Ruled by construal? Framing article choice in English

Romain, Laurence; Hanzlikova, Dagmar; Milin, Petar; Divjak, Dagmar

DOI:
10.1075/cf.22005.rom

Document Version
Peer reviewed version

Citation for published version (Harvard):
Romain, L, Hanzlikova, D, Milin, P & Divjak, D 2024, 'Ruled by construal? Framing article choice in English', *Constructions and Frames*. https://doi.org/10.1075/cf.22005.rom

Ruled by Construal? Framing article choice in English

Laurence Romain[a], Dagmar Hanzlíková [b*], Petar Milin[a], Dagmar Divjak[a,c]

[a] *Department of Modern Languages, University of Birmingham, Birmingham, United Kingdom*

[b] *Independent researcher*

[c] *Department of English Language & Linguistics, University of Birmingham, Birmingham, United Kingdom*

**The work presented here was conducted by this author while she was affiliated with the University of Sheffield as part of the project team.*

Correspondence should be addressed to ooominds@ooominds.org

Data availability Statement

The scripts and data used in this paper can be accessed through the following links XXX and a dummy version of the Qualtrics survey we used can be found here: https://birminghamcoaal.eu.qualtrics.com/jfe/form/SV_8IaA8NWqK7QSGuq

Acknowledgments

The work reported on in this manuscript was funded by Leverhulme Trust Leadership Grant RL-016-001 to Dagmar Divjak which funded all authors. Special thanks go to Marianne Hundt for reading a draft of this paper and providing relevant feedback. We would also like to thank the audiences to which we presented this research (Grammar and Corpora, Ghent 2022) and the three anonymous journal reviewers for their very helpful comments.

Abstract

In cognitive linguistics, grammatical structure is known to be representative of meaning. This is also true of English articles. In this paper, we argue that the choice of article, when the grammar allows it, is dependent on the wider discourse context and most importantly on how the speaker construes this context. Using survey data from 181 native speakers of English, we show that the choice of article depends on the activation of semantic frames and how speakers may choose to highlight different elements of a frame to construe the situation differently. We rely on Entropy to measure the restrictiveness of a context and to identify particular contexts in which choice is allowed or inhibited. We find that some contextual features such as the specificity of the referent are more restrictive while Hearer Knowledge is more open to construal.

Keywords

Articles, construal, survey data, reference

1. Aims & Objectives

Cognitive linguists have a long-standing interest in grammatical and lexical variation. The presence of variation in communal and individual usage, and language users' ability to navigate this variation, provides evidence in support of the hypothesis that language emerges from exposure to usage. This paper examines the role the wider discourse context, as captured by semantic frames, plays in choosing one grammatical variant over another when the grammar allows choice. It also quantifies the extent to which the different properties that are thought to govern the choice affect individuals' freedom to exploit the affordances offered by grammatical variation.

As a case in point, we analyse data on the choices speakers make when establishing and talking about referents. Crucial to the referential process, beside the name/noun chosen to describe the referent, is the choice of article, which speakers use to ground the referent in the current discourse situation. In this paper, we focus on the three main articles of English *the, a(n)* and Ø and trace back variation in their use to the contexts in which they are used. We will show that the restrictions emanating from the wider discourse context often rely on the activation of semantic frames and how different speakers may choose to highlight different elements of a frame to construe the situation differently, justifying different grammatical choices. Semantic frames enable us to focus on lexical semantic issues that are relevant to grammatical structure by narrowing down the number and types of elements that can be construed as expected from context and thus influencing grammatical choice. We will see that in cases where the context allows the article slot to be filled by more than one possible article, the choice boils down to how speakers construe the referent and as such, the way they construe the referent is expected to be reflected in the grammatical structure they prefer.

We also examine to what extent and under which conditions individual speakers' choices may vary. Analysing three-alternative forced choice (3AFC) data collected from 181 native speakers of English, we rely on Entropy to quantify the restrictiveness of the context and to identify types of contexts in which choice is allowed versus inhibited. We will show that there are different types of

contexts; some contextual properties, such as Referent Specificity, are rather restrictive, leaving the speaker with little choice in terms of which article to use while other contextual properties, such as Hearer Knowledge, are such that several articles are possible, albeit with slightly different semantic implications.

2. State of the art

Grammatical choice and individual variation have long been used by cognitive linguists as evidence that it is through exposure to usage that speakers learn which grammatical or lexical forms and structures can be used in a particular context. Because exposure to usage differs for every individual, what they learn may well differ and evidence has started to accumulate of individual differences between speakers of the same language, notably in terms of their lexical and grammatical knowledge (Clark 1997; Dąbrowska 2015: 651; Farmer et al. 2012; Mulder and Hulstijn 2011).

Cognitive linguists have a long-standing interest in understanding, and trying to predict, constructional and lexical choices, notably by measuring the probability that one option will be chosen over another, given a range of contextual properties. It is now accepted that language users are capable of learning not only what is correct/conventional but also what kinds of contexts allow for modulation, based on their experience with language. Speakers' knowledge of which grammatical structures can be used feeds into choices as to how they want to present events and/or entities: speakers will adapt their choices based not only on what information they want to convey but also on how they want to convey this information. Since studies on variation focus on choices made by speakers, they tend to be restricted in their scope; that is, they focus on subsets where both (or more) options are available and interchangeable, omitting cases where this is not the case (e.g., verbs that *cannot* be used in certain argument structure constructions). It is therefore particularly interesting to identify contexts that do not allow variation as these can be used to identify the contextual properties that encourage or suppress variation.

In what follows we will discuss how articles are usually accounted for in the cognitive linguistic literature where they are considered to be grounding elements, used to anchor nominals in the

discourse situation (2.1). We will argue that this grounding depends on the activation of semantic frames (2.2) and will show how the selective focus on elements of a frame facilitates construal (2.3).

2.1 Articles as grounding elements

Within Cognitive Grammar, grounding systems are considered to be key components of language. Grounding is the speaker's "anchoring" of a situation and its participants in the speech situation shared by speaker and hearer. The situation that is described and its participants can be seen as figures while the speech situation functions as ground. The ground includes "the speaker and hearer, the speech event in which they participate, and their immediate circumstances (e.g., the time and place of speaking)" (Langacker 2008: 78).

Grounding is so important to successful communication that the grammar of English, for example, forces its speakers to use grounding elements in every sentence (Radden and Dirven 2007: 49). Grounding elements are typically highly grammaticalized linguistic elements that have nucleus status: they are obligatory grammatical forms linked to the noun(s) and the verb in the sentence and therefore tightly intertwined with the grammatical core of the sentence (Radden and Dirven 2007: 49).

According to Langacker (2008: 263), when seen on the lexico-grammatical continuum, articles are closer to the grammatical pole than to the lexical pole. The article is an element of the nominal grounding system, which encompasses determiners, quantifiers, possessives etc. (Langacker 2004: 87-89; Radden and Dirven 2007: 87-89). These elements provide means for identifying nominal reference relative to the ground. More specifically, they ground the thing described by the noun in the current discourse and make it accessible to the hearer[1] as referents. They indicate whether the things talked about are or are not identifiable in the current discourse (Radden and Dirven 2007: 49).

[1] We use 'hearer' in the broad sense of 'addressee' in this paper and use the two terms interchangeably.

Crucially, the ground is construed subjectively. The subjective nature of grounding elements arises from the point of view taken by the speaker: the speaker assesses whether the situation described is real or potential, and whether the hearer can or cannot identify the participants talked about (Radden and Dirven 2007: 49). Grounding, therefore, pertains to our conceptual organization (Langacker 2008: 272). Langacker (2016) points to the fact that all grounding in use is local and heavily depends on the Current Discourse Space, emphasizing the discursive nature of all grounding. More specifically, Langacker (2004: 103) explains the difference between definites and indefinites in terms of the discourse structure: definites are used when the referent "has some role in the structure being updated" whereas indefinites are "only introduced and identified through the content of the clause containing it".

