Definiteness of Noun Phrases with Body Part Terms

Dmitry Levinson

Qualifying Paper

January 23, 2005

1 Introduction

The uniqueness effects in definite noun phrases have been the subject of much debate (Heim 1983, Kadmon 1990, Szabó 2000, Roberts 2003). Some claim uniqueness is part of the semantics of the definite article, while others derive it pragmatically. Both sides assume that the uniqueness condition is categorical, and so the only thing that matters is whether the number of equally salient candidates is one or more.

In this paper I examine the usage of the definite article with body part terms in a number of constructions in different languages. I check how the cardinality of the body parts affects the choice of the article used with the body part terms. The obtained data are important for evaluating of the different theories of definiteness.

The structure of the paper is as follows. Section 2 contains the survey of the previous literature. Section 3 contains a presentation of Spanish data. Section 4 contains a discussion of similar phenomena in English, and section 5 is devoted to related phenomena in Portuguese and Hebrew. The results are summarized and their importance for the theories of definiteness is explained in Section 6. The distinction between semantic and pragmatic definiteness is addressed in Section 7. Section 8 presents another phenomenon, unrelated to body parts, that shows the same influence of the cardinality on the choice of the article. Section 9 contains the conclusions and suggestions for further research. Finally, Appendix A shows a method for estimating the total size of the text accessible through a Web search engine.

2 Previous Literature - uniqueness and definiteness theories

2.1 Uniqueness and Definiteness

Many analyses of the definite article involve the notion of uniqueness. It has been claimed that felicitous use of the definite article requires the referent of the noun phrase to be unique or uniquely identifiable (Russell 1905; Roberts 2003). According to these claims, uniqueness is part of the semantics of the definite article and its presence or absence determines whether a noun phrase is definite or indefinite. In other theories it is familiarity, the presence of the referent within the discourse (Heim 1982), that leads to unique identifiability and using the definite article.

The definite article is usually accompanied by the uniqueness effect, the implication that the intended referent is unique. For example, (1) suggests that there is exactly one clown in the puzzle.

(1) Teacher, giving directions: On the next page, you will find a puzzle. Find the clown in the puzzle. [ example (3) in (Roberts 2003) ]

The following sentence shows that only the definite article is appropriate if the description is semantically unique:

(1) Teacher, giving directions: On the next page, you will find a puzzle. Find the clown in the puzzle. [ example (3) in (Roberts 2003) ]

The following sentence shows that only the definite article is appropriate if the description is semantically unique:
(2) Last weekend we climbed the biggest mountain in West Virginia. [ example (11) in (Roberts 2003) ]

The following question suggests that there was only one inventor of calculus:

(3) Who is the inventor of the calculus? [ = (8) in Szabó (2000) ]

Many theories of definiteness are based in uniqueness, that is, uniqueness is claimed to be a component of definiteness. Such approaches are presented and discussed in the rest of this section.

2.2 Early uniqueness theories

Russell (1905)

An early semantic account of uniqueness is stated in Russell (1905). According to Russell, the existence and uniqueness of the referent of a definite noun phrase are part of the truth conditions of the sentence. For example, the sentence "x was the father of Charles II" asserts not only that x and Charles II were in a certain relation, but also that no individual other than x was in this relation.

Formally speaking, any phrase of the kind "the F is G" is interpreted as "there is one and exactly one F, and this F is G". If the existence and uniqueness condition does not hold, so that there is no F, or more than one F, then any sentence of the kind "the F is G" is false. Therefore, both the sentence "The present king of France is bald" and the sentence "The present king of France is not bald"1 are false.

Strawson (1950)

Strawson (1950) addresses Russell’s (1905) claim that existence and uniqueness of the referent are part of the assertion of the sentence containing the definite description. He agrees that in order for the sentence to be true the existence and uniqueness condition must hold. The disagreement is in the case those conditions do not hold. Discussing the sentence “The king of France is wise”, Strawson argues that the sentence is not false, but rather is neither true nor false, that is, it has no truth value. The view that existence and uniqueness are part of what the sentence presupposes, and not of what it asserts, has since then become the accepted view. It is worth noting that Frege (1892) also described the existence and uniqueness as the presuppositions of using a definite description.

2.3 Explaining the counterexamples

There are many counterexamples to the uniqueness based theories of definiteness. Such counterexamples are felicitous sentences with definite noun phrase, in which the intended referent is not unique. The articles described in this section discuss sentences in which the definite noun phrase contains a non-unique body part term, and propose explanations for such usage. Löbner (1985) tries to save uniqueness by narrowing the domain so that the intended referent is unique in the domain. Both Ojeda (1993) and Epstein (1999) claim that the definite noun phrase does not refer directly to the non-unique object: according to Ojeda, the referent of the singular noun phrase can also be a group of individuals; for Epstein, the referring is to the role of the object in the frame.

1 in the relevant sense; for Russell, this sentence also has the meaning “It is not the case that the present king of France is bald”, in which case it is true.
Löbner (1985)

Löbner’s explanation of definiteness is based on the distinction between sortal and relational nouns (Löbner 1985, p. 292). The difference is that “Sortal nouns classify objects, whereas relational nouns describe objects as standing in a certain relation to others”. Sortal nouns are one-place predicates, e.g., woman, human, female, adult. Relational nouns are “not one-place predicates, but correspond to predicates with two or more arguments (i.e., to relations)”. Examples: husband, wife, daughter, son.

Functional nouns are a subclass of relational nouns. “For functional nouns, the relation that defines their reference is a function. Functions relate objects unambiguously (or one-to-one) to others. Hence, functional nouns are inherently unambiguous, and this semantic property matters in the present discussion. Functional nouns (if they refer at all) always identify a referent. Sortal nouns, in contrast, only classify their referents. Under certain circumstances it may happen that there is exactly one object that fits this classification. But this would be accidental. Functional concepts, e.g. mother of John, do not allow for more than one referent.”

Löbner’s central claim is:

“In all its uses, the definite article has the meaning of indicating that the noun is to be taken as a functional concept.” (p. 314)

Löbner noticed the inadequacy of his explanation. Dual body parts are not functional nouns, yet they do occur with the definite article. He writes “Not all relational nouns, however, are functional nouns…Hand, eye…refer to parts which can occur more than once.” (p. 294-295). His examples (in German) are as follows:

(4) Er legte ihr die Hand aufs Knie. (= (19’’) in Löbner (1985))

He put her the hand on-the knee.

‘He put his hand on her knee.’ (= (19) in Löbner (1985))

(5) Er war der Sohn eines armen Farmers. (= (17’’) in Löbner (1985))

He was the son a-GEN poor-GEN farmer-GEN.

‘He was the son of a poor farmer.’ (= (17) in Löbner (1985))

Löbner’s explanation of this usage is as follows:

“. . .the NPs in these sentences are not really referential, except for the respective subject NPs. . . Instead of referring directly to a particular situation, such statements are, in a sense, generic. They refer to an abstract situation (in the sense of Barwise and Perry (1983)) in which only those objects figure that are explicitly mentioned. Within that abstract situation and, as we shall see, relative to the way the situation is built up, the concepts used are unambiguous. As abstract situations consist essentially of objects and certain relations between them, I shall refer to them as ‘configurations’” (p. 305).

