7

Conclusion

It was the aim of this thesis to investigate different factors associated with the perception of funniness of punning jokes. The focus hereby lay in the assessment of the relationship between, on the one hand, the phonological distance or semantic similarity between pun and target word, and funniness ratings on the other hand. This was done using a dataset of homographic and heterographic punning jokes rated for funniness. Statistical analyses revealed a negative correlation between phonological distance and funniness of punning jokes, indicating that punning jokes where pun and target are closer in sound are associated with a higher perception of humorousness. There were no significant correlations between semantic similarity of pun and target word and funniness ratings. On the one hand, this may be due to the fact that pun and target word were indeed in most cases not similar enough in order for a effect to be captured. On the other hand, this finding may indicate a need for more fine-grained methods for the assessment of semantic relationships.

In a broader humour-theoretic view, punning is based on linguistic ambiguity and is characterised by resolution of incongruity during the processing of a joke, which leads to the emergence of humorous appreciation. The successful recovery of the target word is a requirement for incongruity resolution. Lower phonological distance between pun and target is hypothesised to facilitate target recovery, which in turn may explain higher funniness ratings. Of course it needs to be kept in mind that only a correlation analysis was executed here, and therefore only limited assumptions can be made regarding the direct causal influence of phonological distance of pun and target on the perception of funniness of punning jokes.

Additionally, a broad range of other features has been hypothesised to play a role for funniness ratings of punning jokes. Amongst these are cultural background and world knowledge, personal taste in humour and other person-related factors, and language-specific features such as proficiency or literacy. Those factors were addressed in a non-exhaustive way in the discus-

sion section. As a result, a model of factors involved in humorous processes associated with punning jokes was proposed, acknowledging the multi-layered character of cognitive phenomena in general, and punning as a form of humorous wordplay in particular.

Language and humour are higher cognitive processes and strongly connected to social communication. By doing research in these areas, new insights about the nature of human communication, interaction, and cognition can be gained. This study therefore contributes to the field by empirically underlining the association of phonological closeness of pun and target word and the perception of funniness of a punning joke.

Bibliography

Aarons, D. (2017). Puns and tacit linguistic knowledge. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 80–94). Routledge. https://doi.org/10.4324/9781315731162-7

Attardo, S. (2014). *Encyclopedia of humor studies*. Sage Publications. https://doi.org/10.1515/humor-2015-0100

Attardo, S. (2017). The general theory of verbal humor. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 126–142). Routledge. https://doi.org/10.4324/9781315731162-10

Attardo, S., Attardo, D. H., Baltes, P. & Petray, M. J. (1994). The linear organization of jokes: Analysis of two thousand texts. *Humor*, 27–54. https://doi.org/10.1515/humr.1994.7.1.27

Attardo, S. & Raskin, V. (1991). Script theory revis (it) ed: Joke similarity and joke representation model. *Humor*, 293–348. https://doi.org/10.1515/humr.1991.4.3-4.293

Attardo, S. & Raskin, V. (2017). Linguistics and humor theory. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 49–63). Routledge. https://doi.org/10.4324/9781315731162-5

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. McGraw-Hill Book Company. https://doi.org/10.1037/11164-000

Binsted, K., Bergen, B. & McKay, J. (2003). Pun and non-pun humour in second-language learning. *Workshop Proceedings, CHI*.

Bird, S., Klein, E. & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Brône, G., Feyaerts, K. & Veale, T. (2006). Introduction: Cognitive linguistic approaches to humor. *Humor*, 203–228. https://doi.org/10.1515/HUMOR.2006.012

Budanitsky, A. & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational linguistics*, *32*(1), 13–47. https://doi.org/10.1162/coli.2006.32.1.13

Buhrmester, M., Kwang, T. & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? In A. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 133–139). American Psychological Association. https://doi.org/10.1037/14805-009

Connolly, J. H. (1997). Quantifying target–realization differences. Part I: Segments. *Clinical linguistics & phonetics*, *11*(4), 267–287. https://doi.org/10.3109/02699209708985195

Covington, M. A. (1996). An algorithm to align words for historical comparison. *Computational linguistics*, *22*(4), 481–496.

Durda, K. & Buchanan, L. (2008). WINDSOR: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, *40*(3), 705–712. https://doi.org/10.3758/BRM.40.3.705

Dynel, M. (2010). How do puns bear relevance? *Relevance studies in Poland*, *3*, 105–124.