2.2 Mental spaces and semantic frames

Epstein (2002) elaborated an account of the definite article in terms of Mental Spaces (Fauconnier 2007), which are not veridical representations of reality, but cognitive models of it. For example, the sentence *We went for coffee at that new independent shop the other day* conjures up a mental space that includes us, having coffee, independent coffee shops, the other day, and so on. Mental spaces are thought to exist in working memory, where they operate on information retrieved from long-term memory. This information is continuously amended to meet the needs of the unfolding conversational situation and can incorporate immediate experience or information obtained during a conversation. Frequent mental spaces can become entrenched in long-term memory; such mental spaces are known as Frames. Frames can be retrieved or activated in their entirety, and as such they can be used to organise mental spaces. For example, the frames for "going to a public place for pleasure", "commercial purchase" and many others can be used to structure the mental space and would do so in different ways.

Frames provide descriptions of concepts, such as the famous RESTAURANT example (Schank and

Abelson 1977), in terms of their association with experience rather than in truth-conditional terms. That is, they activate certain elements and relationships that are expected in the situation. If we take a similar example, such as COFFEE SHOP, the associated concepts include i.e., the presence of cups, staff, a counter, chairs and tables but also events such as putting in your order, paying for your drink and so on. Not all elements are equally activated within the frame, however: some elements are more central (e.g., the coffee, the cups) while others are less so (e.g., some drink that is only found in some coffee shops, some element of decoration etc.). Therefore, while speakers make use of this expected knowledge activated by the frame, they can also easily introduce elements that are new to their addressee, as in (1a). (1a) can be compared to (1b) where the speaker assumes that their addressee knows about the rooftop terrace or that it is potentially expected that coffee shops have a terrace.

(1)

a. In my local coffee shop, the staff are super nice and the coffee is really good. They also have a rooftop terrace that is open throughout the year.

b. In my local coffee shop, the staff are super nice and the coffee is really good. The rooftop terrace is open throughout the year.

Although various kinds of frames may be activated in speech, not all of them are necessarily in focus (see also Lambrecht 1994: 90-92 for a discussion of frames w.r.t. articles). Frames are generally activated in speech by some referent or event, such as a restaurant or a buying event respectively. While it could be argued then that each noun or nominal phrase and each verb denoting an event potentially activates a frame, only some frames are relevant to the context and not all frames are equally activated at the same time/in the same discourse situation. Since frames need some contextual element (linguistic or not) in order to be activated, it implies that before the triggering element is present/mentioned, the frame is not activated. This is what typically happens with generic

statements, which can be made in any context. Generic statements do not need any particular introduction and as such they are not activated by a frame. Take example (2) about Ivory Coast car registration plates from Wikipedia:

(2) Ø Ivory Coast plates are unique because of the color scheme and the location of the identifier band.

The noun phrase *Ivory Coast plates* is used without any sign of grounding, i.e., without an overt article. While the mention of this referent (*Ivory Coast plates*) activates the frame of registration plates, its first mention did not require that the corresponding frame be previously activated in discourse. Interestingly, in the literature, generic statements are considered to be known to the hearer[2] (henceforth HK+) and non-specific (henceforth SR-). The knowledge they require appears to rely on what is known as semantic memory. Semantic memory is a type of long-term memory system that stores general knowledge, i.e., knowledge that is not personal to an individual. Information that depends on personal experience is handled by episodic memory, another type of long-term memory that is used for personal facts (Divjak 2019: 105-106). In these terms, generic statements activate memory traces "stored" in semantic memory while non-generic referents that are HK+ rely on memory traces that belong to episodic memory. Either way, what is relevant to the topic at hand is that generics do not require the activation of a frame prior to being mentioned but they activate a frame themselves. In our example (2), we assume that the mention of *Ivory Coast plates* activates the frame of registration plates, including their general shape, their purpose and the various elements they are made of. It is relevant to note that the noun *identifier band* is introduced by the definite article *the* and thus seems construed by the speaker and understood by the addressee as part of the "registration plate frame".

[2] Note that we use the term hearer as in the phrase 'Hearer Knowledge' but that hearer refers to any sort of addressee.

The possibility for the speaker to modify how they present these elements is related to construal. While the activation of a frame leads to the activation of elements closely associated with it, it is nevertheless and to some extent up to the speaker to decide how they present these elements, either as new (a rooftop terrace, in example (1)) or as expected (the rooftop terrace, in example (1)); this is to some extent but not entirely similar to Ariel's notion of accessibility (Ariel 2001). Fillmore (1982: 127-129) also makes the case that framing can vary depending on the social use of a word. That is, the same word used by people in different communities may trigger different interpretations (related to different frames). Fillmore uses as examples the concept INNOCENT and MURDER, which will have slightly different meanings within the legal community and outside of it (Croft and Cruse 2004: 18). This in-group framing relies on whether speakers consider their addressee to have some shared knowledge with them, as we discuss in the next subsection.

2.3 Construal operations

The presentation of elements as either new or known to the addressee depends on the speaker's conception of what their addressee knows (Authors, under revision; Chafe 1976: 54; Horton and Keysar 1996; Quirk et al. 1985). This expectation can be measured via common ground: Clark (1996: 92) defines common ground as "the sum of [two people's] mutual, common or joint knowledge, beliefs, and suppositions." His definition follows from Stalnaker's (1978) first introduction of the notion of common ground, which was itself inspired by notions such as "common knowledge" (Lewis 1969), "mutual knowledge" or "belief" (Schiffer 1972), and "joint knowledge" (McCarthy 1990). Croft and Cruse (2004: 61) associate Clark's common ground with epistemic perspective: what is assumed to be part of the common ground or not will guide the perspective taken by speakers. Croft and Cruse illustrate this by pointing out the difference between the use of the definite article *the* and the indefinite article *a*, reproduced in (3) below.

(3)

a. Did you see a hedgehog?

b. Did you see the hedgehog?

They argue that the choice of article represents different construals of what the hearer knows: with the indefinite, the referent is construed as new to the hearer whereas with the definite it is construed as known and part of the common ground between speaker and hearer. What is considered common ground is not necessarily fixed or straightforward: speaker and addressee may not share the exact same conceptualization of their common ground. What drives the choice of article is thus what the speaker expects or assumes to be part of this common ground. It is not always the case that these assumptions work out: sometimes, an addressee might not know about a referent that the speaker presents as known.

Building on Clark's notion of common ground and shared knowledge or expertise, Fillmore argues that different groups will construe the same words slightly differently (recall INNOCENT and MURDER). We argue here that this in-group framing also applies to articles. Speakers who consider that their addressee is part of the same group as them will expect certain knowledge from their addressee. Therefore, they can more easily use the definite article *the* to refer to entities that they assume are part of their addressee's knowledge or expertise. As Epstein (2002) puts it: the use of *the* signals that an "access path" is available to the addressee (cf. also Givón 1992 and Ariel 1988 for a discussion of "accessibility"). Let us illustrate this with a real-life example. We argue here that (4a) is expected if one is addressing someone who has at least some knowledge of data analysis and statistics but (4b) is preferred when this is not the case.

(4)

a. Analyses on the data were conducted in R, <u>the</u> widely used software environment for statistics.

b. Analyses on the data were conducted in R, a widely used software environment for statistics.

The difference in the choice of article illustrates the speaker's construal of an entity as part of some in-group knowledge. Choosing *the* reveals the assumption that the addressee is part of a group of people who share this knowledge and that therefore, the "stats in R" frame is activated for them. Should they be confused by the use of *the*, they would show that they in fact are not part of this group and do not have access to this knowledge and are therefore unable to activate the "stats in R" frame. As such, different speakers, due to their different backgrounds and areas of expertise, can be expected to make different choices that also align with their assumptions with regards to their addressee's own knowledge.

