Front Page Information:
Article Omission in Italian Headlines
in light of Information Density

Sjoerd Eilander

Supervisor:

Prof. Dr. Sergey Avrutin

Second reader:

Dr. Denis Paperno

Utrecht University

Research Master Linguistics

Utrecht University

The Netherlands

August 9th, 2020

Front Page Information: Article omission in Italian Headlines in light of Information Density

Sjoerd Eilander

Abstract

When confronted with multiple options to convey the same meaning, speakers tend to choose for the option that distributes information, as measured by surprisal, as evenly as possible over a sentence (Uniform Information Density Hypothesis, or UIDH). In this paper, it is investigated if the UIDH can be used to explain patterns of article omission in Italian newspaper headlines. To study this question, a corpus study has been conducted. As overt articles lower the surprisal of the following noun, it was expected that article omission would occur more frequently before low surprisal nouns. However, while previous research for Dutch and German found results that were in line with the UIDH, the current study found the opposite effect. The UIDH, therefore, cannot be used to account for article omission in Italian newspaper headlines.

Acknowledgments

First and foremost, I would like to thank Sergey Avrutin for his ideas, guidance and supervision. I would also like to thank Denis Paperno for his willingness to help me out with coding problems.

In addition, I would like to thank my friends from "the basement", both on- and offline, in particular Romy van Drie and Rosita van Tuijl, for their assistance with practical problems but primarily for helping me stay on track during an incredibly weird time to write a master's thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The production of language in regularly developed adults involves both knowledge stores and various processes: a thought is conceptualized, formulated with aid from the lexicon, and articulated (Levelt, 1989). Jaeger (2010) hypothesized that language is structured to be efficient at all of these levels of linguistic representation. Based on the idea that the information of a word is measured by its probability in context (Shannon, 1948), Jaeger proposed a principle of efficient language production, the Uniform Information Density Hypothesis (UIDH). This hypothesis proposes that speakers have a desire to distribute information as uniformly as possible over a sentence. Information is measured by the entropy, or surprisal, of a word, which is the probability of a word's occurrence in its context. The reasoning behind this notion is that if a word is unlikely to occur in a specific context, it has a high surprisal, and a high information load.

Jaeger tested this hypothesis by trying to explain omission of the complementizer "that", such as in (1), with it.

(1) I know (that) you want to go to the cinema.

He found that if the surprisal of the onset of the complement clause was high, speakers were more likely to use the complementizer "that", as it distributes the information load over more words. In other words, the UIDH was proven to be a significant predictor of the omission of complementizers.

If choice points in regular speech, where speakers have the option to omit or include words from their utterances, are governed by a tendency towards uniform information distribution, it is interesting to see if omission in deviant speech can be explained by the

same principle. Two notable examples of omission in irregular speech are omissions in the speech of developing children, see (2), from CHILDES (MacWhinney, 2000), and omission in texts in a special register, see (3) (The Guardian, 2019). Special registers, such as newspaper headlines, diary style writing and telegram style writing are characterized by the ellipsis of syntactic elements that are required in other contexts (Oosterhof & Rawoens, 2017).

(2) Daar komt trein. (Niek, 3;00;09)
Here comes train.
'Here comes a/the train.'

(3) Mathieu van der Poel takes final stage and Tour of Britain glory

Interestingly, De Lange (2008) notes that there are crosslinguistic similarities between the omission of articles in child speech and the omission of articles in newspaper headlines: Dutch children omit articles up until a later age than their Italian peers, and Dutch newspaper headlines contain more article omissions than Italian ones. De Lange (2008) argues that another principle based on information theoretic foundations lies behind this similarity: the entropy of the set, which reflects the overall complexity of an article system.

Lemke, Horch, & Reich (2017) agree that information theoretic constraints are relevant in explaining article omission in newspaper headlines. However, they look at individual articles, rather than set entropy, and, based on Horch & Reich's (2016) observation that articles lower the surprisal of the following head noun, they argue that articles in headlines may be yet another category of choice points that is governed by the UIDH. Indeed, omission in German headlines seemed to follow the predictions made by the UIDH: high surprisal nouns were more frequently preceded by an article than low surprisal nouns. This effect was also found for Dutch headlines (Van Tuijl & Paperno, 2020).

However, Lemke et al. (2017) do not explain all patterns of article omission in headlines with the UIDH. For one, as De Lange (2008) notes, there is asymmetry in headlines: omission is more frequent at the sentence initial position than at a sentence internal position. UIDH also fails to provide an account as to why omission is so much more frequent in Dutch headlines than in Italian headlines.

For this reason, this paper is devoted to the study of newspaper headlines in Italian. A corpus study was used to test the hypothesis that the Uniform Information Density

principle governs the omission of articles in Italian newspaper headlines. The Italian article paradigm is an interesting one to look at in light of information density. First of all, De Lange (2008) showed that the set entropy of the Italian article paradigm is lower than the set entropy of the Dutch and German article paradigms. In addition, Italian articles are in part determined by the morphology of the word that succeeds it (Caramazza et al., 2001). As such, there are multiple articles fulfilling the same function. For example, both "i" and "gli" may be used to express a masculine definite plural.

The layout of this paper is as follows. Chapter 2 is devoted to the theoretical background and related work. Starting from a model for the production of language, I will study the UIDH and its foundations. I will also take a closer look at choice points in language, in particular the omission of articles in newspaper headlines, and how these might be explained by information theoretical notions. Then, I will lay out the foundation of the current research. In chapter 3, I will explain the methodology of the corpus study that was used to investigate the research question. In chapter 4, I will present the results of this study, which are discussed in chapter 5. Finally, I will arrive at a conclusion in chapter 6.