Löbner explains that (5) refers to a “configuration” consisting of just two objects, the father and the son. No other sons of the farmer are present in the configuration. “Hence the relational concept ‘son’ in fact provides a functional link from the father to the son.”

---

2 I think this is a mistake. It should be many-to-one.
The explanation for (4) is similar: only one hand and knee are present in the configuration.

This explanation is hard to accept, because Löbner gives no criteria for distinguishing between the objects included in the configuration and the ones excluded from it. For example, although the referent of the NP *the apple* in (6) might be the only one in the configuration, the sentence is infelicitous if the apple was not previously mentioned.

(6) *I am eating the apple.*

**Ojeda (1993)**

Ojeda (1993) addressed the problem of definite noun phrases denoting non-unique body parts. He gives the following examples (for (8)-(10) it is assumed that noun phrases with possessive pronouns are definite):

(7) John was hit on the arm. [= (1) in Ojeda (1993)]

(8) Claire skinned her knee. [= (2) in Ojeda (1993)]

(9) Susan pinched her finger. [= (3a) in Ojeda (1993)]

(10) I always kiss my mother on the cheek when I go home. [= (11a) in Ojeda (1993)]

Such usage is also found with non-body part NPs:

(11) Bill broke the headlight on his car. [= (3b) in Ojeda (1993)]

(12) Ann had to pick her son up at the airport. [= (3c) in Ojeda (1993)]

Ojeda deals with this issue by changing the assumed theory of the noun phrases. Instead of the assumed view that a noun denotes a set of individuals, he proposes that a noun denote a *pairwise disjoint* set of groups of individuals. For example, instead of assuming the following denotation of the noun *arm*:

(13) \[ [[\text{arm}]] = \{ x \in U \mid x \text{ is an arm of some individual in } U \} \]

he proposes the following:

(14) \[ [[\text{arm}]] = \{ x \in U \mid x \text{ is the group of arms of some organism or device in } U \} \] [=his(5)]

Then the definite description in (7)-(10) is explained as referring to the unique group of the body parts, i.e., (7) refers to John’s unique group of arms and so on.

This analysis might be reasonable for the examples given by Ojeda. Indeed, John’s being hit on the arm means his being hit on some part of the arm, and this part is also part of his group of arms. As Ojeda states it, “being hit on an armgroup does not entail being hit on each one of the arms which constitute it” (p. 249). However, this analysis does not seem to give correct predictions in sentences like the following:

(15) John raised his arm.

If in this case *his arm* referred to the unique group of his arms, the sentence (15) as a whole would mean that John raised both his arms, which is clearly not the case.

**Epstein (1999)**

Epstein discusses a number of cases in which a definite noun phrase contains a non-unique description. In some cases the noun is a body part term.
(16) Lee kissed Pat on the cheek.
(17) The dog bit him on the finger.
(18) She put her hand on his knee.
(19) I broke my toe.
(20) He was the son of a poor farmer.
(21) I read this in the paper.
(22) Waiter, I demand to see the menu.

Epstein's claim is that the definite noun phrases refer to unique ‘role’ functions, rather than to any of the non-unique values of the roles. The role referred to is salient within a frame. For example, in (22) the menu is a salient role within the 'restaurant' frame. This is somewhat similar to Löbner's (1985) claim, discussed in Section 2.3, that when a noun is used with the definite article, it denotes a functional concept.

However, in most cases of roles discussed by Epstein, there is typically one object belonging to the role in any particular frame. If there is obviously a number of possible objects satisfying the role in a particular frame, the definite article seems inappropriate. Epstein discussed Birner and Ward's (1994) example of a dinner table with three baskets of rolls. "As they point out..., in such a situation it is possible to say Pass the rolls (with a plural noun) but it would be strange to use the singular form Pass the roll (even if the speaker wants only one roll and it does not matter which one)" (p. 127). His explanation is that the 'dinner table frame' contains a salient, uniquely identifiable role for a set of rolls, but not for an individual roll. That is, the table typically has a basket of rolls, but not an individual roll.

With body parts, the situation is different. It would be strange to say that a person typically has a finger, or a hand. A person typically has 10 fingers and two hands, and this is an aspect in which the body parts differ from the other cases discussed by Epstein. This difference makes the claim that the noun in a definite noun phrase refers to a role less plausible.

2.4 DRT-based refinement of uniqueness - Kadmon (1990)

In response to Heim (1982, 1983), Kadmon (1987, 1990) defends the view that uniqueness is always present when the definite article is used, and claims that every felicitous use of a definite NP must satisfy the Uniqueness Condition she formulates.

Kadmon’s analysis is based on Cooper (1979) and is expressed along the lines of the Discourse Representation Theory (DRT) proposed in Kamp (1981). Kadmon’s Uniqueness Condition is that “…the variable of a definite NP must be unique relative to the set of variables that are bound higher up” (p. 293). This is formally expressed by the generalized version of the condition shown in (23)

(23) \[ B_K := \{ X : \exists K' \text{ accessible from } K \text{ s.t. } K' \neq K \text{ and } X \in U_K \} \]

The Uniqueness Condition

Let \( \alpha \) be a definite NP associated with a variable \( Y \), let \( K_{loc} \) be the local DRS of \( \alpha \), and let \( K \) be the highest DRS s.t. \( K \) is accessible from \( K_{loc} \) and \( Y \in U_K \). \( \alpha \) is used felicitously only if for every model \( M \), for all embedding functions \( f, g \) verifying \( K \) relative to \( M \),
\[
\text{if } \forall X \in B_K \ f(X) = g(X) \text{ then } f(Y) = g(Y). \text{ (= (31) in Kadmon (1990))}
\]

This can be informally described as follows. \(Y\) is the variable associated with the definite phrase to which the condition is applied. \(K\) is the DRS (Discourse Representation Structure) in which \(Y\) was introduced. \(B_K\) is the set of variables on which the definite depends; it contains variables that are bound higher than \(Y\). The uniqueness requirement is relative to \(B_K\): “\(Y\) must be unique relative to a choice of values for all the variables in \(B_K\)” (p. 293). This allows to explain examples like the following (p. 292):

(24) Every chess set comes with a spare pawn. It is taped to the top of the box.

In this case the intended spare pawn is unique relative to the chess set, and \(B_K\) for the NP it contains the variable representing the chess set.

2.5 A hierarchical approach - Gundel, Hedberg and Zacharsky (1993)

Gundel et al. (1993) describe the use of the different determiners by the cognitive status required for each determiner. The statuses are ordered in the Givenness Hierarchy given in Table 1.

<table>
<thead>
<tr>
<th>in focus &gt;</th>
<th>activated &gt;</th>
<th>familiar &gt;</th>
<th>uniquely identifiable &gt;</th>
<th>referential &gt;</th>
<th>type identifiable</th>
</tr>
</thead>
<tbody>
<tr>
<td>it</td>
<td>that</td>
<td>this N</td>
<td>the N</td>
<td>indefinite</td>
<td>a N</td>
</tr>
</tbody>
</table>

Table 1. The Givenness Hierarchy of Gundel et al. (1993).