Fellbaum, C. (1998). A semantic network of English verbs. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 153–178). The MIT Press. https://doi.org/10.7551/mitpress/7287.003.0008

Fleischhacker, H. A. (2005). *Similarity in phonology: Evidence from reduplication and loan adaptation* (Doctoral dissertation). University of California, Los Angeles.

Flynn, T. N. & Marley, A. A. (2014). Best-worst scaling: Theory and methods. In S. Hess & A. Daly (Eds.), *Handbook of choice modelling* (pp. 178–201). Edward Elgar Publishing. https://doi.org/10.4337/9781781003152.00014

Francis, W. N. & Kucera, H. (1979). Brown corpus manual [Online; accessed 01-01-2022]. http://korpus.uib.no/icame/brown/bcm.html

Freud, S. (1961). Humour. *The standard edition of the complete psychological works of Sigmund Freud, Volume XXI (1927-1931): The future of an illusion, civilization and its discontents, and other works* (pp. 159–166). London: Hogarth Press.

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, 183–206. https://doi.org/10.1515/cogl.1997.8.3.183

Giorgadze, M. (2014). Linguistic features of pun, its typology and classification. *European Scientific Journal*, 271–275.

Gomes, P., Pereira, F. C., Paiva, P., Seco, N., Carreiro, P., Ferreira, J. L. & Bento, C. (2003). Management and reuse of software design knowledge using a CBR approach. *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI'03 Workshop*.

Goodger, D. (2013). Wordnet interface [Online; accessed 01-01-2022]. https://www.nltk.org/howto/wordnet.html

Guidi, A. (2017). Humor universals. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 17–33). Routledge. https://doi.org/10.4324/9781315731162-3

Hausmann, F. J. (1974). *Studien zu einer Linguistik des Wortspiels*. Max Niemeyer Verlag. https://doi.org/10.1515/9783111710747

Hempelmann, C. F. (2003). *Paronomasic puns: Target recoverability towards automatic generation* (Doctoral dissertation). Purdue University.

Hempelmann, C. F. (2017). Key terms in the field of humor. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 34–48). Routledge. https://doi.org/10.4324/9781315731162-4

Hempelmann, C. F. & Attardo, S. (2011). Resolutions and their incongruities: Further thoughts on logical mechanisms. *Humor*, 125–149. https://doi.org/10.1515/HUMR.2011.008

Hempelmann, C. F. & Miller, T. (2017). Puns: Taxonomy and phonology. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 95–108). Taylor & Francis. https://doi.org/10.4324/9781315731162-8

Hirst, G., St-Onge, D. et al. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 305–332). The MIT Press. https://doi.org/10.7551/mitpress/7287.003.0020

Honnibal, M. & Montani, I. (2017). spaCy 2: Natural language understanding with bloom embeddings. *Convolutional neural networks and incremental parsing*, *7*(1), 411–420.

Jaech, A., Koncel-Kedziorski, R. & Ostendorf, M. (2016). Phonological pun-derstanding. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 654–663. https://doi.org/10.18653/v1/N16-1079

Jatnika, D., Bijaksana, M. A. & Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in English words. *Procedia Computer Science*, *157*, 160–167. https://doi.org/10.1016/j.procs.2019.08.153

Jiang, J. J. & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of ROCLING X*. https://arxiv.org/abs/cmp-lg/9709008v1

Kao, J. T., Levy, R. & Goodman, N. D. (2016). A computational model of linguistic humor in puns. *Cognitive science*, *40*(5), 1270–1285. https://doi.org/10.1111/cogs.12269

Kawahara, S. & Shinohara, K. (2009). The role of psychoacoustic similarity in Japanese puns: A corpus study. *Journal of linguistics*, *45*(1), 111–138. https://doi.org/10.1017/S0022226708005537

Koestler, A. (1964). *The act of creation: A study of the conscious and unconscious processes of humor, scientific discovery and art*. Dell Book.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 288–295.

Kondrak, G. (2002). *Algorithms for language reconstruction* (Doctoral dissertation). University of Toronto, Toronto.

Lagerquist, L. M. (1980). Linguistic evidence from paronomasia. *Papers from the Sixteenth Regional Meeting. Chicago Linguistic Society*, (16), 185–191.