Chapter 2

Theoretical Background and Related Work

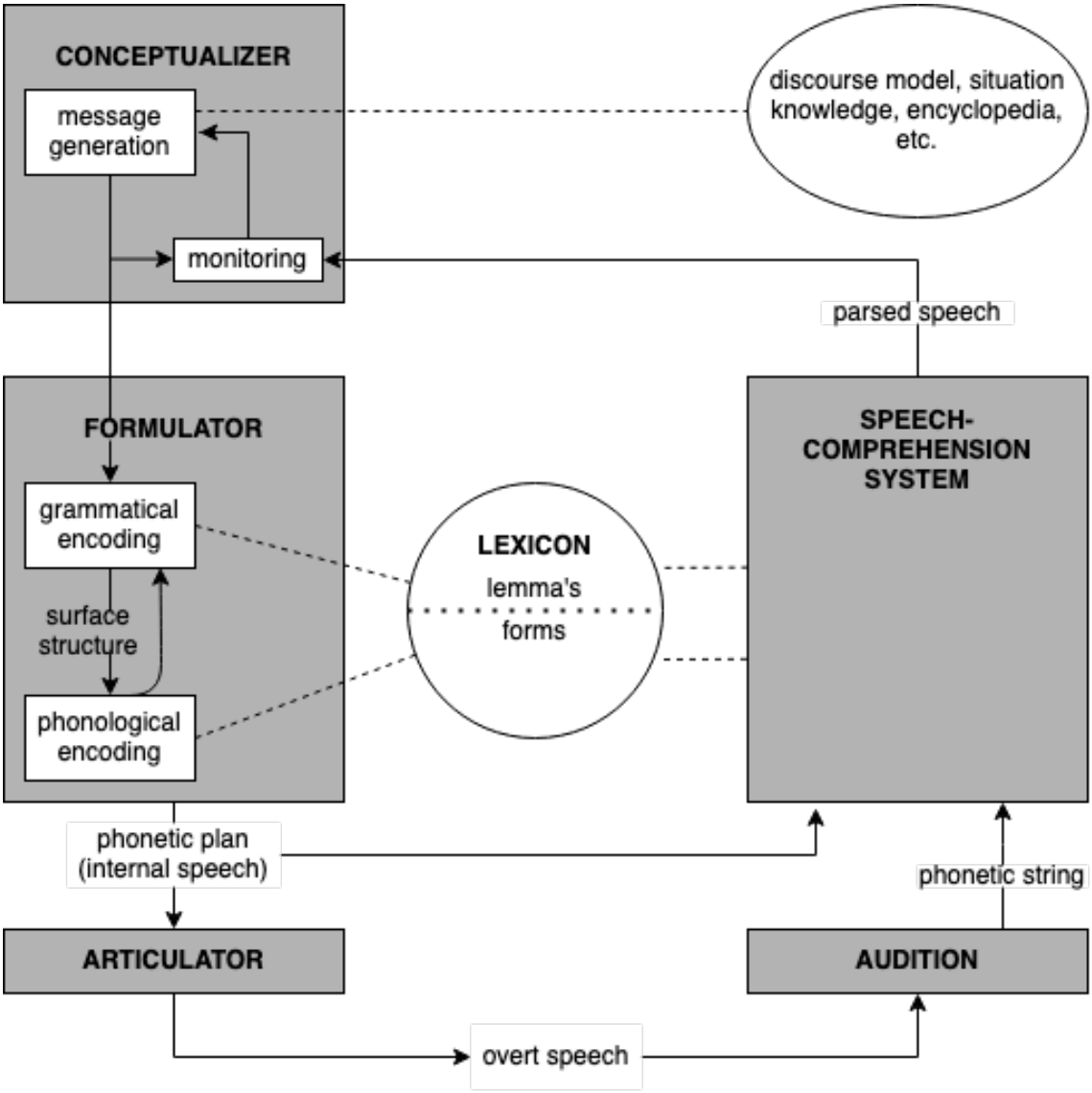
In this chapter, Levelt's (1989) speech production model is discussed. Then, Jaeger's 2010 claim that language is structured to be communicatively efficient is introduced, along with a principle that follows this claim, the Uniform Information Density Hypothesis. This hypothesis may be observed by looking at choice points, which are introduced after. Finally, the scope of the current research is laid out.

2.1 The production of language

A model for the production of language of typically developed adults has been proposed by Levelt (1989) and is shown in figure 1. This model contains knowledge stores, as well as different processes involved in the production of language: conceptualizing, formulating, articulating, and self-monitoring. Each of these processes receives some kind of input, and produces a different kind of output. Below, these processes are further explained.

The first procedure in the model is the conceptualizer, which is defined by Levelt as the sum of different activities pertaining to talking as an intentional activity: conceptualizing an intention, selecting the relevant information for the realization of that intention, ordering this information, keeping track of what has been said, and so on. In addition, a speaker will monitor their own utterances, which also influences the generation of the message. The output of the conceptualizing stage is then a *preverbal message*.

Figure 2.1: Levelt's (1989) speech production model



The formulator translates the preverbal output of the conceptualizer into a linguistic structure, in two steps. First, the message is grammatically encoded. This encoding consists of procedures for accessing lemmas in the lexicon and of procedures for syntactic building. A speaker's lemma information is stored in their mental lexicon. Lemma information contains the lemma's meaning, as well as its syntax. According to Levelt (1989), a lemma is activated once its meaning matches part of the preverbal message. Then, its syntax is made available, which in turn activates syntactic building procedures. Once all relevant lemmas have been accessed, the grammatical encoder has produced a *surface structure*. The second step is the phonological encoding of the message. This step builds an articulatory plan for each lemma and for the whole utterance. It retrieves morpholog-

ical and phonological information from the lexicon. Furthermore, modification or further specification of form information, such as stress on the sentence level, is laid out.