According to Gundel et al. (1993), the definite article can be used felicitously only if the referent is uniquely identifiable. As they state it,

UNIQUELY IDENTIFIABLE: The addressee can identify the speaker’s intended referent on the basis of the nominal alone. This status is a necessary condition for all definite reference, and it is both necessary and sufficient for the appropriate use of the definite article \(\text{the}\). (p. 277)

2.6 More counterexamples - Birner and Ward (1994)

Birner and Ward (1994) give some examples in which the definite NP has a non-unique referent. For example,

(25)[At a table containing four pitchers of milk, all equidistant from the hearer] Please pass the milk. (= (10) in B&W)

According to Birner and Ward, the use of the definite article in this case is possible because the differences between the units of milk are not relevant. As they explain, “...the relevant unit of milk...is not relevantly differentiable from any other unit of milk present. If the units are individuated in some relevant way, the non-unique definite reference becomes infelicitous.”

This is usually not allowed with singular non-mass nouns:

(26)[At a table containing four baskets of rolls] # Please pass the roll (= (11) in B&W)
However, there is a class of singular nouns that can be used with the definite article without requiring uniqueness. For example,

(27) a. As soon as my cousin arrived in Santiago, she broke her foot and had to spend a week in the hospital. (= (13) in B&W)

b. Your 10:00 appointment - a Mr. Johanson - said he’d be late because he had to stop at the bank first.

c. My history professor announced to the class today that he wasn’t going to give us a final. He said that, while waiting in line at the grocery store, he realized that he already had enough information to assign us a grade.

The definite NPs refer to non-unique entries. The explanation given for (25) is valid here as well: “…these NPs are used to refer to locations that are not relevantly differentiable from other locations denoted by the same NP.”

Interestingly, adding a modifier results in infelicity:

(28)# While in Santiago, Bill broke his foot and was rushed to the big hospital.” (= (14) in B&W)

In route directions, nouns denoting objects moving along a regular constant path allow non-unique definite usage. In these cases the noun can be said to refer to the unique route. This does not happen with a vehicle in which the person has a control over the route. Examples:

(29) [Hotel concierge to guest, in a lobby with four elevators] You are in Room 611. Take the elevator to the sixth floor and turn left. (= (19) in B&W)

(30) To get to Dr. Smith’s office, I suggest taking the stairs/bus/train/*car/*bike/*taxi. (= (20)-(21) in B&W)

2.7 Approaches based on notions other than uniqueness

Familiarity and File Change Semantics

Heim (1982, 1983) proposed a theory of definite and indefinite noun phrases called File Change Semantics. It is based on the notion of familiarity used by Christophersen (1939). The metaphor Heim uses for modeling the interpretation of the utterances is a mental file containing cards. A card contains information about a particular object. The difference between definite and indefinite noun phrases is in their relation to the cards: an indefinite noun phrase creates a new card; a definite noun phrase refers to an already existing one. These conditions are called, respectively, Novelty and Familiarity. In Szabó’s (2000) formulation,

(31) Novelty: For every indefinite description, start a new card.

(32) Familiarity: For every definite description, if there is an appropriate old card in the file, update it; otherwise start a new card.

Heim does not stipulate uniqueness as a necessary condition for using a definite description. The uniqueness presuppositions are explained by the following update constraints, (again, in Szabó’s formulation):

(33) Non-redundancy: When filing an utterance, don’t create redundancy.
(34) Non-arbitrariness: When filing an utterance, don’t make arbitrary choices.

The non-redundancy constraint limits the usage of the indefinite noun phrases; without this constraint indefinites would be felicitous in any case. The non-arbitrariness constraint is the one explaining the uniqueness presupposition of the definite noun phrases; indeed, in order for the choice of the card not to be arbitrary, it must be known to the hearer which card to choose; that is, the intended object must be uniquely identifiable. These constraints can be seen as a special case of Grice’s (1975) conversation maxims.

*Lyons (1999)*

Lyons (1999) distinguishes between the definiteness as a grammatical notion and the semantic/pragmatic definiteness (p. 274). The grammatical definiteness is a grammaticalization of the semantic/pragmatic definiteness, but "the correspondence between a grammatical category and the category of meaning it is based on is never one-to-one." This means that not all the cases of definiteness can be explained by the same semantic/pragmatic notion. This is an expected result of grammaticalization, as the newly created grammatical category can acquire new usages.

Lyons discusses whether the grammatical definiteness is a grammaticalization of *identifiability* or *uniqueness/inclusiveness*. According to the identifiability condition, the definite article signals that the hearer can identify the referent of the noun phrase (p. 5-6). It is a refining of the familiarity condition, stating that the definite article signals that the denoted entity is familiar (p. 3). According to Lyons, "This view of definiteness does not altogether reject familiarity. Rather, familiarity, where it is present, is what enables the hearer to identify the referent."

Lyons gives two reasons in favor of identifiability (p. 278). First, the semantic/pragmatic notion of definiteness plays a role even in the languages with no overt marking of definiteness. In these languages it is the identifiability that matters, and not uniqueness. Second, demonstratives and personal pronouns are typically treated as definite, and this is better explained by identifiability rather than uniqueness.

*Roberts (2003)*

Roberts (2003) adopts Heim’s familiarity condition for the use of definite noun phrases. However, the familiarity she proposes is different from the Heim’s view of familiarity, requiring the referent to be mentioned or known by the hearer. Instead, Roberts proposes the notion of *weak familiarity*. An entity is *weakly familiar* if its existence is entailed in the context (p. 298). For Heim, the felicity of *it* (assumed to be a definite noun phrase) in (35) is explained by accommodation, since it has no prior familiar discourse referent. For Roberts, the referent is weakly familiar because of the information introduced in the first sentence, and no accommodation is required.

(35) Every motel room has a copy of the Bible in it. In this room, it was hidden under a pile of TV Guides. [ =(15) in Roberts (2003)]

The uniqueness effect is explained by Gricean principles (Grice 1975), similar to the way it is explained within Heim’s approach.

2.8 *Definiteness theories and body part terms*

Typically, body part terms are not given special treatment in uniqueness-based definiteness theories. Therefore, Russell (1905), Strawson (1950), and Kadmon (1990)
predict that the definite article would only be used in the case the referent is unique. This would happen with unique body part terms (head, nose), and with non-unique body parts terms (hand, finger), when the referent is salient in the context. The definite article is not supposed to occur with a non-unique body part term when the referent is not salient; different non-unique body part terms are predicted to have the same behavior.

Löbner (1985), Ojeda (1993), and Epstein (1999) predict that the definite article can always be used with the body part terms. According to Löbner, the body part referred to is unique in the situation of the utterance. According to Ojeda and Epstein, the body part noun phrase does not refer to a particular body part token, but rather to the uniquely identifiable body part group (Ojeda 1993), or role (Epstein 1999). Therefore, the definite article can always be used in a body part noun phrase. These analyses make no distinctions between the different body part terms, so the predictions for the different body part terms are the same.