2.4 This study

Articles are used as grounding elements in the discourse, and more specifically in the current discourse situation. Speakers use articles to build meaning and to situate entities in relation to their discourse space. This grounding participates in the building of mental spaces and is correlated to the activation of frames: some elements activate frames, which in turn trigger a certain number of related concepts and entities which guide the choice of article, based on whether the corresponding entity is assumed to be part of an activated frame. Whether an entity or concept is part of a frame is not entirely fixed but is dependent on the speaker's conceptualisation or construal, not only of the frame but of the amount of shared knowledge and expertise they assume to have with their addressee(s). This shared knowledge or expertise, also known as common ground, can be modulated for social and stylistic purposes, e.g., to create a difference between in-group and out-group knowledge. Yet this modulation, while to some extent dependent on speakers' individual decisions, is not always available.

In this study, we will use survey data to measure to what extent the context allows multiple construals. Looking more specifically at the two properties that have dominated research on articles, namely Hearer Knowledge (HK) and Specificity of the Referent (SR), we explore whether these variables play a supporting or inhibiting role in allowing speakers to exploit the affordances offered by grammatical variation. Through the measure of a given context's potential openness to construal, we also identify which contextual elements are more or less modulable, based on speakers' conceptualisation of the situation.

3. Data collection and annotation

The study reported on was conducted on data obtained from online surveys created with the online survey builder Qualtrics. The data were collected between April and May 2018 among native speakers of English. We recruited participants online via the research group's social media, the University newsletter and via leaflets posted on information boards at the University or handed out during events. Our respondents were aged between 17 and 73 (average 35.7), varied in their education from GCSE to postgraduate level of education (GCSE or 6th form level: 18, Some university education but did not complete or is in progress: 44, Undergraduate degree: 39, Postgraduate degree: 80). We had 130 female participants, 48 male participants, and 3 participants who chose not to disclose their gender. In total, 181 participants completed the survey.

The survey comprised 12 texts in total and participants were shown four of these texts chosen at random by Qualtrics. These texts were all online articles on various topics, mostly opinion pieces or commentaries. Some parts of the original texts were omitted so that each text was 230–300 words long and formed a complete story. To prepare the stimuli, each text was divided into smaller chunks of approximately 1-2 sentences with one gap to fill at a time. In each chunk, one article (or zero article) was replaced with […..] and below the chunk were all three options in the form of multiple choice, i.e., *a, an / the / --.* Participants could only choose one of the three options. If the participants changed

their mind after reading further context, they could go back and change their previous responses. Each text featured 12 to 18 article gaps to fill, yielding 171 gaps in total (see the underlying dataset for the complete texts and a summary).

Occasionally, if a sentence was too long or there would be two gaps to fill, it was divided into two parts and the participant would first see the beginning of the sentence. Then, they would see the whole sentence, including the article they chose, in the following question. We illustrate this with the examples below, where participants would be shown (5) first and the next question would feature the article they had chosen, as in (6):

(5) Meanwhile, [.....] study found that most of the heroic characters in their research sample were American-sounding;

(6) a. Meanwhile, a study found that most of the heroic characters in their research sample were American-sounding; only two heroes had [.....] foreign accents.

b. Meanwhile, the study found that most of the heroic characters in their research sample were American-sounding; only two heroes had [.....] foreign accents.

c. Meanwhile, study found that most of the heroic characters in their research sample were American-sounding; only two heroes had [.....] foreign accents.

Participants could choose between three options: “a/an”, “the” or “—“ as a “lack” of article to avoid confusing them with the marker “Ø” which is usually preferred among linguists. For the purpose of this paper, we assume that Ø is an article of English rather than an indication of the lack thereof. This approach lets us compare instances of Ø with the other two articles and makes our explanation clearer. We refer the reader to Sommerer (2018) for a discussion of Ø as a lack of article.

The original stimuli were also manually annotated for certain features, including Hearer Knowledge (HK), Specificity of the Referent (SR), and Set Phrase (Authors, under revision). As presented in Table 1 the Hearer Knowledge variable had two values: either the referent was considered known to the hearer (HK+) or unknown to the hearer (HK-). This was also the case for Specificity where the Referent was either specific (SR+) or non-specific (SR-). As to Set Phrase, we considered a noun phrase a set phrase if it was deemed idiomatic or if the combination of the various elements of the noun phrase were considered to be used together frequently enough to be considered a chunk. Most of these variables are highly dependent on context and were thus annotated accordingly. We discuss our annotation process in more detail in SupMat 1.

Table 1. Variables used for the annotation of the data and their values

Variable	**Values**
Hearer Knowledge (HK)	yes (+HK), no (-HK)
Specificity of the Referent (SR)	yes (+SR), no (-SR)
Set phrase	yes, no

4. Method

In this section, we first summarise how we categorised articles based on our annotation of the data (Section 4.1). Then, we explain how we quantify construal through the measure of Entropy (Section 4.2.). Finally, we explore ways to measure variation among participants both across all questions and across participants (Section 4.3.).

4.1 Data classification

The variables used to annotate our data were based on previous work in the literature, notably Huebner's semantic wheel (1983, 1985) which identified four types, as illustrated in Table 2. Note that this table does not include Set Phrases which we annotated as Type 5 as they do not necessarily fit the usual description of articles. We use these types to differentiate between generics and different types of non-generic referents.

Table 2. Huebner's (1983, 1985) semantic wheel as summarized by Thomas (1989).

	SR+	SR-
HK+	Type 2: referential definites Possible article(s): *the*	Type 1: generics Possible article(s): *the, a/an, Ø*
HK-	Type 3: referential indefinites Possible article(s): *a/an, Ø*	Type 4: non-referentials Possible article(s): *a/an, Ø*

The distribution of our stimuli according to Type and then in terms of SR and HK values individually is presented in Table 3: instances of Type 4 make up about half of all our stimuli and we only have 5 instances of Type 5 or Set Phrases.

Table 3. Number of stimuli per Type of article use

Distribution of stimuli (171)					
Types			**SR/HK**		
Type 1	16	9%	**SR+**	64	37%
Type 2	40	23%	**SR-**	107	63%
Type 3	23	14%	**HK+**	59	35%
Type 4	87	51%	**HK-**	112	65%
Type 5	5	3%			

4.2 Entropy

To quantify the extent to which a context allows the speaker to construe their message freely, we rely on the concept of Entropy, often also called more familiarly – Uncertainty. Entropy plays a key role in Information Theory -- for an overview of the main concepts of IT and its use in morphology, see Milin et al. (2009a); Milin et al. (2009b). Essentially, Entropy mathematically models information transmission (and difficulties associated with it). In simple terms, Information Theory assumes that the probability of an event allows us to determine how accurately a message will be reproduced: accurate reproduction is easy if the message is expected, but hard if the message is unexpected. If there is no choice, then there is no uncertainty (Entropy is zero). With more choices, things naturally become more complicated or uncertain, less so if one choice is much more likely, but maximally uncertain if all choices are equally likely. Entropy expresses this mathematically, as numeric quantity.

Much to the disappointment of linguists, Information Theory does not concern itself with

meaning (Divjak 2019: 89). Information Theory is only concerned with the problem of "reproducing at one point, either exactly or approximately, a message selected at another point" (Shannon 1948: 379). This, however, makes it particularly suitable for our purposes: in order to identify whether there exists a type of context that is more amenable to construal, we can establish to what extent a specific article will be reproduced in a particular context: the more constraining the context, the more one particular article will be expected, and hence the more likely it is to be used or reproduced, leaving less freedom for the speaker to construe the message alternatively. This is pertinent to our general remark about choices and their likelihood.

Quantification relies on the likelihood (probability) of choices or, more broadly, occurrences of events. Mathematically, information is defined as the logarithm of the inverse of the probability that the event will occur:

$$H = -\sum_{i=1}^{n} p_i \times log\ p_i$$

where i is the i-th event, out of n. With all these logarithms and inverses, the concept might appear rather elusive, but we can achieve clarity with a relevant linguistic example: proper nouns and articles.