This articulatory plan has not yet been realized as overt speech. Articulation is the execution of this plan by the human voice system. However, articulation does not occur immediately after the articulatory plan has been created. In order to cope with this, a phonetic plan needs to be temporarily stored. The storage device is called the *articulatory buffer*. The articulator retreives chunks of internal speech from this buffer and executes them, resulting in overt speech.

A speaker has access to both their internal and overt speech. They can listen to their own overt speech, which involves an audition processing component. The speaker can then understand what they are saying. This processing is indicated in Figure 2.1 as the *speech-comprehension system*. That, too, consists of various subcomponents. The system has access to the form and lemma information in the lexicon, in order to recognize words and retrieve their meanings. The output of this system is *parsed speech*.

2.2 The Uniform Information Density hypothesis

The previous section concerned a model for the production of language, showing different processes and knowledge stores involved in this production. Jaeger (2010) hypothesized that human language could be organized to be efficient at all of these levels of linguistic processing, leading him to propose the UIDH, which is the focal point of this paper.

One of the earliest and most famous observations related to efficient language production is Zipf's (1949) finding that frequent words generally have shorter forms. This observation led Zipf to propose the *Principle of Least Effort*, according to which speakers try to communicate efficiently with the least expense of effort. Piantadosi et al. (2011) found that an even better predictor of word length is the average predictability of a word in context.

This predictability in context is an information theoretical notion: the more probable a word is in its context, the less information it carries in that context. The foundation for information theory was laid in 1948, when Shannon proposed a mathematical theory of communication. It was originally developed as a way to study and solve problems that arose in the transmission of signals over communication channels. According to Shannon,

a communication system consists of five parts: an information source, a transmitter, a channel, a receiver and a destination. While it was previously thought that an increase in transmission rate of information would increase the probability of error, Shannon found that the complexity level of the encoded information determined this probability. He named this complexity level of the encoded information H, or *entropy* (Shannon, 1948).

Shannon argued that the most important aspect of communication is that the chosen message is one that is selected from a set of possible messages. If the number of messages is finite, then this number can be seen as a measure of the information produced when one message is chosen from the set. The informative value of an individual element can be described with the formula below:

$$I(X) = -\log_2 p(X)$$

It is important to note that the concept of information in information theory does not equate to the notion of information in colloquial speech, and these should not be confused. In the context of Shannon's information theory, information is a measure of the freedom that one has when one selects a message. More freedom of choice leads to a greater uncertainty, which translates to a greater amount of information. Tribus (1961) called this measure *surprisal*, as it measures the surprise we get when we see the element. Surprisal, entropy and information are all the same concept, but to avoid confusion, the term surprisal will be used in this study.

Genzel & Charniak (2002) have stated that the successful transfer of information through a noisy channel with limited bandwidth is maximized by transmitting this information uniformly, and close to the channel's capacity. The human brain can be characterized as a channel that transmits information when processing language. If communication is to be efficient, it balances the risk of transmitting too much information, which would increase chances of information loss, against the desire to transmit as much information as possible. Jaeger (2010) stated that linguistic communication would be optimal if, on average, each word adds the same amount of information, and the rate of information transfer is close to the channel capacity. Even though it is unlikely that every aspect of language is maximally informative, due to other linguistic constraints, language may still be efficient: speakers may try to communicate efficiently within the bounds set by grammar.

The UIDH states that if speakers have a choice between several variants to encode their message, they are predicted to prefer the variant with a more uniform distribution of information (Jaeger, 2010). To avoid large fluctuations in the distribution of information, Jaeger (2010) claims that speakers may be managing the amount of information per amount of linguistic signal when they can. The information density, then, is the amount of information per amount of time.

If language production is organized to be efficient, and speakers prefer to distribute the information load uniformly across the linguistic signal, it should be possible to observe the effects in the preference of speakers at certain choice points. For example, the choice between the full or contracted form ("he is" versus "he's") (Frank & Jaeger, 2008), or the choice between full or reduced constituents, of which examples are given in (4) (Ferreira & Dell, 2000) and (5) (Rohdenburg, 1998):

(4) The coach knew (that) you missed practice.

(5) It helps you (to) focus where your money goes.

This optionality has been explained by various accounts. Ferreira & Dell (2000) found that speakers do not disambiguate sentences by mentioning optional function words, for example in regard to a temporary ambiguity, which is when words like "selected" may be initially interpreted as both active and passive, in sentences like "The astronauts selected for the Apollo missions made history" (Ferreira & Dell, 2000, p. 310). Rather, they found that the use of optional words is sensitive to the availability of the material. Omission was found to be more likely if embedded subjects were repeated, or when the recall consisted of the embedded clause material. Race & MacDonald (2003) found that omission of "that" was more likely before less common words. They argued that this was to alleviate production difficulty.

Jaeger (2010) claims that such a word might be omitted because the speaker wants to keep the distribution of information within a sentence as uniform as possible. Indeed, he shows the hypothesis to be an accurate predictor of omission of the complementizer "that" in sentences such as 4, even when considering other accounts.

The UIDH has been corroborated in other studies: Maurits, Navarro, & Perfors (2010) found that the distribution of word order across languages could at least in part be ex-

plained by the UIDH. Collins (2014) investigated syntactic alternations in English, and found that uniform information density was a powerful predictor of human preferences.