In familiarity-based approaches (Heim 1982, Szabó 2000, Roberts 2003) the uniqueness effect is explained by Gricean principles. For example, in Heim (1982) it is formalized as the non-arbitrariness condition. If this condition is taken categorically, the prediction is still that the definite article should be used only in the case the referent is unique, similar to the predictions of the uniqueness-based theories. However, if the condition is taken to be gradual, the prediction would be that the definite article would be used with the unique body part terms and those of the non-unique body part with which the arbitrariness of the choice, and the chance for a mistake in identification, is low. For example, the usage of the definite article should be higher with dual body parts (hand) than with those whose number in the human body is higher (finger). Lyons' (1999) explanation of definiteness as grammaticalized identifiability has similar predictions: the higher the number of the body part in the human body, the lower is the identifiability of the intended referent.

The following sections present the data of the actual use of the articles with body part terms. Comparison between the usage data and the predictions of the different definiteness theories is performed in Section 6.4.

3 Definiteness of Body Part Terms in Spanish

3.1 Body parts in previous literature and in this paper

Most examples in the literature include names of body parts that are either unique (head, nose) or dual (hand, leg) in the human body (Demonte 1988, Kliffer 1983, Kliffer 1984). The body part terms discussed in this paper and the cardinality of the body parts they denote are listed in Table 2. They include the words finger and tooth, denoting body parts with cardinality higher than two.

<table>
<thead>
<tr>
<th>Name</th>
<th>Cardinality</th>
</tr>
</thead>
<tbody>
<tr>
<td>nose, head</td>
<td>1</td>
</tr>
<tr>
<td>hand, arm, leg, foot, eye, ear</td>
<td>2</td>
</tr>
<tr>
<td>finger</td>
<td>10</td>
</tr>
<tr>
<td>tooth</td>
<td>up to 32</td>
</tr>
</tbody>
</table>

Table 2. Body part terms discussed in this paper
3.2 Preliminary discussion: native speakers judgments

In this section I present the results of a preliminary examination of the articles used with body part terms in Spanish. I asked native speakers of Spanish to construct sentences with the verb *doler* 'hurt' reporting a new pain in a particular body part. This made it possible to obtain comparable data for different body parts. The results were as follows:

For unique body parts, only the definite article can be used (36). For dual body parts, the definite article is preferred (37). For fingers, both the definite and the indefinite article can be chosen (38). Interestingly, if the finger is named so that there are only two body parts with that name, the definite article is preferred (39). For teeth, the definite article suggests that the tooth that hurts was salient in some way (40).

(36) *Me duele la/*una cabeza/nariz.*

Me hurts the/a head/nose.
‘My head/nose hurts.’

(37) *Me duele la/*una mano.*

Me hurts the/a hand.
‘My hand hurts.’

(38) *Me duele el / un dedo.*

Me hurts the / a finger.
‘My finger hurts.’

(39) *Me duele el / un dedo índice.*

Me hurts the / a finger index.
‘My index finger hurts.’

(40) *Me duele #el / un diente.*

Me hurts the / a tooth.
‘My tooth hurts.’

3.3 Corpus experiment: data

In order to get more precise data regarding the usage of the articles with the body part terms, I conducted a corpus search. The search was performed on the World Wide Web using the Google search engine. I chose to use the Web instead of the traditional corpora because of the difference in the corpus size. The total size of the Spanish web pages accessible to Google is to the order of magnitude of 10 billion words (see Section 0 for the description of the technique used to appraise this number). For comparison, one of the biggest Spanish corpora, www.corpusdelespanol.org, contains 100 million words, i.e. 100 times less.

I looked for the phrases of the template "*duele* 'hurts' + article + body part term" and "*se rompió* 'he/she broke' + article + body part term". Some corpus examples:

(41) Si te *duele el diente* insoportablemente, un endodoncista puede quitarte el dolor sin quitarte el diente. [www.amad21.org/Documentos/Articulos/Endodoncia.html]
‘If your tooth hurts unbearably, an endodontist can remove the pain without removing the tooth.’

(42) ¿Si me duele un diente debo tomar sólo analgésicos?
[http://www.geocities.com/sixtodont/sixto.htm]

‘If my tooth hurts, do I only have to take analgesics?’

Table 3 and Table 4 below contain the results of the corpus search. In some cases the combination “article + body part term”, e.g., la mano ‘the hand’, were part of a larger uniquely identifying noun phrase, such as la mano derecha ‘the left hand’ or la mano izquierda ‘the right hand’. In this case the definite article is obligatory. I excluded such cases from the results.

<table>
<thead>
<tr>
<th>Body part</th>
<th>Occurrences with the indefinite article</th>
<th>Occurrences with the definite article</th>
<th>Percentage of indefinite use</th>
</tr>
</thead>
<tbody>
<tr>
<td>cabeza</td>
<td>1</td>
<td>808</td>
<td>0%</td>
</tr>
<tr>
<td>nariz</td>
<td>0</td>
<td>68</td>
<td>0%</td>
</tr>
<tr>
<td>mano</td>
<td>5</td>
<td>206</td>
<td>2%</td>
</tr>
<tr>
<td>brazo</td>
<td>6</td>
<td>172</td>
<td>3%</td>
</tr>
<tr>
<td>pierna</td>
<td>20</td>
<td>152</td>
<td>12%</td>
</tr>
<tr>
<td>oído</td>
<td>13</td>
<td>70</td>
<td>16%</td>
</tr>
<tr>
<td>pie</td>
<td>29</td>
<td>138</td>
<td>17%</td>
</tr>
<tr>
<td>ojo</td>
<td>27</td>
<td>112</td>
<td>19%</td>
</tr>
<tr>
<td>dedo</td>
<td>18</td>
<td>80</td>
<td>18%</td>
</tr>
<tr>
<td>diente</td>
<td>47</td>
<td>30</td>
<td>61%</td>
</tr>
</tbody>
</table>

Table 3. Use of articles in the template “duele ‘hurts’ + article + body part term”.

<table>
<thead>
<tr>
<th>Body part</th>
<th>Occurrences with the indefinite article</th>
<th>Occurrences with the definite article</th>
<th>Percentage of indefinite use</th>
</tr>
</thead>
<tbody>
<tr>
<td>cabeza</td>
<td>1</td>
<td>79</td>
<td>1%</td>
</tr>
<tr>
<td>nariz</td>
<td>0</td>
<td>89</td>
<td>0%</td>
</tr>
<tr>
<td>mano</td>
<td>26</td>
<td>43</td>
<td>37%</td>
</tr>
<tr>
<td>brazo</td>
<td>103</td>
<td>111</td>
<td>48%</td>
</tr>
<tr>
<td>pierna</td>
<td>273</td>
<td>171</td>
<td>61%</td>
</tr>
<tr>
<td>dedo</td>
<td>70</td>
<td>37</td>
<td>65%</td>
</tr>
<tr>
<td>diente</td>
<td>39</td>
<td>4</td>
<td>91%</td>
</tr>
</tbody>
</table>

Table 4. Use of articles in the template “se rompió ‘he/she broke’ + article + body part term”.