In English most proper nouns, and more specifically proper nouns that refer to people, do not take an article. For example, if you want to refer to the current British Prime Minister, you can use the NP *Boris Johnson.* The phrase without any article is quite common: the query "* * Boris Johnson" (i.e., *Boris Johnson* with two wildcards before) returns 131,179 hits in the News on the Web (NOW) corpus (Davies 2016). However, out of these 131,179 instances, only 197 contain an article before the noun, these are instances of the string "a/an * Boris Johnson", for example, and they make up less than 0.17% of all instances. It thus follows that the occurrence of the phrase *Boris Johnson* without any article is not surprising at all, and thus uncertainty (Entropy) is low. To the contrary, a phrase such as *A sombre Boris Johnson has put Britain under lockdown* is not expected. This is where the inverse

relationship comes into play: the higher the probability of an event, the less surprising it is (mathematically: $1/Pr(Event)$). Now, if an event's probability were maximal ($Pr(Event) = 1$), suddenly, we would have the paradox that surprise in this case would also be 1 (because $1/1 = 1$). As it would be much more consistent for such an extreme case to reflect no surprise (0), the logarithmic transformation is used: the logarithm of 1 gives 0, an intuitively logical value in this case. For the study of English articles, we have a slightly more complex system of three possible events or outcomes. Information-theoretic principles, however, apply equally well.

The stimuli in our task were seen by between 57 and 64 respondents, who each selected one of three article options (Ø, a/n, the) to fill the gap left after the originally used article had been removed in about 55 stimuli (out of a potential 171). Importantly, the choice originally made by the author of a text is not always the only possible one; often, it is but one option out of many. These options or choices created the basis for our data analytic approach. First, for each stimulus we estimated uncertainty (Entropy), given the respective frequencies of each of the three possible answers. This allowed us to capture how much room for construal there was for a given stimulus. For example, if all participants' answers were identical, that would be reflected in low Entropy, signalling no freedom of choice.

4.3 Variation at participant and stimulus level

For the analysis of participants' individual choices, we devised a straightforward procedure, consisting of four steps. First, for each stimulus we determined the most likely choice (i.e., the dominant value or mode) as the usage-based operationalization of the "norm". That is, we considered as the "norm" the option preferred by the majority of participants for each stimulus.

Next, individual participants' choices for each stimulus were defined in terms of matches (1) or mismatches (0) with the mode. If a participant chose the same article as the majority of the participants, this constitutes a match (1); if not, it is considered a mismatch (0).

We then ran a fully random logistic regression model, using the created binary indicator of Match (with the mode) as dependent variable, and Participants and Items (our stimuli) as random factors. A fully random model is justified by the fact that participants as well as items are *randomly sampled* from the larger populations of all potential participants and items. Crucially, this allows us to draw conclusions about those respective populations of Participants and Items (not just about those individuals and texts that we used in our study). We used the adjustments to the model estimates to group participants and stimuli based on how well they match the mode. Agreement (i.e., match with the mode) could be low, average or high, depending on whether the upper limit on the 95% confidence intervals on the adjustments remained below (low), above (high) zero, or crossed (average) zero.

Finally, we ran Log-Linear Modelling, LLM, which is implemented as one of the base routines in the R statistical computing environment (R Core Team 2021), to predict the dependent variable MatchDominant (match vs. mismatch with the mode) from our set of predictor variables which include QuestionAgreement (levels: low, average, high) and ParticipantAgreement (levels: low, average, high) as well as DominantArticle (Mode, with levels: *Ø , a/an, the*). This method allows us to see to what extent articles differ in how predictable they are and in which situations (low, average or high agreement), that is, which articles have more matches or mismatches with the mode and in what type of stimuli (low, average or high agreement). LMM analyses crossed frequency tables, typically 3-way or higher, and it does not require any particular distributional assumption to be satisfied (cf., Rudas 2018).

5. Results

In this section we will first show how Entropy differs depending on the type of variables that may guide article choice (Section 5.1). Then, we will examine differences in variation among participants and among items (Section 5.2).

5.1 Entropy: constraint and construal

We calculated Entropy over the article choices made by our participants per stimulus. For example, in (7), all 61 respondents who saw the stimulus chose the option "a/n", while in (8) – which is the first sentence of a text - the choices were more equally balanced with, out of the 59 respondents who saw the stimulus, 15 selecting Ø, 20 selecting "a/n" and 24 selecting "the". In (7), Entropy (uncertainty) is low (0.000332193), while in (8), Entropy is high (1.559235916).

(7) Don't be shocked at how gormless students can be (they'd have to be, or they wouldn't cheat, right?). One left the sales receipt from the Essay Mill in his book.
Another sent <?> army of male students pretending to be him to sit his exams, all equipped with fake IDs.

(8) <?> free-school advocate and journalist, Toby Young, recently joined other business executives to co-head the government's initiative, the Office for Students (OfS).

In addition to identifying individual stimuli with low or high Entropy, we also compared Entropy across groups of stimuli that share properties, and more specifically HK and SR[3]. Table 4 contains the mean, median, standard deviation, and range of Entropy values for each of the five Types of the well-known semantic wheel (Huebner 1983, 1985, Thomas 1989)

Table 4. Entropy values (mean, median, standard deviation and range) for each Type

	Mean	**Median**	**Std. Dev.**	**Range**
Type 1	0.75	0.94	0.34	[0.12, 1.28]

[3] Since HK and SR are binary values, we used the tetrachoric correlation coefficient. The correlation coefficient equals 0.66.

Type 2	0.53	0.58	0.34	[0.00, 1.56]
Type 3	0.54	0.55	0.44	[0.00, 1.28]
Type 4	0.76	0.82	0.34	[0.00, 1.49]
Type 5	0.41	0.36	0.35	[0.00, 1.00]

Type 5, the set phrases, has the lowest mean Entropy as well as lowest median and also shows the narrowest range, as expected: set phrases are, by definition, expressions that are rather fixed and do not allow much variation. This is illustrated in (9), where 98% of respondents chose *the*.

(9) Millions enjoy going out at the weekend and 'killing some brain cells' by downing a few drinks. Most of us would assume, especially when feeling tender <u>the morning after</u>, that booze is not good for your brain.

Types 1 and 4 score highest on Entropy while Types 2 and 3 score in-between. Note that Types 1 and 3 had a relatively low frequency in our survey whereas Types 2 and 4 were highly frequent; therefore, frequency does not play a causal role in the differences in Entropy.

In the following sub-sections we will examine more closely the relationship between Entropy and Specificity of the Referent (5.1.1), Entropy and Hearer Knowledge (5.1.2) and finally Entropy and other constraints (5.1.3).

5.1.1 Entropy and Specificity of the Referent

Interestingly, Types 1 and 4 are both negative for Referent Specificity, while Types 2 and 3 are positive. In other words, when the referent is marked as specific, respondents tend to prefer the same article. When the referent is not specific (i.e., the context does not point to a specific referent), there tends to be more variation in the article choices respondents make. See for example (10), where the referent was annotated as SR-: 60.7% of respondents preferred *the* while Ø (the article used in the original text) was chosen by 39.2% of respondents (Entropy = 0.966784729). Note that this sentence is the

very first sentence from that particular text called ‘Foreign accents’.

(10) In many of the cases studied, Ø/the villains were given foreign accents.

It is not too surprising that our respondents should hesitate between Ø and ‘the’ in this situation. Their choice boils down to how they interpret, or construe, the referent. Importantly, however, the situation *allows* choice which “creates” room for uncertainty (Entropy), which is exactly what makes the utterance, and more specifically the article used, *informative* – meaningful and/or important in the communicative sense.

There are indeed at least two ways in which the reference can be construed in (10), depending on how prominent the respondents considered the villains to be. The use of the Ø article, as chosen by the author in the original text, tends to make the reference to villains less expected and the statement may appear more general. When a speaker chooses to use *the* instead, they seem to establish a more direct link between the cases studied (introduced in the first clause) and the villains mentioned in the second clause. With *the*, it appears that the film frame has been activated, and that villains are therefore expected. In (10), the context is not strongly constrained towards an SR+ interpretation as it is not clear from the context whether this is a general statement about villains or about a specific set of villains from a specific set of films (cases studied), which could explain our participants’ choices and the distribution of answers.