2.3 Irregular omission

The UIDH may be used to explain choices made by the speaker. This is most visible at choice points. The omissions that so far have been studied concern omissions that are allowed within the rules of a grammar. However, sometimes omission also occurs outside of these rules. This section is devoted to this last kind, irregular omission.

2.3.1 Omission in deviant speakers

Deviant speakers, such as children, or patients suffering from Broca's aphasia (a group of people that are linguistically impaired as a result of brain damage) may sometimes omit words from sentences where this is not allowed. In (2), an example of omission in the speech of developing children has been given. Omission in the speech of a patient suffering from Broca's aphasia is shown in (6) (Avrutin, 2001).

(6) Rekening is voldaan. (Patient HB)

To account for omission in children, an initial intuition may be that omission is due to them not having learned all rules of the language. Indeed, this is the explanation used by Poeppel & Wexler (1993). However, Grela (2003) showed that children with specific language impairment omitted subjects more frequently as linguistic complexity increased. This effect was not found for the typically developed control children. This supports the idea that grammatical errors in children may be due to problems with processing complex linguistic information rather than limitations in linguistic knowledge.

2.3.2 Omission in special registers

Regularly developed speakers may sometimes omit articles and other functional elements from a sentence, within a special register. These registers, such as newspaper headlines, diary style, and telegram style, are characterized by the ellipsis of syntactic elements that are otherwise required (Oosterhof & Rawoens, 2017).

Baauw, De Roo, & Avrutin (2002) note the similarities between omission in deviant speakers and in some forms of adult speech. In (7) and (8), it is shown that some contexts allow omission of determiners, and (9) shows that some contexts allow root infinitives (Baauw et al., 2002):

(7) Deur dicht!
Door closed!
'Close the door!'

(8) Leuk huisje heb je.
Nice little house have you.
'You have a nice house.'

(9) Ik een huis kopen? Nooit!
I a house to buy? Never!
'Me buying a house? Never!'

With this evidence from adult speech, Baauw et al. (2002) assume that omission in typical adult speech does not differ much from that in deviant speech, and propose a unified account of determiner drop and root infinitives, which takes the role of context into account.

In order to explain determiner drop and root infinitives in child, aphasic, and adult speech, Baauw et al. (2002) use a model of syntax-to-discourse mapping that is based on Heim's *file card semantics* (Heim, 1983). In this model, syntax and discourse are connected by functional elements, which are represented by metaphorical file cards reflecting entities used in discourse. The determiner establishes a connection between the noun phrase and an individual file card. File cards may be introduced extra-syntactically, by means of discourse presupposition. Baauw et al. (2002) argued that in cases of determiner and tense omission in special registers, context provides enough information to allow the listener to introduce the appropriate file cards. While typically developed adults can make use of this register, children and aphasics are argued to rely on discourse presupposition more often as it is an easier, more economic way to introduce file cards (Baauw et al., 2002).

Schumacher & Avrutin (2011) point out that different special registers vary in their makeup. Diary style often features the omission of subject, while newspaper headlines allow omission of articles and tense, see (10) (Trouw, 2020):

(10) Friesland als laatste provincie nu ook regenboogprovincie
Friesland as final province now also rainbow province
'Friesland is the final province to become a rainbow province'

The mechanism behind omission in newspaper headlines is subject to ongoing debate. One might entertain the thought that the reason for omission is to keep headlines short and concise Reich (2017). Indeed, this is what production guides to journalistic writing provide as advice, as Oosterhof & Rawoens (2017) mention. However, as Reich (2017) notes, this does not explain Stowell's (1991) observation that there is asymmetry in newspaper headlines. See the examples below:

(11) a. Man bites a dog.

b. * A man bites dog.

Stowell (1991) posits that c-command is the key here: article omission in a DP is impossible if that DP is c-commanded by a DP with an overtly realized article. This would predict a strict rule in which no exceptions would be allowed. However, those exceptions were found by De Lange (2008) in practice.

2.4 Information theory and irregular omission

This section will explore two theories that have been proposed in regard to newspaper headlines and information theory. Section 2.2 introduced the UIDH (Jaeger, 2010), which was adapted by Lemke et al. (2017) to account for newspaper headlines. The other theory that is explored is one by De Lange (2008), who argued that it was entropy of the system rather than the surprisal of individual elements that account for omission patterns.

2.4.1 Uniform information density in newspaper headlines

The UIDH has been used to account for article omission in German headlines. Article omission can be seen as similar to the omission of complementizers, in that there are two alternatives that convey the same meaning. Horch & Reich (2016) found that the use of articles significantly lowered the surprisal of the following head noun. Based on this discovery, Lemke et al. (2017) conducted a corpus study and found that articles were significantly more frequent when they preceded a less predictable head noun. In addition, they found that participants were more likely to accept omitted articles if the head noun was more predictable. These results are in line with the predictions made by the UIDH. A reduplication study for Dutch found the same effect (Van Tuijl & Paperno, 2020).