Figure 1 shows the percentage of the indefinite article usage with each of the constructions.
3.4 Corpus experiment: discussion

The usage of the indefinite article with the *doler* construction is as follows:

1. *mano* ‘hand’ and *brazo* ‘arm’: used almost exclusively with the definite article
2. *pierna* ‘leg’, *pie* ‘foot’, *ojo* ‘eye’, *oído* ‘ear’, *dedo* ‘finger’ : 10%-20% of the sentences have the indefinite article.
3. *diente* ‘tooth’: more than 50% of the sentences have the indefinite article.

In this case, the major trend in the data is due to *cardinality*: the higher the cardinality, the more frequent is the indefinite article. However, another factor intervenes. Why are *mano* ‘hand’ and *brazo* ‘arm’ less used with the indefinite article than the other dual body parts, while *dedo* ‘finger’ has approximately the same percentage as the dual body parts?

A possible explanation would be the different *visual saliency* of the body parts. Hands, arms and fingers are close to the speaker’s view. On the other hand, feet, legs, eyes and ears are further from the speaker. Both factors affecting the article frequency in this case, cardinality and visual saliency, can be subsumed by the notion of *distinguishability*, the ability to *individuate* (Birner and Ward 1998) the intended referent from the others. The distinguishability is higher for body parts with lower cardinality and for those that are more visually salient (‘hand’ and ‘arm’ vs. the other). The higher the distinguishability, the less probable the choice of the indefinite article.

In this case, and in the other cases discussed in this paper, I checked the influence of the cardinality and the visual saliency of the body part term on the indefinite article usage. This was done by performing a linear regression analysis with SPSS. The cardinality and the visual saliency were the independent factors, and the article used in each case was the dependent variable. For Spanish, both factors have a statistically significant contribution; for a detailed description of the statistical analysis, see Section 6.1.

4 Definiteness of Body Part Terms in English

For English, I have checked the sentences with the body part terms as a direct object of the verbs *raised* and *broke*. These verbs were among the most frequent cooccurring

Figure 1. Percentage of indefinite article usage with body part terms in some Spanish constructions.
with different body part terms in the British National Corpus. Since the size of the searchable English Web is much larger than the Spanish, I was able to obtain results for different subjects separately.

For English, the variation of the determiner with the verb break is not between the definite and indefinite article, but rather between the indefinite article and the possessive:

(43) I broke a/*the/my finger.

Some corpus examples:

(44) Well, to start with, I broke my finger, but the riot cops were really nice and helpful about it. [http://www.mathnews.uwaterloo.ca/Issues/mn8704/riotcops.php]

(45) I broke a finger playing basketball. [http://www.childbirthsolutions.com/articles/daddy/fingertest.php]

I checked the determiners used with the different body part terms in the construction "I broke + determiner + body part term". The results are in Table 5 below.

<table>
<thead>
<tr>
<th>Body part</th>
<th>Indefinite percentage</th>
<th>Definite percentage</th>
<th>Possessive percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>hand</td>
<td>1% [13]</td>
<td>4% [37]</td>
<td>95% [946]</td>
</tr>
<tr>
<td>arm</td>
<td>2% [114]</td>
<td>1% [82]</td>
<td>97% [5600]</td>
</tr>
<tr>
<td>leg</td>
<td>4% [301]</td>
<td>1% [57]</td>
<td>95% [6960]</td>
</tr>
<tr>
<td>foot</td>
<td>1% [21]</td>
<td>2% [41]</td>
<td>97% [2270]</td>
</tr>
<tr>
<td>finger</td>
<td>29% [277]</td>
<td>2% [19]</td>
<td>69% [665]</td>
</tr>
<tr>
<td>toe</td>
<td>18% [194]</td>
<td>3% [34]</td>
<td>78% [829]</td>
</tr>
<tr>
<td>tooth</td>
<td>73% [901]</td>
<td>1% [13]</td>
<td>26% [326]</td>
</tr>
</tbody>
</table>

Table 5. English: percentages of determiner usage for sentences of the kind (43). [In brackets: the number of occurrences]

As can be seen from the table, the major variation is between the indefinite article and the possessive pronoun. I will consider the noun phrases with possessive pronouns to be definite (as is generally assumed, see Abbott (2004)), so the variation is similar to the Spanish one.

The sentences with the definite article that were found demonstrate the imperfection of the search method. They are of a different structure than the construction I am checking. Some reasons for using the definite article are relative clauses (46) and usage of the word hand as a noun modifier (47).

(46) After a fatal car accident, in which I broke the hand I write with in about seven places, I never stopped writing.

(47) I broke the hand mirror.

However, it can be seen from Table 5 that such sentences appear in a similar small amount (less than 5%) with all the body part terms, and their major trends in the data are not influenced by them.

Figure 2 below contains the data of Table 5; it also contains the corpus data for phrases with he and she as a subject.
The percentage of the indefinite article usage is substantially higher for higher cardinality body parts, and the results are statistically significant. For a detailed discussion of the statistical analysis, see Section 6.1.

The results for the verb *raised* are in Figure 3. In this case as well as with the other verb, the indefinite article is more frequent with the high cardinality body part term.

In addition, I checked sentences with the *possessor raising* construction. The article in this construction shows no variation, and the definite article is used with all the body part terms. A possible explanation is that in this case the definite article became grammaticalized.

Examples:

(48) The dog bit him on the/*a finger.

(49) I kicked him on the/*a knee.

The data on the article usage with the "*bit him on* + determiner + body part" is in Table 6 below. The indefinite usage is almost nonexistent.
Table 6. English: determiner usage for sentences with "bit him on DET BP".

For comparison, the English data for different constructions in concentrated in Figure 4 below.

![Figure 4](image-url)

Figure 4. Percentage of indefinite article usage with body part terms in English sentences.

5 Other languages

5.1 Brazilian Portuguese

The structure of the Brazilian Portuguese sentences I checked is similar to Spanish:

(50) *Me dói a/\*uma mão.*
    Me hurts the/a hand
    ‘My hand hurts.’

(51) *Quebrou o/\*um braço.*
    Broke-3sg the/a arm.
    ‘He broke his arm’ or ‘She broke her arm’

The Web search results are summarized in Table 7 and Figure 5.

<table>
<thead>
<tr>
<th>Body part name</th>
<th>Portuguese: dói 'hurts'</th>
<th>Portuguese: <em>quebrou</em> 'broke-3SG'</th>
</tr>
</thead>
<tbody>
<tr>
<td>mão</td>
<td>0% [0/21]</td>
<td>1% [1/97]</td>
</tr>
<tr>
<td>braço</td>
<td>0% [0/23]</td>
<td>17% [66/390]</td>
</tr>
<tr>
<td>perna</td>
<td>---</td>
<td>21% [128/592]</td>
</tr>
<tr>
<td>ouvido</td>
<td>0% [0/32]</td>
<td>---</td>
</tr>
<tr>
<td>pé</td>
<td>4% [1/23]</td>
<td>2% [6/384]</td>
</tr>
<tr>
<td>olho</td>
<td>0% [0/26]</td>
<td>---</td>
</tr>
</tbody>
</table>
Table 7. Percentage of indefinite article usage with body part terms in Portuguese constructions.

Although the tendency is the same, namely, higher usage of the indefinite article with higher cardinality body part terms, the percentages are much lower than in Spanish. It is possible that in the Portuguese construction the definite article is becoming grammaticalized.