Larkin-Galiñanes, C. (2017). An overview of humor theory. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 4–16). Routledge. https://doi.org/10.4324/9781315731162-2

Leacock, C. & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 265–283). The MIT Press. https://doi.org/10.7551/mitpress/7287.003.0018

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, *10*(8), 707–710.

Lin, Y.-S., Jiang, J.-Y. & Lee, S.-J. (2013). A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering*, *26*(7), 1575–1590. https://doi.org/10.1109/TKDE.2013.19

Little, C. C. (2018). Abydos documentation [Online; accessed 01-02-2022]. https://abydos.readthedocs.io/en/latest/

Maraev, V., Breitholtz, E. & Howes, C. (2020). How do you make an AI get the joke? Here's what I found on the web. *First AISB Symposium on Conversational AI (SoCAI)*.

McHugh, T. & Buchanan, L. (2016). Pun processing from a psycholinguistic perspective: Introducing the model of psycholinguistic hemispheric incongruity laughter (M. PHIL). *Laterality: Asymmetries of Body, Brain and Cognition*, *21*(4-6), 455–483. https://doi.org/10.1080/1357650X.2016.1146292

Mihalcea, R., Strapparava, C. & Pulman, S. (2010). Computational models for incongruity detection in humour. *International Conference on Intelligent Text Processing and Computational Linguistics*, 364–374. https://doi.org/10.1007/978-3-642-12116-6_30

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv e-prints*, *1301.3781*. https://arxiv.org/abs/1301.3781

Miller, T., Hempelmann, C. F. & Gurevych, I. (2017). Semeval-2017 task 7: Detection and interpretation of English puns. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 58–68. https://doi.org/10.18653/v1/S17-2005

Pedersen, T., Patwardhan, S., Michelizzi, J. et al. (2004). WordNet:: Similarity-measuring the relatedness of concepts. *Demonstration Papers at HLT-NAACL 2004*, 38–41. https://doi.org/10.3115/1614025.1614037

Raskin, V. (1985). Semantic theory of humor. *Semantic mechanisms of humor* (pp. 99–147). Springer. https://doi.org/10.1007/978-94-009-6472-3

Raskin, V. (2017). Script-based semantic and ontological semantic theories of humor. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 109–125). Routledge. https://doi.org/10.4324/9781315731162-9

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. https://arxiv.org/abs/cmp-lg/9511007

Ritchie, G. (2004). *The linguistic analysis of jokes*. Routledge. https://doi.org/10.4324/9780203406953

Ritchie, G. (2005). Computational mechanisms for pun generation. In G. Wilcock, K. Jokinen, C. Mellish & E. Reiter (Eds.), *Proceedings of the tenth european workshop on natural language generation (ENLG-05)* (pp. 125–132). ACL Anthology.

Ritchie, G. (2018). *The comprehension of jokes: A cognitive science framework*. Routledge. https://doi.org/10.4324/9781351232753

Roberts, A. (2017). Funny punny logic. *dialectica*, *71*(4), 531–539. https://doi.org/10.1111/1746-8361.12200

Sanders, N. C. & Chin, S. B. (2009). Phonological distance measures. *Journal of quantitative linguistics*, *16*(1), 96–114. https://doi.org/10.1080/09296170802514138

Shahaf, D., Horvitz, E. & Mankoff, R. (2015). Inside jokes: Identifying humorous cartoon captions. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1065–1074). Association for Computing Machinery. https://doi.org/10.1145/2783258.2783388

Simpson, E., Do Dinh, E.-L., Miller, T. & Gurevych, I. (2019). Predicting humorousness and metaphor novelty with Gaussian process preference learning. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5716–5728. https://doi.org/10.18653/v1/P19-1572

Simpson, E. & Gurevych, I. (2018). Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, *6*, 357–371. https://doi.org/10.1162/tacl_a_00026

Smith, R. H., Hoogland, C. E. & Brown, E. G. (2020). Once a pun a time: Exploring factors associated with perceptions of humorous punning. *Humor*, *33*(1), 7–28. https://doi.org/10.1515/humor-2018-0058

Sobkowiak, W. (1991). *Metaphonology of English paronomasic puns*. Lang.

Somers, H. (1998). Similarity metrics for aligning children's articulation data. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, 1227–1232.