2.4.2 Entropy of the system

De Lange (2008) studied omission of articles in Italian and Dutch. She notes the crosslinguistic similarities in omission patterns in newspaper headlines and child speech, and argues that these patterns are governed by the same principle. She uses Information Theory (Shannon, 1948) to account for the omission patterns. However, rather than using surprisal of individual elements, De Lange uses relative entropy of a set to account for omission rates. Relative entropy can be used as a measure of set complexity and is obtained by dividing the sum of the absolute entropy values of the members of the set by the sum of the theoretical maximum entropy values of each member of the set. According to De Lange (2008), set entropy is tied to a limitation, either a limited reading time in newspaper headlines, or a cognitive limitation in a developing brain. She found that this idea was corroborated by the difference between the Dutch and Italian article paradigms, with the set entropy of the Italian article paradigm being lower, thus reflecting a lower complexity. She found that this principle may be used to explain both the higher rates of omission in newspaper headlines in Dutch compared to Italian, as well as the higher omission rate that Dutch children have compared to their Italian peers.

2.5 The current study

The first section of this chapter has shown a model for the production of speech. According to Jaeger (2010), different modules of speech are governed by a tendency to distribute information as evenly as possible. Indeed, in typical speech, it was found that such a tendency seems to exist. However, it is unclear if all speakers and utterances are sensitive to this distribution. If omission of child speech were to be explained via the UIDH, it would require them to be sensitive to information load. This is a rather bold requirement.

Writers of newspaper headlines are typical speakers in other contexts, and as such may be sensitive to information load, as Lemke et al. (2017) suggest. However, they did not consider the asymmetry of omissions. In addition, Lemke et al. (2017) fail to explain why article omission is only allowed in special registers. Finally, they only studied a rather limited set of headlines (a total number of 131) in only one language, German, and while their results were corroborated by a corpus study on Dutch headlines (Van Tuijl &

Paperno, 2020), this latter language has a limited article paradigm. Furthermore, neither language contains overt articles for indefinite plural nouns.

The current study is devoted to Italian, as its article paradigm is different from that of Dutch and German. The relation between surprisal of the noun and article omission will be researched via a corpus study, where it is expected that, in line with German and Dutch, the omission rates of articles will be higher for low surprisal head nouns.

2.5.1 The Italian article paradigm

The Italian article paradigm is interesting to contrast with the Dutch and German paradigm. While Dutch and German have articles that serve multiple functions (for example, the Dutch article "de" serves as a marker for the definite common gender singular, and the definite plural of both genders), a single function is never fulfilled by multiple articles. However, this is the case for the Italian articles, although their use depends on the morphology of the words that follows it (Caramazza, Miozzo, Costa, Schiller, & Alario, 2001). These articles may in turn fulfill multiple roles, making the Italian article paradigm an interesting one. Furthermore, while indefinite plurals in both Dutch and German are by default without article, there are no such gaps in the Italian article paradigm. The full paradigm is given in Table 2.1.

Table 2.1: Morphosyntactic forms of Italian articles

	Definite		**Indefinite**		**Partitive**	
	Masculine	Feminine	Masculine	Feminine	Masculine	Feminine
Singular	il lo l'	la l'	un uno	una un'	del dello dell'	della dell'
Plural	i gli	le	dei degli	delle	dei degli	delle

2.5.2 The role of definiteness in article omission

De Lange (2008) found that in Italian, the preference for omission is higher when it comes to indefinite articles, although she notes that, in the condition with definite articles, the nouns were inherently unique nouns, such as "prime minister". These are more often associated with a definite determiner than common, not inherently unique nouns, such as "journalist". In other words, the use of an inherently unique noun increases the prob-

ability of a definite determiner. The relation between definiteness and surprisal will be investigated as well.

Chapter 3

Method

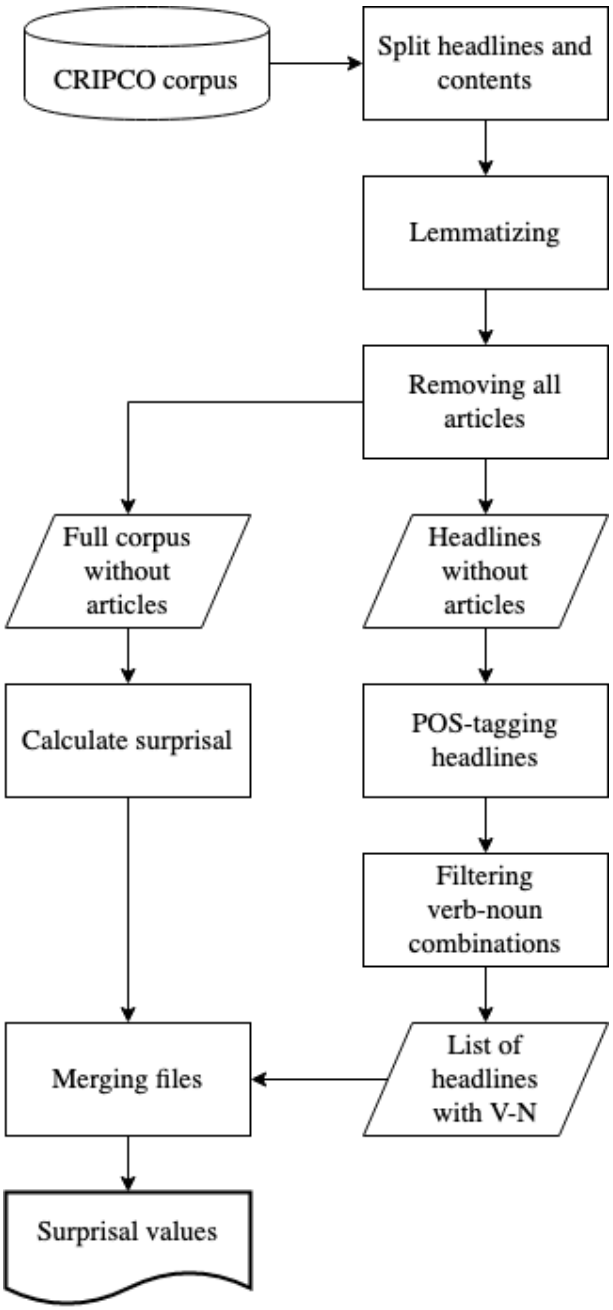
In the previous chapter, we have seen that the Uniform Information Density Hypothesis has been shown to be promising in accounting for the omission of articles in newspaper headlines in both German (Lemke et al., 2017) and Dutch (Van Tuijl & Paperno, 2020), but that the Italian paradigm warrants a replication.