Figure 5. Percentage of indefinite article usage with body part terms in Portuguese constructions.

5.2 Italian

For Italian, I checked usage of the articles in sentences of the template “mi fa male + article + body part term”, i.e. (52):

(52) Mi fa male la mano.

Me does bad the hand.

‘My hand hurts’

The results are in Table 8.

Table 8. Italian: article usage in sentences like (52)
The statistically significant correlation between the percentage of the indefinite article use and the cardinality of the body part was observed for the Italian example as well.

5.3 Hebrew

The only article existing in Hebrew is the definite article ha-. Indefiniteness is expressed by a noun phrase without any article. For Hebrew, I checked sentences of the template “koevet ‘hurts’ li ‘me’ + (ha-) + body part term”, e.g., (53).

(53) Koevet li (ha-)yd / (ha-)regel / (ha-)ecba/ (ha-)šen.
    Hurts me (the)hand / (the)leg / (the)finger / (the)tooth.
    ‘My hand/leg/finger/tooth hurts’

The intuitions of native speakers I asked were that the definite article was obligatory for all the body part terms. The corpus search supported these intuitions: almost in all the sentences I found the definite article was present. The only sentences without the indefinite article were in the antecedent of a conditional/quantificational phrase. I checked the intuitions of the native speakers on corpus sentences. I deleted the definite article from the sentences in which it was present and asked two native speakers of Hebrew to insert the article whenever appropriate. The native speakers made the same choices, using the definite article in all but quantificational contexts. One sentence from the Web contains the verb in the past, and the noun was without the article. Both native speakers I consulted thought that it’s better to use the article. Since there were considerably less sentences found for Hebrew than for the other languages, I do not present quantitative data.

Interestingly, sometimes rephrasing the sentence changes the optionality of the article. In (54) the article is optional. In (55) the definite article can only be used if the tooth was already salient in the context.

(54) Nišbera li (ha-)šen.
    became-broken me (the)-tooth
    ‘My tooth has broken’

(55) Šavru li (*et ha-)šen.
    broke-3Pl me (ACC the) tooth.
    ‘They broke my tooth’

Sentence (56) occurred in the corpus without the article, and the native speakers gave the same judgment:

(56) Nafla li (*ha-) šen.
    Fell me (the-) tooth.
    ‘My tooth fell out’

Therefore, the definite article is obligatory in some constructions and optional in others.
6 The results and their theoretical implications

6.1 Statistics

The statistical method I used in order to check the correlation of the body part cardinality and visual saliency with the choice of the article was the linear regression analysis, performed with the SPSS package. The independent factors were the cardinality and the visual saliency of the body part. The article chosen in every particular sentence was the dependent variable. Table 9 below contains an example of the data in the SPSS representation.

<table>
<thead>
<tr>
<th>body part name</th>
<th>body part id</th>
<th>cardinality</th>
<th>visual saliency</th>
<th>article_id</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>hand</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>206</td>
</tr>
<tr>
<td>hand</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>arm</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>172</td>
</tr>
<tr>
<td>arm</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>leg</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>152</td>
</tr>
<tr>
<td>leg</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>foot</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>138</td>
</tr>
<tr>
<td>foot</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>finger</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>finger</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>tooth</td>
<td>6</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>tooth</td>
<td>6</td>
<td>32</td>
<td>0</td>
<td>1</td>
<td>47</td>
</tr>
</tbody>
</table>

Table 9. Representing the data in SPSS. article_id: 0 - definite, 1 - indefinite; n - the number of sentences found in the corpus.

In all the cases discussed in the previous sections, the cardinality of a body part was shown to be a statistically significant regressor for the indefinite article frequency. The visual saliency was statistically significant in most cases. In this section I would like to address the following question: to what extent is the difference between the frequencies explained by these factors?

I checked the contribution of the various factors in terms of variance. First, I calculated the variance in the data between the different body parts using the ANOVA procedure. Then, I checked how much of this variance is explained in terms of the difference in the regressors. The variance explained by the regressors is given by SPSS as part of the linear regression analysis. The results are in Table 10:

<table>
<thead>
<tr>
<th>Case</th>
<th>Variance explained by cardinality (percentage)</th>
<th>Residual variance explained by visual saliency (percentage out of total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish “duele”</td>
<td>79%</td>
<td>18%</td>
</tr>
<tr>
<td>Spanish “se rompió”</td>
<td>54%</td>
<td>32%</td>
</tr>
<tr>
<td>Portug. “dói”</td>
<td>85%</td>
<td>-</td>
</tr>
<tr>
<td>Portug. “quebrou”</td>
<td>73%</td>
<td>-</td>
</tr>
<tr>
<td>English “I broke”</td>
<td>99%</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 10. The importance of the regressors in terms of variance

| English “he broke” | 85% | 0.4% |
| English “she broke” | 77% | 4% |
| English “I raised” | 91% | 1.4% |
| English “she raised” | 94% | 2.1% |
| Italian “mi fa male” | 61% | 18% |

It can be seen from the table that the difference in cardinality accounts for most of the variance in all the cases. On the other hand, the factor of visual saliency is relatively important only for the Spanish and Italian examples. My conclusion is that the cardinality, but not the visual saliency, is crosslinguistically important.

6.2 Results - the factors that influence the choice of the article

The results obtained for the different languages were mostly similar. For unique body parts, only the definite determiner is allowed. For non-unique body parts, there is a variation between the definite and the indefinite determiner. The body part cardinality was shown to have a statistically significant correlation with the percentage of the indefinite article usage. The visual saliency was shown statistically significant in most cases, although its impact was comparable to that of cardinality only for the Spanish and Italian examples. Some of the data is shown in the combined graph in Figure 6.

The combination of various factors governing the choice of the article results in gradual changes between preferring the definite article and preferring the indefinite one. This supports the view that the conditions for the article use are not absolute. The choice of the article is affected by the distinguishability of the body part, affected by its cardinality and visual saliency (see Section 3.4). The higher the distinguishability, the less probable the choice of the indefinite article.

![Figure 6. Percentage of indefinite article usage in different languages.](image)
6.3 Asymmetry between the articles

Another interesting result is the asymmetry between the uniqueness effect of the definite and the indefinite article. Unique body parts can only be used with the definite article; the indefinite article is not allowed. Non-unique body parts can usually be used with either of the articles. A possible explanation is that the indefinite article has the implication of choice, of the necessity of choosing the intended referent from the set of candidates. It is therefore not accepted with the unique body part terms. The higher indefinite usage with the higher cardinality body parts can be explained by the more prominent necessity of the choice in these cases.

It seems that in the cases discussed here that the definite article is the unmarked choice, and the indefinite article is the marked one. In a similar case, Du Bois (1980, p. 232) claims that in (57) using the indefinite article would be “unnecessary precision”, that is, giving more information than necessary.

(57) The boy scribbled on the/a living-room wall.

Indeed, when used with a body part term, the indefinite article conveys more information about the body part cardinality, since the definite article can be used with all the body part terms, and the indefinite one only with the non-unique ones. In definite-only constructions the marked choice is not possible at all, and the additional information is not given in any case.