As mentioned above, SR+ contexts are less likely to allow different construals: average Entropy for these contexts is 0.54. For example, in (11) below, the context is strongly constrained towards a specific referent interpretation with Entropy at 0.000332193. In this particular case, 100% of our respondents chose *the.* It is indeed very difficult to imagine a situation where the referent “2016 presidential election” could be interpreted as non-specific: the very low Entropy shows that there is hardly any room for alternative construals.

(11) […] all the talk of competency during the 2016 presidential election, qualifications, be they ideological or political, are mere pretexts for their choice of candidate.

The fact that SR+ contexts are more likely to lead to agreement as to what the "preferred" article would be has been taken as an indication that Referent Specificity must be the primary property for article assignment, if not an innate property (Bickerton 1981, 1984). Instead of considering Referent Specificity as the primary property, taking a cognitive linguistic perspective, we hypothesise that Referent Specificity is a fixed property, i.e., the options for construing referent specificity are limited and by and large determined by the context.

5.1.2 Entropy and Hearer Knowledge

Hearer Knowledge, on the other hand, appears to be much more susceptible to construal operations, allowing the speaker to set the value of HK regardless of the context. Let us consider example (12).

(12) In between arguments about McCarthyism and an alleged Remainer bias in academia, many professors responded with grander claims of academic freedom and of the embracing of a wide diversity of opinion in the lecture hall.

The Entropy of (12) is 1.48557958, we originally annotated it as HK-/SR- and the author's choice was the indefinite article. It seems that the choice made by our respondents was a matter of whether they considered that the addressee knew about the existence of a Remainer bias among academics. Our assumption is that if they considered that it was a well-known fact that academics tend to be Remainers, then they chose *the*. If, on the other hand, they consider this to be new information for the addressee, *an* would be their preferred choice (which was the original choice). This decision is not guided by the immediate linguistic context; it is open to our respondents' interpretation of what common ground or common knowledge they have with the addressee. This is particularly interesting

since we have shown in Authors (under revision) that Hearer Knowledge is the most crucial variable when it comes to article choice. Overall, for this referent, our respondents showed a slight preference for the HK+ construal with *the* (46%) whereas 37% chose *an*.

Additional support for our interpretation of Referent Specificity as relatively more constrained by the context than Hearer Knowledge stems from contexts where properties other than HK do not constrain the article choice fully. That is, contexts that clearly signal SR+ may still allow choice between *a/an* and *the* for example (cf. *Most cars have a/the steering wheel on the left-hand side.)*; therefore it is HK that governs the choice between the two articles, and HK is subject to construal, which yields more variation in the choices made by participants. In (13) below, participants disagreed as to which article was the most appropriate. They were shown the sentences here in brackets just before and had to choose an article to go with *study*.

(13) (Now, a new study on alcohol and cognitive decline is being used to suggest that Ø official guidelines on alcohol consumption, already laughably low, should be lowered still further. As with all such claims, some serious scepticism is required.) <?> study, published in the Journal of Public Health claims that alcohol consumption of more than 10 grams per day […]

In this particular example, there is no question as to the specificity of the referent: the context makes it clear that the noun 'study' refers to a specific study (cf. the following clauses which include the journal where it was published and the study's main claim), and this stimulus's Entropy is 0.984406085. Interestingly, we found that more than half of our participants chose *a* here: 57% preferred *a* while 43% chose *the*. As can be seen from the preceding sentences, there had already been mention of "a" study. Whether they chose *the* or *a* thus reflects their interpretation as to whether this was a reference to the study previously mentioned, which was therefore HK+ (*the*) or a reference to a different study that they did not know of at this point, hence HK-, and which would thus require the use of *a*. This example perfectly illustrates how hearer knowledge depends on assumptions

on the part of the speaker. However, not all contexts allow the presupposition of hearer knowledge to be (as) open to interpretation, as in (14) whose Entropy is really low at 0.000332193:

(14) In San Juan Capistrano, California, there is a summer camp for disabled children that pairs each camper with a counsellor who attends to their needs, […]

In (14), the presence of the phrase *there is* constrains the interpretation towards the introduction of a new referent and thus of the article *a* in this context (HK-, singular). All respondents (100%) agreed here that *a* was the preferred article. There were also a handful of cases where a vast majority of respondents agreed on the indefinite article, as in (15) below, where 97% of respondents chose *a*.

(15) At Scripps College in Claremont, California, a publication called The Unofficial Scripps College Survival Guide is made available to all students.

In (15), several elements in the context seem to constrain the HK- interpretation: there has been no mention of a publication before, this is also the first introduction of this particular college and as such, it is hard to expect addressees to know about this publication. While these two examples could be considered to invoke a semantic frame such as "city" or "college" respectively, the following elements (*a summer camp* and *a publication*) are too specific to be prominent or easily accessible elements of the frame. While publications are part of the university frame, the specificity of the publication in (15), indicated by the modifier *called The Unofficial Scripps College Survival Guide,* makes it very difficult to construe as HK+ in this context. Elements such as a town hall or students would be assumed to be part of the frame and could thus take *the* as they are more directly expected/are more saliently part of the frame. Not all towns host summer camps for example.

Through the measure of Entropy, we have been able to ascertain that certain variables are more open to interpretation than others. That is, Entropy offers a measure of how much of a context is open

to construal. Through a more detailed analysis of examples, we have also shown that construal is often related to the activation, or not, of a semantic frame. As we have seen so far, SR+ contexts are less open to construal operations than SR- contexts. Hearer Knowledge, whether positive or negative, is not as fixed a property and seems more open to interpretation than other variables and this interpretation partially depends on whether or not a semantic frame is considered to be activated. Before moving on to a discussion of how participants varied in their individual choices and overall tendencies, we take a brief detour to examine a couple of examples where HK and SR's role in determining article choice is not as straightforward.

5.1.3 Entropy and other constraints

So far, we have seen that SR and HK can be more or less fixed and thus lead to the choice of one article or another. However, some uses of Ø, for example, make it more difficult to decide which of HK or SR is more crucial, or which is modified when Ø is replaced by *the*. For instance, a number of our stimuli were originally instances of HK- and number:plural. Since the plural allows both Ø and *the*, it is the value of HK that decides on the article, and HK is to some extent the speaker's prerogative. Uses of Ø with a plural do not necessarily entail HK-. It is very often the case that Ø and plural are combined for generic statements which are considered to be HK+, as they supposedly refer to an entire group, for example, and thus to the representation of a concept rather than to individual instances of a category, e.g., *Ø Cats sleep about 18 hours a day*. But it is interesting to note that some examples are rather vague as to whether the referent is construed as generic or not. Take example (16): if the referent is considered to be part of a previously activated frame (i.e., construed as such), the article of choice would be *the*, but it would be Ø if the speaker decides to present the referent as either unknown to the hearer and as referring to a particular set of guidelines or as a generic referent, i.e. a reference to the entire category of 'official guidelines'. In (16) (Entropy = 0.747560543), a majority of respondents

(79%) chose *the* over Ø (21%), thus construing the referent (official guidelines) as known to the addressee.

(16) Now, a new study on alcohol and cognitive decline is being used to suggest that Ø official guidelines on alcohol consumption, already laughably low, should be lowered still further. As with all such claims, some serious scepticism is required.

In (16), there are several elements in the context that may activate specific frames: one is the mention of alcohol, which is associated with alcohol consumption and regulations, and the other is the fact that "official guidelines" normally entail guidelines specific to a given legal geographical entity, which different speakers might interpret differently. It is most likely the invoking of these frames that guided participants toward the use of *the*.

There are also cases where a HK+ interpretation is constrained by the context or maybe even the referent itself. In (17) and (18), both noun phrases can only take *the*, which entails that their referents are generally conceived as being HK+. We could even consider them phrases as they are highly frequent collocations.

(17) the degree to which the study group is reflective of the general population, and so on – it's hard to believe this is a problem worth worrying about for the vast majority of drinkers.

(18) a post-adolescent transition from the family to society, the postponement of entering the labour force, and primarily the university posited as job training.