If language users try to distribute the information in a sentence as evenly as possible, as the UIDH postulates, this should be reflected in a corpus of the text type that allows omissions, in this case newspaper headlines. To determine whether or not the UIDH can explain article omission in Italian newspaper headlines, the corpus study by Lemke et al. (2017) has for a large part been replicated, although the method that has been used varied slightly. An overview of the methodology is given in Figure 3.1.

3.1 Corpus

The corpus that has been used for this study is the Cross-document Italian People Coreference corpus (CRIPCO) (Bentivogli, Girardi, & Pianta, 2008). This corpus is composed of a subset of news reports published by the Italian local newspaper "L'Adige" from 1999 to 2006, consisting of 43,328 documents. The original aim of the corpus was to create a gold reference for cross-document coreference resolution of person entities. However, as the corpus constitutes an unbiased selection of newspaper articles when it comes to article omission in headlines, it is also well suited for the current study.

Figure 3.1: Overview of the study



3.1.1 Preprocessing

In order to make the corpus suitable for use in this study, it first had to be preprocessed by teasing apart the headlines from the rest of the article.

Then, in order to calculate the surprisal of the base form rather than the surface form, the lemmatizer from the Python package SpaCy (Honnibal & Montani, 2017) was

employed. This way, the surprisal reflects the probability of a word's occurrence regardless of any inflection.

To reflect the chance of a noun occurring after a specific verb, regardless of eventual placement of articles in between, all articles were removed from the corpus. Information about where the articles were removed was stored in a separate file. Then, two separate corpora were created: one with all source material excluding articles, consisting of 536,999 lines and 16,785,630 tokens, and one with just the headlines, consisting of 33,307 lines and 245,654 tokens.

Only nouns that immediately followed a verb were taken into account in order to keep confounding variables to a minimum. In order to create a subset of only those headlines where nouns precede a verb, a Part of Speech tagger was used (Honnibal & Montani, 2017). Furthermore, only verb lemmata that occurred at least 10 times in the corpus were used. This generated a list of 2952 headlines that contained an article between the verb and the noun, and 1007 headlines where at least one article was omitted.

3.2 Training a language model

The aim of the research was to study the relation between omission of articles and the surprisal of the noun. Following the procedure Lemke et al. (2017) used for German headlines, bigram surprisal of a noun was used to investigate this relation. Bigram surprisal captures the probability of a word occurring, given only the previous word. For this experiment, this comes down to $-log_2(noun|verb)$, or the probability of a noun occurring after a given verb.

These surprisal values were calculated by means of the SRILM toolkit (Stolcke et al., 2011). In line with Lemke et al. (2017), the model was trained using Kneser-Ney smoothing (Kneser & Ney, 1995) and interpolated.

3.3 Subsetting the data

As shown in table 2.1, the Italian article paradigm contains definite, indefinite and partitive articles. As there is overlap between these types in the plural forms, a subset was made of only the headlines that contained a singular definite or singular indefinite article to study

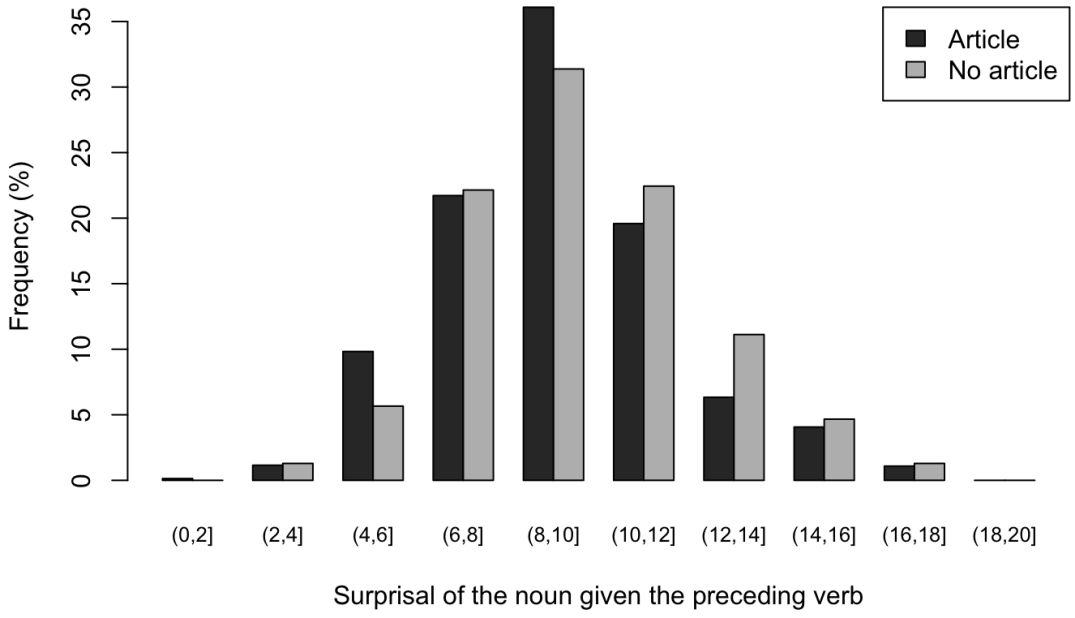
the effect of definiteness. This subset consisted of 1803 headlines with a definite singular article, and of 631 headlines with an indefinite singular article.