Uekermann, J., Daum, I. & Channon, S. (2007). Toward a cognitive and social neuroscience of humor processing. *Social Cognition*, *25*(4), 553–572. https://doi.org/10.1521/soco.2007.25.4.553

Vitz, P. C. & Winkler, B. S. (1973). Predicting the judged "similarity of sound" of English words. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 373–388. https://doi.org/10.1016/S0022-5371(73)80016-7

Wilson, D. & Sperber, D. (2002). Relevance theory. In G. Ward & L. Horn (Eds.), *The handbook of pragmatics* (pp. 606–632). Blackwell. https://doi.org/10.1002/9780470756959.ch27

Wu, Z. & Palmer, M. (1994). Verb semantics and lexical selection. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, 133–138. https://doi.org/10.3115/981732.981751

Yus, F. (2003). Humor and the search for relevance. *Journal of pragmatics*, *35*(9), 1295–1331. https://doi.org/10.1016/S0378-2166(02)00179-0

Zwicky, A. M. & Zwicky, E. D. (1986). Imperfect puns, markedness, and phonological similarity: With fronds like these, who needs anemones. *Folia Linguistica*, *20*(3-4), 493–503. https://doi.org/10.1515/flin.1986.20.3-4.493

Appendix A: Abstract

Punning jokes are a form of humorous wordplay based on semantic ambiguity between two phonologically similar words – the pun and the target – in a sentence context where both meanings are more or less acceptable.

Previous research attempted to quantify and compare phonological features of pun and target, looking at correlations with acceptability and understandability. Additionally, semantic features are to be considered when examining the success and humorousness of a punning joke. It was the goal of this study to quantify phonological and semantic distance between pun and target words, and assess possible correlations with funniness ratings of the respective punning joke. Statistical analyses revealed a significant negative correlation between phonological distance and perceived funniness for two of the four phonological distance measures applied. This is in line with previous phonological analyses of puns which found lower phonological distance between pun and target to be associated with higher humorousness. None of the seven semantic distance measures applied showed significant correlations with funniness ratings, which leaves space for a number of interpretations.

However, other factors such as situational context or cultural norms may also influence the perception of funniness of punning jokes. Further studies should attempt to take these additional aspects into account, by collecting detailed demographic data or strictly controlling for possible confounding variables during assessment of funniness ratings.

Keywords: humour studies, linguistics of humour, computational linguistics, psycholinguistics

Appendix B: Zusammenfassung

Kalauer oder „punning jokes“ sind eine Form von humoristischem Wortspiel oder Wortwitz. Der humoristische Charakter wird dabei durch semantische Ambiguität zwischen zwei phonologisch ähnlichen Wörtern hervorgerufen. Diese beiden Wörter - „pun“ und „target“ - müssen dabei in einen syntaktischen Kontext eingebettet sein, welcher beide Lesarten zulässt.

In bisherigen Studien wurden phonologische Aspekte von „pun“ und „target“ quantifiziert und verglichen und mögliche Korrelationen mit Verständlichkeit von Kalauern erhoben. Zusätzlich spielen auch semantische Aspekte eine Rolle für das Gelingen von Kalauern. Das Ziel dieser Arbeit war es, phonologische und semantische Distanz zwischen „pun“ und „target“ zu berechnen und mögliche Korrelationen mit Lustigkeitebewertungen der jeweiligen Kalauer zu erfassen. Für zwei der vier angewandten Messansätze ergaben statistische Analysen eine signifikante negative Korrelation zwischen phonologischer Distanz und Lustigkeit. Dies bestätigt frühere Forschungsergebnisse, welche ergaben, dass geringere phonologische Distanz zwischen „pun“ und „target“ mit höherer Lustigkeit einhergeht. Keine der sieben Messmethoden zur semantischen Distanz ergab signifikante Korrelationen mit Lustigkeit, was verschiedene Interpretationen zulässt.

Neben phonologischen und semantischen Aspekten spielen auch andere Faktoren wie situativer Kontext und kulturelle Normen eine Rolle bei Lustigkeitsbewertungen von Kalauern. Zukünftige Studien sollten demnach auch jene Aspekte miteinbeziehen indem beispielsweise ausführlich demographische Daten erhoben oder andere Einflussvariablen während der Datenerhebung kontrolliert ausgeschlossen werden.

Appendix C: List of figures

Appendix D: List of tables