For example, a quick query on the web interface of the NOW Corpus (News on the Web, Davies 2016) shows that the combination "general population" occurs 36,534[4] times in the corpus (which is the

most frequent 'ADJ population' combination in this corpus): out of these 36,534 occurrences, 33,239 are combined with the definite article *the* (90% of all instances). As to *labour force,* which is the fifth most frequent 'NOUN force' combination in the corpus (the American spelling *labor force* comes 6th, the two combined are thus the most frequent 'NOUN force' combination), it is also its use with *the* that is the most frequent in the corpus. Entropy for these stimuli is quite low with 0.208347042 for (17) and 0.124115964 for (18); the majority of participants chose *the* in these two sentences (97% and 98% respectively). There thus seems to be a strong association between the article *the* and both *general population* and *labour force,* thus creating an almost fixed lexical string.

5.2 Variation at participant and stimulus level

Because some contexts appear to be more open to construal operations than others, different participants make different choices. In this section, we explore these differences in more detail, both at participant and item (stimulus) level (in 5.2.1), and we analyse contexts of stimuli where the mode (the article preferred by participants overall) differs from the original article chosen by the author (in 5.2.2).

5.2.1 Participants and items

The goodness-of-fit of our fully random logistic regression model alone was already moderately high. This is expected as random effects often account for a significant part of variation in the dependent variable ($R^2 = 0.39$, on 10,231 datapoints; expected probability of matches of a typical individual of 0.86 vs. observed probability of 0.80). The results are summarized in Figure 1, which represents by-Participant (left panel) and by-Item (right panel) adjustments to the global level of matches/mismatches in a so-called caterpillar plot.

Recall that, based on these results, we categorised both participants and sentences into three categories, given the magnitude of the required adjustments: low, average or high. In effect, this created groups of participants and items with lower, average, and higher agreement (as expressed by the number of matches) with the usage-based norm (the dominant value or mode). The frequencies in those three categories were rather uneven for Participants ($Low = 15\%$; $Average = 81\%$; $High = 4\%$), but more balanced for Items ($Low = 39\%$; $Average = 40\%$; $High = 21\%$). This is visible in Figure 1: only a handful of the most extreme Participant adjustments (i.e., the lowest and the highest of all, on the left panel) remain entirely above zero (i.e., their confidence intervals do not cross zero). So overall, the bulk of participants show a tendency to agree in most cases. It is also obvious that Participants vary much less than Items: quite some number of items, in fact, induced moderate to high disagreement regarding the article chosen: low and average per-sentence agreement, together, make up 80% of all cases.

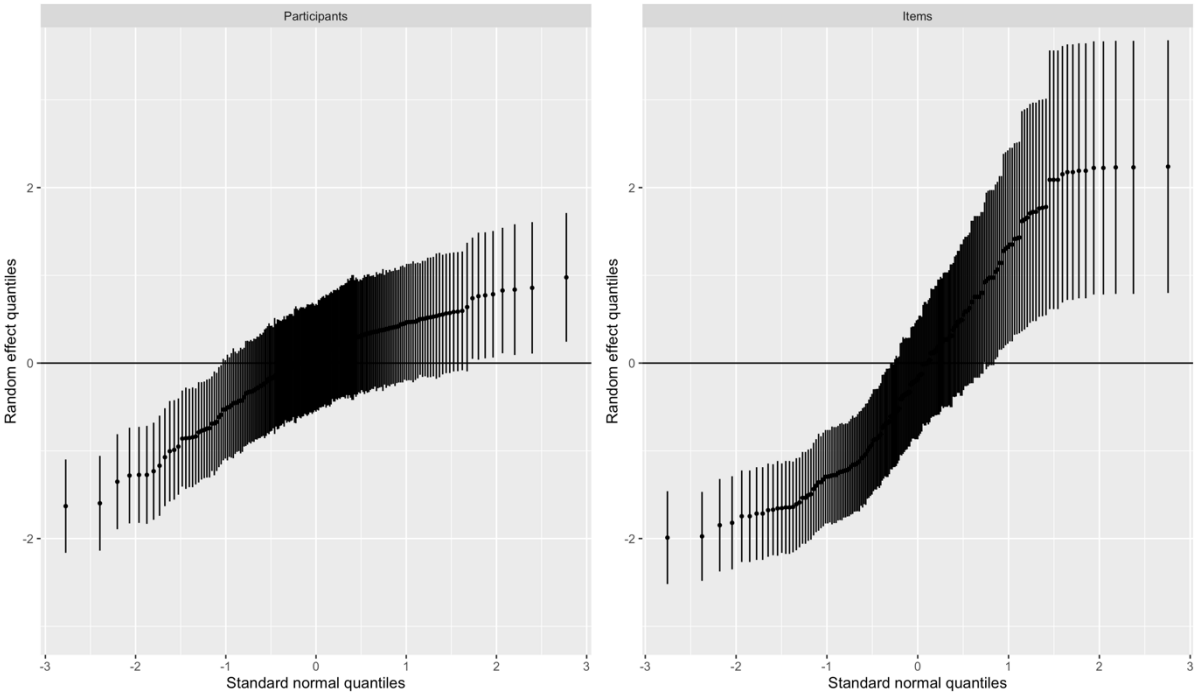

FIGURE 1: by-Participant (left panel) and by-Item (right panel) adjustments to the global level of matches/mismatches

As mentioned in 4.3, and given the properties of our variables, we used Log-Linear Modelling to predict (mis)matches with the mode in various contexts (e.g., high or low agreement among

participants). Given the fact that our data was heavily biased towards matches ($\sim 80\%$), LLM was particularly appealing in comparison with alternatives such as Logistic Regression.

We ran a series of LLMs, and the simplest one that showed a good model fit is the model with two direct effects, of DominantArticle and QuestionAgreement, on Match tallies ($Likelihood\ Ratio = 6.779$; $df = 4$; $p > 0.14$; with reduction of deviance from the non-saturated model of 1499.676: $1506.455 - 1499.676 = 6.779$).[5] The results are summarised in Figure 2, where the X-axis depicts QuestionAgreement and the Y-axis shows Frequency. The coloured bars represent the observed frequencies while the black horizontal lines represent the predicted frequencies.

By inspecting Figure 2 we can see that there are more Matches (left panel) than Mismatches (right panel) overall ($Param_{Match} = 1.05$ vs. $Param_{Mismatch} = -1.05$). The predominantly chosen definite article (*the*) prevails ($Param_{the} = 0.37$), with the indefinite article (*a/an*) remaining close to the expectations ($Param_{a/an} = -0.05$), and a considerably smaller proportion for the zero article ($Param_{0} = -0.32$). Stimuli with low and average agreement are also more frequent ($Param_{low} = 0.95$; $Param_{average} = 0.53$); jointly, they balance out the rare stimuli with high agreement ($Param_{the} = -1.48$).

With respect to Match (left panel), DominantArticle reveals a small match-bias for Ø and *a/an* ($Param_{0 \rightarrow Match} = 0.02$; $Param_{a/an \rightarrow Match} = 0.03$) and similarly a small mismatch-bias for *the* ($Param_{the \rightarrow Mismatch} = 0.05$). QuestionAgreement shows a mismatch-bias (right panel) for low and average agreement ($Param_{low \rightarrow Mismatch} = 0.81$; $Param_{average \rightarrow Mismatch} = 0.11$), and a somewhat more pronounced match-bias for high agreement stimuli ($Param_{high \rightarrow Match} = 0.92$).

Visual inspection of Figure 2 reveals an increase in mismatches (right panel) as QuestionAgreement decreases (right panel). Conversely, for matches (left panel), we see that the three articles are distributed differently: the definite article (*the*) has a rather symmetric distribution, with the majority of counts for stimuli with average agreement; the indefinite article (*a/an*) peaks for

[5] Note that the test-statistics and its p-value reflects how well the model fits the data. Thus, a non-significant p-value indicates a good model.

stimuli with high agreement, and shows little difference between average and low agreement; finally, the zero article has almost no instances in stimuli with high agreement, but it accrues under stimuli with lower (average and low) agreement.