Chapter 4

Results

To investigate the effect of surprisal on article omission in Italian newspaper headlines, a corpus study has been conducted. Figure 4.1 shows a bar plot of the distribution of article omission across surprisal values. The surprisal of nouns that are not preceded by an article is higher than the surprisal of nouns that are preceded by an article.

Figure 4.1: Surprisal values of nouns with and without articles

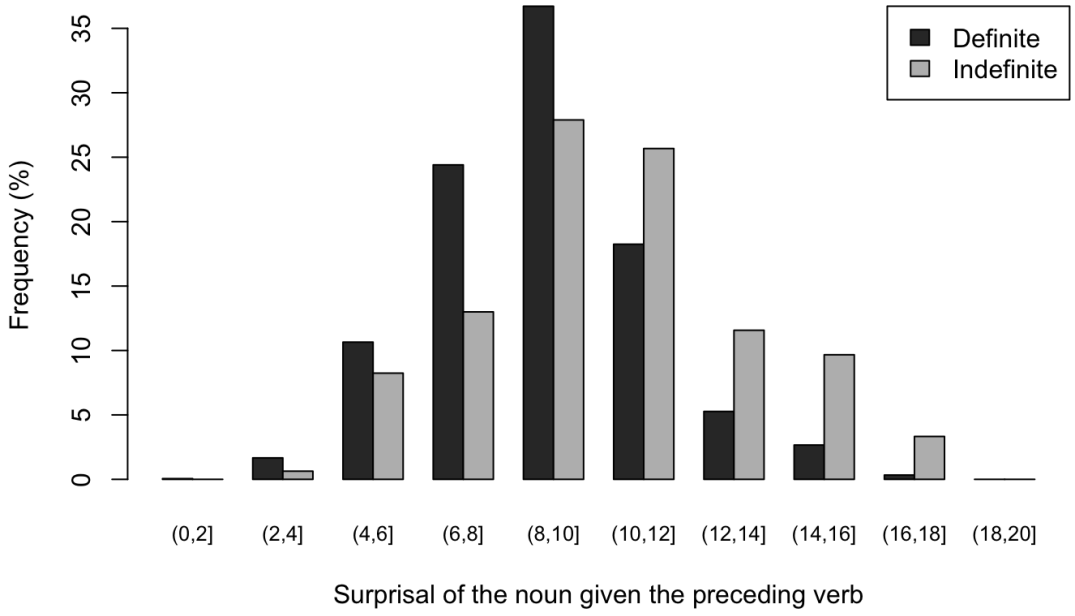


This effect was found to be statistically significant using a generalized linear mixed model using the lme4 (Bates et al., 2015) package in R (R Core Team, 2020) with random intercepts for noun and verb lemmata. The model that included surprisal as main effect

fitted the data significantly better than a baseline model without such an effect (χ^2=7.3324, p=0.007).

As De Lange (2008) showed an effect of definiteness on omission, the relation between surprisal and definiteness in a subset of the corpus, as defined in section 3.3 was studied as well. A bar graph containing these results is given in figure 4.2.

Figure 4.2: Surprisal values of nouns after definite and indefinite articles



As definiteness is not predicted by a noun's surprisal, nor the other way around, an independent samples t-test assuming unequal variances was conducted. This test showed that the relation between definiteness of overt articles in Italian newspaper headlines and the surprisal of the following noun was significant, $t((935.93)=10.891$, $p<0.0001$.

Chapter 5

Discussion

In the previous chapter, findings were shown that run directly counter to the expectations set by the Uniform Information Density Hypothesis, which predicted that omission would be more frequent before low surprisal nouns, as as an overtly realized article lowers the surprisal of the following noun (Horch & Reich, 2016). However, the opposite effect was found: high surprisal words were less likely to be preceded by an article than low surprisal words. This effect was statistically significant, and therefore directly contradicts the UIDH.

In addition, the role of definiteness was investigated. It was found that the surprisal of nouns that were preceded by an overt definite article was significantly lower than the surprisal of nouns that were preceded by an indefinite article. However, it is important to note that this only concerns overt articles, as the definiteness of omitted articles is inherently unknown. This makes a study into the exact role that definiteness plays in article omission a challenging effort.

A significant effect of noun surprisal on the realization of articles in Italian newspaper headlines was found. This result is consistent with the account proposed by De Lange (2008), who uses set entropy to explain crosslinguistic patterns of omission and thus makes no claims about the effect of surprisal of individual nouns on article omission.

The findings in this study compared to those of Lemke et al. (2017) and Van Tuijl & Paperno (2020) strongly suggest that the difference in article paradigms may be a reason for the difference in mechanisms that are used for article omission. While set entropy may be used explain general patterns of omission in newspaper headlines, future research should establish the effect of the surprisal of the individual noun on article omission.

The finding that the UIDH does not govern article omission in Italian headlines does not necessarily mean that the hypothesis does not hold at all, as it has been shown to be promising in regard to regular speech (e.g. Frank & Jaeger, 2008; Jaeger, 2010; Maurits et al., 2010; Collins, 2014). More likely, it shows that the production of language in this special register cannot be modelled in the same way regular speech can be (Levelt, 1989). Therefore, a new model might be needed to capture the processes involved in language production in newspaper headlines. Moreover, it is quite possible that this is not only the case for newspaper headlines, but for special registers in general. In this light, it is interesting to investigate the effect of information density on omissions in spontaneous language in a special register, such as colloquial speech, as in (8).