6.4 Relevance to definiteness theories

The results described above show that the theories requiring every definite noun phrase to be unique (Russell 1905, Kadmon 1990) cannot properly account for the data. This, however, was already shown by the counterexamples given in the previous literature. The main difference between the results in this paper and the other counterexamples is in showing that the uniqueness effect does exist, but the transition from the definite to indefinite usage is gradual.

The results also constitute evidence against the explanations of Ojeda (1993) and Epstein (1999). Those analyses claim that the body part noun phrase does not refer to a particular body part token, but rather to the uniquely identifiable body part group (Ojeda 1993), or role (Epstein 1999). Therefore, the use of the definite article is sanctioned in any reference to a body part. The difference between the usages with different body parts is not predicted by these theories.

The results can be better accounted for within the familiarity-based theories in which uniqueness is a by-product explained by Gricean principles (Heim 1982, Szabó 2000, Roberts 2003). The usage of the indefinite for non-unique reference is explained by non-arbitrariness, the need to avoid the mistake in the choice of the appropriate referent required when the definite description is used. If this is taken as a gradual measure, the indefinite article should be used more frequently in the cases in which the probability of the misidentification is higher. These are indeed the results I obtained: the indefinite article is more frequent with less distinguishable body parts, which would more likely lead to misidentification if used with the definite article³.

---

³ See (Chen 2004, fn 5) for another demonstration of gradual changes of distinguishability.
This is also consistent with Lyons' (1999) description of definiteness as grammaticalization of identifiability. Definiteness, basically a grammatical category showing identifiability, is also being used in cases, in which the identifiability is less than absolute. The further from complete identifiability, the less frequent is the definiteness. Such an explanation would require the identifiability to be a quantitative measure, instead of a supposedly binary category is Lyons' description. Further research is required to find and evaluate the different ways in which this can be accomplished. This is also true for Heim's non-arbitrariness condition.

The results also raise a methodological issue: theories about article use should be concerned with both the definite and the indefinite article, if both are present in the language. Papers about uniqueness (Kadmon 1990, Roberts 2003) typically discuss only the definite article. Some papers (Gundel et al. 1993) seem to assume the indefinite article can always be used, if syntactically possible. In our case, such an approach would miss the difference between the unique body part terms, for which the definite article is the only option, and the dual ones, allowing the indefinite one. In order to fully account for the facts, conditions governing the use of both articles should be examined.

6.5 Comparison to another non-unique usage of definite article

Birner and Ward (1994) present a number of examples in which definite NPs denote a non-unique object (see Section 2.6). It is important to note that their examples and the examples discussed here are different.

In their examples the definite usage is more likely if the objects are similar to each other in the relevant properties, i.e., not relevantly distinguishable. The body parts constructions discussed here show the opposite correlation: the less distinguishable body parts are less likely to appear with the definite article than the more distinguishable ones.

Also, while using a modifier in (28) makes the definite article infelicitous, the body parts construction shows the opposite effect. In Spanish, while *diente* ‘tooth’ is usually used with the indefinite article, “*diente* + adjective” is usually a definite noun phrase. This suggests that the phenomenon discussed in this paper is different from the one discussed in Birner and Ward (1994). The explanation proposed by Epstein (1999), that the noun definite phrase refers to a unique role, is more plausible in their case.

6.6 Syntactic variation and grammaticalization

With respect to the article usage with body part terms, I found two kinds of constructions in English and Hebrew. In some constructions, there is a variation between the definite and the indefinite article. In others, only the definite article is possible. One possible explanation would be to say that the second kind is a result of the grammaticalization of the definite article. The reason is that even in the constructions allowing variation, the vast majority of the cases have the definite article, since the body parts most frequently mentioned are the unique and the dual ones. The unique body part terms are only used with the definite article, and the dual ones are used with the definite article in most cases. The dominance of the definite article may cause the speaker to make a generalization and start using the definite article in all the remaining cases as well.

This can also explain the different results for similar constructions in different languages: the grammaticalization of the definite article in a particular construction in
one language can be in a more advanced stage than in the corresponding construction in the other language.

7 Semantic and pragmatic definiteness - evidence from German dialects

The data showing the usage of the definite article with non-unique body parts may lead to the following objection: maybe the definite article is used because the intended body part was somehow made salient, i.e., by being mentioned in the previous context. Evidence from German dialects brought in this section suggests this is not the case.

Some German dialects have two paradigms of the definite article. North Frisian (Ebert 1971), Bavarian (Scheutz 1988) and the dialects of Armen (Heinrichs 1954) and Mönchengladbach (Hartmann 1982) are among such dialects. One of the articles, called A-article by Ebert (1971), is used for referring to unique groups/objects and generic concepts. The other one, called D-article, is used for anaphoric and deictic use. According to Scheutz (1988), the two articles represent two kinds of definiteness, semantic, or $W$-definiteness, based on the knowledge of the world, and pragmatic, or $K$-definiteness$^4$, based on the textual or situational context. Examples:

(58) $a$ san (example from Ebert (1971), cited in Lyons (1999), p. 162)
the-A sun
'the sun'

(59) [jistərə hadət an də pəəə jəkləpt. ɛx lu:ə dəəə fənste ɛn zə: enə mən də ˈmən tən. ɛx ma:k di pəəə ɔp ɛn ˈmən əz: t fəəə məj... ] (= (16) in (Hartmann 1982))

'Yesterday someone knocked at the door. Looking out of the window I saw a man standing there. I opened the door and the man told me…'

Regarding the usage with body parts, Hartmann writes as follows (p. 196): "In relation to the speaker’s body as a possible context, parts like head, nose, etc. can be conceived as unica. Linguistic expressions containing the article $dər$ which refer to parts of a body like head, arms, eyes, legs, etc. usually refer to the speaker’s own head, etc. ". Examples:

(60) [də ˈkɔp mɛχ viːə.] (= (12a) in (Hartmann 1982))
The-A head me aches
'My head is aching'

(61) $Oke$ hat a fut breegen (= (37) in Ebert (1971))
Oke had the-A foot broken
'Oke broke his/her foot.'

As noticed by Hartmann, the dual body parts are used with the semantic A-article, and not with the pragmatic D-article. This would not have been possible if the reason for definiteness was being mentioned in previous context; this would create $K$-definiteness,

---

$^4$ The German terms $W(elt)$-Definitheit and $K$ontext)-Definitheit are from Krifka (1984).
expressed with the D-article. The usage of the A-article shows that there is W-definiteness, and the speaker perceives the intended body part to be “unique enough” for the definite article to be used.

8 A different case with the same dependency: cooperation words

The previous sections contained the discussion of the article usage with body part terms. In this section I examine another case showing a similar article distribution.

Words co-author, co-founder, partner, and others describe an individual as a part of a group devoted to a common activity. I wanted to check if the choice of the article used with such words depends on the actual size of the group. Unfortunately, for a given occurrence of the word co-founder the size of the group, that is, the total number of the co-founders is not easy to find. In addition, it is often unclear whether the person who wrote the sentence knew this number.

A systematic way for finding the cardinality of the group was found for the word co-author. Author descriptions on the online bookstore website Amazon.com contain many sentences like the following:

(62) He is a co-author of a college-level textbook in corporate finance.