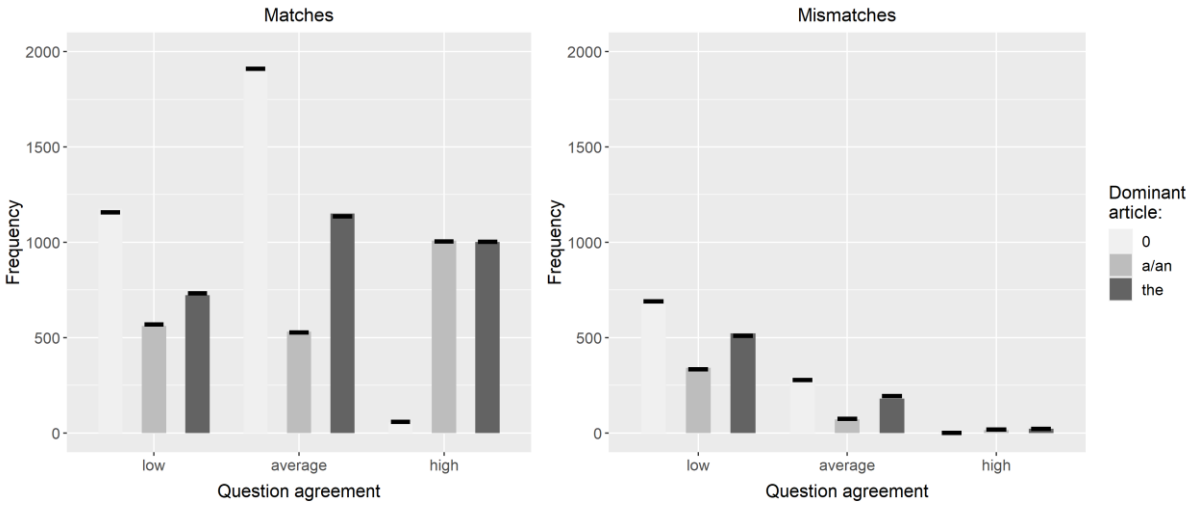

FIGURE 2: Plot of observed (coloured bars) versus predicted (black horizontal line) frequencies for matched and mismatched responses with the dominant article (mode), distributed over the article itself (DominantArticle) and the level of agreement (QuestionAgreement).

5.2.2 Deviation from the mode

Identifying the mode for each stimulus also allowed us to establish on which stimuli the mode differs from the originally used article. There are 29 stimuli on which the mode differs from the originally used article and all 181 participants made at least one deviant choice (i.e., if we subset the stimuli in which the mode differs from the original, we find that all participants do this at some point). Half of these stimuli are stimuli where the original article is Ø and the mode is *the*. As we showed in the examples above and as we argued in section 2.2, the choice of article in these stimuli has to do with frame activation and construal. Another example is given in (19) below. In this particular example, 70% of our respondents chose *the* instead of the original Ø. We annotated the original version as HK- and SR-, which corresponds to Type 4 which has the highest Entropy (see section 5.1). Let us note that out of the 15 sentences where the mode was *the* and the original article Ø, 10 were instances of HK- and number:plural.

(19) (When I used to present Ø programmes on English usage on Radio 4, people would write in and complain about the pronunciations they didn't like. In their hundreds. (Nobody ever wrote in to praise the pronunciations they did like.) It was the extreme nature of the language that always struck me.) Ø Listeners didn't just say they "disliked" something. They used the most emotive words they could think of. They were "horrified", "appalled", "dumbfounded", "aghast", "outraged", when they heard something they didn't like.

We can also compare (19) to (20), which is also a HK-/SR- example and where 51% of respondents chose *the* against 49% for Ø.

(20) The questions after conference papers can be incredibly useful means of identifying Ø flaws in your arguments, or of finding ways to strengthen what you want to say.

Both (19) and (20) were originally HK- and SR- but when participants choose to use *the* instead of Ø, they, in essence, assign a different value to these features. The interpretation of the referent is thus different and the scope of potential referents is narrower (than in the original). That is, in (19), the listeners are not just any listeners, they are the listeners that have been included in the "radio" frame. In (20) the presence or absence of flaws is no longer open to interpretation, with *the,* there are expected to be flaws in arguments.

We find the same type of shift in interpretation (compared to the original) with other instances where the mode differed from the original article. There are in total six questions among the 29 where respondents preferred *the* over the original *a/an*. We present such an example in (21).

(21) (But Ø debates need to be based on factual truth and reasoned assessment,) rather than a desire to be heard loudest.

In (21), 51 % of respondents chose *the* against 49% for *a/an*. Again, when the article is *the*, the referent is then expected to be known to the hearer.

This shows that in all stimuli where the mode differs from the originally used article, the same type of shift in interpretation is at stake, i.e., whether or not the participant considers the referent to be part of the common ground.

6. Discussion

As we argued in the introduction to this paper, on a cognitive linguistic approach to language, articles are considered to be grounding elements, used to anchor referents in the discourse situation. The ground, however, is not fixed and is open to construal. Therefore, it comes as no surprise that article choice also depends on construal. As shown by our survey of 181 native speakers of English, certain contexts are more open to interpretation, and thus the choice of article is less constrained, whereas other contexts contain some elements that will push the choice towards a specific article. We quantified construal through the measure of Entropy, where a context that is more constrained has low Entropy while a context that is more open to interpretation has high Entropy.

Thanks to this measure we were able to see that certain factors make the choice of article more or less favourable to construal operations. For example, we showed that two types of article use had higher Entropy than others (and were therefore less constrained): Types 1 and 4, which are both SR-. To the contrary, SR+ contexts are more constrained towards a particular article (either *a* or *the*) as our respondents tended to agree on the choice of article in these contexts. As to Hearer Knowledge, we

found that none of the two values for this variable were particularly prone to low or high Entropy. HK thus strikes us as less fixed of a feature than SR and unless some specific element in the context clearly constrains HK toward one of the two values, HK is open to interpretation. As we showed in section 5.2, the contexts in which the majority of respondents' choices did not match the original article were cases where HK was amenable to different construals. Out of 29 stimuli where the mode does not match the original article, 12 were contexts in which the key difference is HK modulation. Also, for 15 out of these 29 stimuli, the difference in the choice between *the* and Ø brings out a difference in the interpretation of what is considered as part of the common ground between speakers or activated by a frame. This leads us to conclude that the choice of article depends on what is expected to be part of the knowledge shared by the speaker and their addressee(s) (but cf. Ariel 2001 for related but slightly different conclusions).

Interestingly, Ø is not necessarily constrained in terms of HK, as it can occur with both HK- and HK+; in the latter case it is mostly used for generic reference. As our results show, we also found that Ø is the article that has the most mismatches overall and very few matches when agreement about which article to use given the context is high. As we argued above, the use of Ø with generic reference does not require the prior activation of a specific frame. What is interesting to note is that the use of *the* in contexts where the original article was Ø and the referent was generic does not shift HK entirely, but it does narrow down the scope of potential referents. It is questionable whether this is a true matter of specificity vs. genericity, as *the* can also be used with generics, and as we showed in (13) for example, the choice between *a* and *the* does not depend on specificity either. What seems to matter more is whether the referent is expected or not, which is a matter of construal, notably whether a corresponding frame has been activated or not, and/or whether the referent is considered part of the common ground. It can be argued that a generic use of Ø signals that the addressee is not assumed to expect the mention of the referent. However, when respondents chose to use *the* instead of Ø in such contexts, the difference in interpretation led to an understanding that the addressee was expected to know about the referent.

This expectation, while depending on the respondent's/speaker's construal of the situation, relies on what is considered to be part of the common ground shared by speaker and addressee. Whatever knowledge is part of this common ground is guided by various elements such as potential shared experience, potential shared expertise, activation of a particular frame or even elements present in the context, whether it be physical or linguistic context. As we see from our analysis of individual preferences, this expectation very much depends on individual speakers' interpretation of the situation. While we found that most of our respondents were within the average (i.e., few participants made unusual choices compared to the majority of the group), we also find that the degree of agreement varies depending on the type of context found in the original sentence.