Chapter 6

Conclusion

To study the effect of noun surprisal on the omission of articles in newspaper headlines, a corpus study was conducted. In this study, the surprisal of the noun in context of a verb was researched in relation to omission of articles in Italian newspaper headlines. The Uniform Information Density Hypothesis states that speakers efficiently communicate by distributing information load as evenly as possible over a sentence.

As Horch & Reich (2016) established that overt articles lower the surprisal of the noun that follows it, it would be expected on the basis of the UIDH that articles occur more frequently before high surprisal nouns. Indeed, such an effect was found in previous studies for German and Dutch headlines (Lemke et al., 2017; Van Tuijl & Paperno, 2020).

In this study, the same effect was studied for Italian newspaper headlines. Contrary to the earlier findings, however, high surprisal nouns in Italian headlines were less likely to be preceded by an article, rather than more. This effect was statistically significant, which directly contradicts the UIDH.

While future studies should determine the exact role that surprisal plays in relation to the omission of articles in newspaper headlines, a uniform information density account fails to provide an explanation for article omission in Italian newspaper headlines. As the UIDH is well-established in other domains of speech, it ultimately follows that the production of newspaper headlines does not follow the same path regular speech does.

Bibliography

Avrutin, S. (2001). Linguistics and agrammatism. *Glot international*, *5*(3), 1–11.

Baauw, S., De Roo, E., & Avrutin, S. (2002). Determiner omission in language acquisition and language impairment: Syntactic and discourse factors. In *Bucld proceedings* (Vol. 26, pp. 24–35).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: doi:10.18637/jss.v067.i01

Bentivogli, L., Girardi, C., & Pianta, E. (2008). Creating a gold standard for person cross-document coreference resolution in Italian news. In *The workshop programme* (p. 19).

Caramazza, A., Miozzo, M., Costa, A., Schiller, N. O., & Alario, F.-X. (2001). A cross-linguistic investigation of determiner production. In E. Dupoux (Ed.), *Language, brain, and cognitive development: Essays in honor of jacques mehler* (pp. 209–226). MIT Press Cambridge, MA.

Collins, M. X. (2014). Information density and dependency length as complementary cognitive models. *Journal of psycholinguistic research*, *43*(5), 651–681.

De Lange, J. (2008). *Article omission in headlines and child language: A processing approach*. Netherlands Graduate School of Linguistics.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, *40*(4), 296–340.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 30).

Friesland als laatste provincie nu ook regenboogprovincie. (2020, May). *Trouw*. Retrieved from `https://www.trouw.nl/nieuws/friesland-als-laatste-provincie-nu-ook-regenboogprovincie~b4df1ba9/`

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 199–206).

Grela, B. G. (2003). The omission of subject arguments in children with specific language impairment. *Clinical linguistics & phonetics*, *17*(2), 153–169.

Heim, I. (1983). File change semantics and the familiarity theory of definiteness. *Semantics Critical Concepts in Linguistics*, 108–135.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.* (To appear)

Horch, E., & Reich, I. (2016). On "article omission" in German and the "uniform information density hypothesis". *Bochumer Linguistische Arbeitsberichte*, 125.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23–62. doi: doi:10.1016/j.cogpsych.2010.02.002

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing* (Vol. 1, pp. 181–184).

Lemke, R., Horch, E., & Reich, I. (2017, 4). Optimal encoding! – information theory constrains article omission in newspaper headlines. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics* (Vol. 2, pp. 131–135).

Levelt, W. J. (1989). *Speaking: From intention to articulation.* MIT press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

Mathieu van der Poel takes final stage and Tour of Britain glory. (2019, Sep). *The Guardian*. Retrieved from `https://www.theguardian.com/sport/2019/sep/`

14/mathieu-van-der-poel-wins-final-stage-and-tour-of-britain-cycling-manchester

Maurits, L., Navarro, D., & Perfors, A. (2010). Why are some word orders more common than others? a uniform information density account. In *Advances in neural information processing systems* (pp. 1585–1593).

Oosterhof, A., & Rawoens, G. (2017). Register variation and distributional patterns in article omission in Dutch headlines. *Linguistic Variation*, *17*(2), 205–228.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Poeppel, D., & Wexler, K. (1993). The full competence hypothesis of clause structure in early German. *language*, 1–33.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Race, D. S., & MacDonald, M. C. (2003). The use of "that" in the production and comprehension of object relative clauses. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 25).

Reich, I. (2017). On the omission of articles and copulae in German newspaper headlines. *Linguistic Variation*, *17*(2), 186–204.

Rohdenburg, G. (1998). Clausal complementation and cognitive complexity in English. In *Anglistentag* (pp. 101–112).

Schumacher, P. B., & Avrutin, S. (2011). Register affects language comprehension: ERP evidence from article omission in newspaper headlines. *Journal of Neurolinguistics*, *24*(3), 304–319.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at sixteen: Update and outlook. In *Proceedings of ieee automatic speech recognition and understanding workshop* (Vol. 5).

Stowell, T. (1991). Empty heads in abbreviated English. *GLOW abstract, GLOW Newsletter*, *26*.

Tribus, M. (1961). *Thermostatics and thermodynamics.*

Van Tuijl, R., & Paperno, D. (2020). *Article omission in Dutch newspaper headlines decreases as nouns become more informative and article entropy increases.*

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* addison-wesley press.