(63) Cal Thomas is the co-author of Blinded by Might.

It is worth noting that if definite article only appeared when the referent is unique, it would never occur in such sentences.

I checked whether the usage of the article depends on the total number of co-authors. I sampled 20 sentences with the definite article and 20 sentences with the indefinite article. The sentences were taken from book descriptions on the Amazon.com website and were all of the form “author-NP is a/the co-author of book-NP”. The results are in Table 11 and Table 12.

<table>
<thead>
<tr>
<th>Co-authors</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

Table 11. Number of co-authors of books in “the” sentences.

<table>
<thead>
<tr>
<th>Co-authors</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

Arnold Zwicky and students in the Syntactic Variation seminar suggested that the choice of the article may depend on whether the author is the topic of the sentence. The common source of all the sentences, namely, author descriptions, eliminates the influence of this factor, since, naturally, the author is the topic in all these sentences.
Table 12. Number of co-authors of books in “a” sentences.

Books with 2 co-authors appeared in 15 out of 20 sentences (75%) with the indefinite article, but only in 2 out of 20 sentences (10%) with the definite article. For the purposes of a statistical analysis, I grouped together all the cases in which the number of co-authors was more than 2. The combined results are in Table 13.

<table>
<thead>
<tr>
<th>Number of co-authors \ Article used</th>
<th>definite</th>
<th>indefinite</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>more than 2</td>
<td>5</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 13. Article usage with the word co-author: combined results

ANOVA F-test shows that the two samples are different with significance level less than 0.001.

The results are similar to those obtained for the body part terms: the usage of the indefinite article gradually decreases for candidate sets of higher cardinality.

9 Further Research

The data of article usage with body part terms give rise to the following question: to what extent are the results due to the actual properties of the body part, and to what extent do they reflect the conventional image of it? The data of the article usage with the word co-author show that, at least in that case, the actual cardinality of the candidate set affects the article choice. This does not necessarily mean that the same holds for the body part terms. This could be checked by an experiment in which speakers are shown pictures with creatures having a non-conventional number of body parts, such as multiple heads and arms. The speakers then would have to write, or to complete, sentences involving the relevant body part. It would be interesting to see if the actual number of the body parts in the picture will affect the article usage frequencies.

Apart from cardinality, another factor that can potentially influence the identifiability is the similarity between the possible candidates. For example, if all the candidates look similar, the intended one is harder to identify than if it looks different. One way to check the influence of the similarity between candidates could be to make a production experiment in a number of settings that would differ in the level of saliency of the intended referent and to see if gradual changes in the frequency of article use will be obtained.

For Hebrew and Spanish I collected judgments of a small number of native speakers on some examples taken from the corpora. In order to check that the speakers indeed do not see any difference between the sentences with the definite and the indefinite article, a full completion study can be run, in which the subjects are shown sentences with the article taken out and then asked to fill in the appropriate article.

Similar phenomena can be examined in languages that do not necessarily mark definiteness. In Russian there is neither definite nor indefinite article; however, the definiteness is sometimes expressed by the word order. In Chinese there is no definite
article; however, the construction “yi ‘one’ + classifier” becomes grammaticalized as the indefinite article (Chen 2004).

As noticed above, in many cases the article usage is discussed in terms of where the definite article can or cannot be used. Some of the cases reported as counterexamples to uniqueness can actually be cases of variation.

Another possible direction for further research is to check the influence of the syntactic construction and the verb on the article usage frequency. For English and Hebrew I showed that in some constructions the definite article is obligatory. For Spanish and Portuguese there is significant variation between constructions in the article percentages. Examining a larger number of constructions can lead to understanding the syntactic factors influencing the article usage.

One possible direction for such an examination is suggested by Du Bois’ (1980) observations (see Section 6.3) that the identifiability is required to a reasonable precision. This suggests that the definite article would be less frequent with verbs that denote change of state that with those that do not. The definite article would be less frequent with main argument (direct object) and more frequent with indirect objects and adjuncts.

A diachronic research of the constructions in which the definite article is obligatory can show whether these cases are indeed a result of gradual grammaticalization, or they were used only with the definite article. The relative infrequency of the constructions can be a serious obstacle for such research.

Appendix A

Choosing the languages: Web as a corpus

A standard corpus usually allows the user to know the number of the words a corpus contains. This is not the case with using a Web search engine. Kilgarriff and Grefenstette (2003, section 3) describe a technique introduced and tested by Grefenstette and Nioche (2000) to find the total size (in words) of text in a particular language accessible via a search engine. A small corpus whose size is known is required. The technique is as follows:

Let \( C_1, C_2 \) be two corpora, and \( w_i \) a frequent word. Then, if the corpora are similar, we assume, that

\[
\frac{\text{tokens}(w_i, C_1)}{\text{size}(C_1)} = \frac{\text{tokens}(w_i, C_2)}{\text{size}(C_2)}
\]

\( \text{tokens}(w_i, C_j) \) is the number of occurrences of the word \( w_i \) in corpus \( C_j \).

If \( \text{size}(C_2) \) is unknown, it can be estimated as

\[
\text{size}(C_2) \approx \frac{\text{tokens}(w_i, C_2) \cdot \text{size}(C_1)}{\text{tokens}(w_i, C_1)}
\]

As noticed in (Kilgarriff and Grefenstette, 2003, fn 7), currently the search engines do not give the number of occurrences of a word, but rather the number of documents containing the word. Using this number, while using the number of occurrences for the calibrating corpus \( C_1 \), leads to an underestimation of the size of \( C_2 \).

Kilgarriff and Grefenstette’ estimation of the size of the web in March 2001 for the 10 languages with the highest estimated web size is given in Table 14. Kilgarriff and
Grefenstette checked a total of 30 Latin-script languages. Languages with a non-Latin script were not examined.

<table>
<thead>
<tr>
<th>Language</th>
<th>Web Size (millions of words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>76,600</td>
</tr>
<tr>
<td>German</td>
<td>7,040</td>
</tr>
<tr>
<td>French</td>
<td>3,840</td>
</tr>
<tr>
<td>Spanish</td>
<td>2,660</td>
</tr>
<tr>
<td>Italian</td>
<td>1,850</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1,330</td>
</tr>
<tr>
<td>Dutch</td>
<td>1,063</td>
</tr>
<tr>
<td>Swedish</td>
<td>1,000</td>
</tr>
<tr>
<td>Norwegian</td>
<td>610</td>
</tr>
<tr>
<td>Czech</td>
<td>520</td>
</tr>
</tbody>
</table>

Table 14. Kilgarriff and Grefenstette’s Web size estimated, March 2001

Using the method described above, I estimated the size of the Spanish web to be about 10 billion words, and the approximation for Italian is about the same. Assuming a similar web growth, the numbers in Table 14 should be multiplied by 4 to obtain a current picture. In 2001, when the size of Spanish and Italian corpus was approximately 2.5 billion words, results discussed in this paper could not have been obtained for these languages. The size of the web corpus limits the languages for which significant results can be obtained at a given time. As the web grows, the number of such languages will grow as well.

References