Our argument is thus that the activation of a frame (or not) and the construal of what is considered as common ground between speaker and addressee alter meaning and thus influence grounding and hence the choice of article. It has been argued elsewhere in the general cognitive literature that construal influences/is revealed in the choices speakers make to present situations. We find that this is also true of article use and we offer a way of measuring this by means of Entropy.

7. Conclusion

In this study, we took an empirical approach to mapping out and understanding individual differences in linguistic choice, as illustrated by a case study on English articles. Although the bulk of our participants tended to agree in the majority of cases on which article fit the context best, a non-trivial number of stimuli yielded substantial variation among participants. We argue that this is due to two factors.

First, we have shown that frames play a crucial role in the establishment of reference in the discourse situation and we have argued that the choice of article relies heavily on what speakers construe as being part of their addressee's immediate knowledge. The variation (or lack thereof) observed in the choice of article is a product of a speaker's decision to construe reference differently.

When the context is not strictly constrained towards one specific article, language users show variation in the choices they make, and this variation is the result of the differential activation of one or more semantic frames and the addressee's familiarity with different elements in the frame.

Second, we have proposed a way to measure the context's potential for construal through Entropy where low Entropy corresponds to no or very little variation and high Entropy to high variation. Via the measure of Entropy we were able to identify which types of contexts were more or less likely to be open to variation. Focusing on the two properties that have dominated research on reference, Referent Specificity and Hearer Knowledge (Huebner 1983, 1985; Quirk et al. 1985; Thomas 1989), we found that Hearer Knowledge, which has been considered the property that dominates article selection (Authors, under revision), is in general more open to interpretation, and hence construal, than Referent Specificity. Furthermore, contexts where the referent is specific are less likely to exhibit variation in the choice of article than contexts where the referent cannot be considered specific.

This state of affairs leaves the English article system in the awkward position of being highly dependent on a contextual property that is open to construal. But this finding goes a long way to explaining the variation that exists within the system and sheds light on the intricate nature of the knowledge competent language users must possess to navigate the system: language users need to know not only what the relevant dimensions are that construct the space within which choices need to be made, but also which dimensions are rigid and which ones are soft, available for exploitation in the service of the expression of meaning. Our study of English articles illustrates language users' ability to interpret linguistic properties and use them creatively to express meaning and adapt communication to their target audience, within the constraints imposed by the context. Competent language users thus show their awareness that grammatical structures are meaningful by choosing the structures that best fit their construal of events and situations. As such, it becomes clear that lexical elements gain meaning from the grammatical structures in which they are used. The

establishment of reference and the associated modulation of meaning in language thus appears to rely on elements all along the lexico-grammatical continuum.

Authors' info

Laurence Romain, l.m.y.romain@bham.ac.uk
Dagmar Hanzlíková, dagmar.hanzlik@gmail.com
Petar Milin & Dagmar Divjak, oominds@ooominds.org
All authors' address: Department of Modern Languages, Ashley Building, University of Birmingham, B15 2TS, Birmingham, United Kingdom

References

Ariel, Mira. 1988. Referring and accessibility. *Journal of Linguistics* 24(1). 65-87.

Ariel, Mira. 2001. Accessibility theory: an overview. In Sanders, Ted, Joost Schilperoord & Wilbert Spooren (eds.), *Text representation: linguistic and psycholinguistic aspects*, 29-88. Amsterdam: John Benjamins.

Authors. under revision. Details omitted for blind reviewing.

Bickerton, Derek. 1981. Roots of language.

Bickerton, Derek. 1984. The Language Bioprogram Hypothesis. *Behavioral and Brain Sciences* (173-188).

Chafe, Wallace L. . 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, Charles N. (ed.), *Subject and Topic*. New York: Academic Press.

Clark, Herbert H. 1996. *Using Language.* Cambridge: Cambridge University Press.

Clark, Herbert H. 1997. Communal lexicons. In Malmkjaer, Kirsten & John Williams (eds.), *Language Learning and Language Understanding*, 63-87. Cambridge: Cambridge University Press.

Croft, William & D. Alan Cruse. 2004. *Cognitive Linguistics.* Cambridge: Cambridge University Press.

Dąbrowska, Ewa. 2015. Individual differences in grammatical knowledge. In Dąbrowska, Ewa & Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*, 650-668. Berlin/Boston: De Gruyter.

Davies, Mark. 2016. Corpus of News on the Web (NOW).

Divjak, Dagmar. 2019. *Frequency in language.* Cambridge: Cambridge University Press.

Epstein, Richard. 2002. The definite article, accessibility, and the construction of discourse referents. *Cognitive linguistics* 12(4). 333–378.

Farmer, Thomas, Jennifer B. Misyak & Morten H. Christiansen. 2012. Individual differences in sentence processing. In Spivey, Michael, Ken McRae & Marc Joannisse (eds.), *Cambridge Handbook of Psycholinguistics*. Cambrdige: Cambridge University Press.

Fauconnier, Gilles. 2007. Mental Spaces. *The Oxford Handbook of Cognitive Linguistics*, 351-376. Oxford: Oxford University Press.

Fillmore, Charles J. . 1982. Frame Semantics. In Korea, The Linguistic Society of (ed.), *Linguistics in the morning calm*, 111-137. Seoul: Hanshin.

Givón, Thomas. 1992. The grammar of referential coherence as mental processing instructions. *Linguistics* 30. 5-55.

Horton, William S. & Boaz Keysar. 1996. When do speakers take into account common ground? *Cognition* 59(1). 91-117.

Huebner, Thom. 1983. *A longitudinal analysis of the acquisition of English.* Ann Arbor, MI: Karoma.

Huebner, Thom. 1985. System and variability in interlanguage syntax. *Language Learning* 35(2). 141-163.

Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus, and the mental representations of discourse referents.* Cambridge: Cambridge University Press.

Langacker, Ronald W. 2004. Remarks on nominal grounding. *Functions of Language* 11(1). 77-113.

Langacker, Ronald W. 2008. *Cognitive Grammar: A Basic Introduction.* Oxford: Oxford University Press.

Langacker, Ronald W. 2016. *Nominal Structure in Cognitive Grammar. The Lublin Lectures.* Lublin: Maria Curie-Skłodowska University Press.
Lewis, David K. . 1969. *Convention: A philosophical study.* Cambridge, MA: Harvard University Press.
McCarthy, John. 1990. Formalization of two puzzles involving knowledge. In Lifschitz, Vladimir (ed.), *Formalizing common sense: Papers by John McCarthy*, 158-166. Norwood, NJ: Ablex Publishing.
Milin, Petar, Dusica Đurđević & Fermin Moscoso del Prado Martín. 2009a. The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Language and Memory* 60(1). 50-64.
Milin, Petar, Victor Kuperman, Aleksandar Kostic & Harald Baayen. 2009b. Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In Blevins, James P & Juliette Blevins (eds.), *Analogy in grammar: Form and acquisition*, 214-252. Cambridge: Cambridge University Press.
Mulder, Kimberley & Jan H. Hulstijn. 2011. Linguistic skills of adult native speakers, as a function of age and level of education. *Applied Linguistics* 32. 475-494.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language.* New York: Longman.
R Core Team. 2021. R: A language and environment for statistical computing. 4.1.1 edn. Vienna, Austria: R Foundation for Statistical Computing.
Radden, Günter & René Dirven. 2007. *Cognitive English Grammar.* Amsterdam/Philadelphia: John Benjamins.
Rudas, Tamás. 2018. *Lectures on categorical data analysis.* New York: Springer.
Schank, Roger C. & Robert P. Abelson. 1977. *Scripts, plans, goals and understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Schiffer, Stephen R. 1972. *Meaning.* Oxford: Oxford University Press.
Shannon, C. E. 1948. *The processing of lexical sequences*. University of Alberta PhD Dissertation.
Stalnaker, Robert C. . 1978. Assertion. In Cole, Peter (ed.), *Syntax and Semantics 9: Pragmatics*, 315-332. New York: Academic Press.
Thomas, Margaret. 1989. The acquisition of English articles by first-and second-language learners. *Applied psycholinguistics* 10(3). 335-355.