5-2014

A Unified Annotation Scheme for the Semantic/Pragmatic Components of Definiteness

Archna Bhatia
Carnegie Mellon University

Mandy Simons
Carnegie Mellon University, simons@andrew.cmu.edu

Lori Levin
Carnegie Mellon University, levin@andrew.cmu.edu

Yulia Tsvetkov
Carnegie Mellon University

Chris Dyer
Carnegie Mellon University, cdyer@cs.cmu.edu

See next page for additional authors

Published In

Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 910-916.

A Unified Annotation Scheme for the Semantic/Pragmatic Components of Definiteness

Archna Bhatia* **Mandy Simons†** **Lori Levin*** **Yulia Tsvetkov*** **Chris Dyer*** **Jordan Bender‡**

*Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA
{archna, lsl, ytsvetko, cdyer}@cs.cmu.edu

†Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA
simons@andrew.cmu.edu

‡Department of Linguistics, University of Pittsburgh, Pittsburgh, PA
jab311@pitt.edu

Abstract

We present a definiteness annotation scheme that captures the semantic, pragmatic, and discourse information associated with noun phrases, which we call *communicative functions*. A survey of the linguistics literature suggests that definiteness does not express a single communicative function but is a *grammaticalization* of many such functions, for example, identifiability, familiarity, uniqueness, and specificity. Our annotation scheme unifies ideas from previous research on definiteness while attempting to remove redundancy. The scheme encodes the communicative functions of definiteness rather than the grammatical forms of definiteness. We assume that the communicative functions are largely maintained across languages while the grammaticalization of this information may vary. Corpora that are annotated using communicative functions can be used to train classifiers, offering data-driven insights into the grammaticalization of definiteness in different languages. We release our annotated corpora for English and Hindi as well as sample annotations for Hebrew and Russian, together with an annotation manual.

Keywords: definiteness, annotation scheme, communicative functions

1. Introduction

We follow Lyons (1999) in that definiteness is a morphosyntactic category which represents the grammaticalization of some category of meaning.[1] The difference in meaning in (1) and (2) illustrates the meaning associated with definiteness. While it appears that (1) with the indefinite article *a* would be uttered in a context where the addressee is not aware of the referent of the noun phrase (NP) *a car*, in fact he is being introduced to it here, (2) with the definite article *the* would be used in a context where the knowledge about the referent of the NP *the car* is shared by both discourse participants, the speaker as well as the addressee.

(1) I bought **a car** this morning.

(2) I bought **the car** this morning.

The use of articles in English NPs indicates to the addressees what the intended referent might be (a referent already known to the addressee or a new one being introduced).[2] These articles, which are considered to be expressing definiteness in English, thus denote semantic, pragmatic, and discourse information related to the NPs which helps the discourse participants in identifying appropriate referents for the NPs from all the possible referents available in the discourse context. We use the term **communicative functions of definiteness** to refer to such semantic, pragmatic, and discourse information.[3] It should be noted, however, that the articles are not the only grammatical constructions in English that express the communicative functions (or their combinations); there are other constructions as well, such as bare plurals (*dinosaurs*), possessives (*John's daughter*), and pronouns (*she*), all of which express various communicative functions associated with definiteness.

In addition to the variability within a language with respect to the grammatical form of various communicative functions, there can also be variability across languages to express the same communicative functions. For example, Czech, Hindi, Indonesian, Russian do not have articles; the same communicative functions that English articles express have to be conveyed using different grammatical strate-

[1] In §2, we explore what this category of meaning is.

[2] We use the term *article* to refer to non-deictic determiners like English *the* and *a*.

[3] These *communicative functions* include the functions associated with known discourse entities (definites) such as uniqueness, and familiarity, as well as functions associated with new entities being introduced (indefinites) such as nonanaphoricity. See §2 and §3 for more details.

gies (if the language expresses those communicative functions formally).[4] This creates a problem for the machine translation systems since the languages on the source side and the target side may be using different grammatical strategies. For example, Tsvetkov et al. (2013); Stymne (2009) have observed that translating from an article-language to an article-less language is problematic. Furthermore, the communicative functions or their combinations that are selected to be expressed formally may differ across languages. For example, Schwarz (2013) shows that Hausa expresses the distinction between an entity that has been previously mentioned in the discourse (BASIC_ANAPHORA in fig. 1) and an entity that is inferrable but is not mentioned previously (UNIQUE_PHYSICAL_COPRESENCE in fig. 1) as illustrated in (3) and (4) respectively. English, on the other hand, does not seem to make this distinction formally and uses the same form (*the*) to denote both situations.

(3) yáayáa garî**dî-n**.
town DEF
'how's **the** town [that you've visited]?'
(There is a prior discussion of the town in the discourse and the speaker asks the addressee this question.)

(4) yáayáa garî-**n** .
town-DEF
'how's **the** town?'
(The speaker arrives at the addressee's place, and first asks him this. Note here *the town* is the first mention but its referent is inferrable from common ground, they are talking about the town in which the place is located.)

To add yet another complication for the MT systems, the same grammatical form across languages may not express (exactly) the same communicative functions in these languages. Croft (1991) points out many differences in article usage between English and French depending on the meanings (communicative functions) that are being expressed.

However, for machine translation, a target language sentence ideally expresses the same meaning/ communicative functions as the source language sentence.[5] Hence, if we know the mapping between the communicative functions and the grammatical constructions in the source language as well as in the target language, this knowledge can be leveraged to help the MT system select the correct grammar in the target language (e.g., by assigning more weight to the target language constructions that express the relevant communicative function).

We therefore have two goals: to identify the communicative functions of definiteness that are preserved across translations, and to correlate those meanings with the grammatical constructions that express them in the source and target languages. This paper addresses the first goal via the creation of an annotation scheme to express the communicative functions of definiteness. The second goal will be the target of future work.

§2 provides theoretical background on functions of definiteness. We operationalize the semantic, pragmatic, and discourse components of definiteness with a novel taxonomy and an annotation scheme presented in §3. Our annotated corpus and details of the annotation process are described in §4. We conclude in §5 with suggestions for future research, and discuss how the annotated corpus that we release can be used in downstream applications.

All the annotated materials, along with the annotation manual, are available at `http://www.cs.cmu.edu/~ytsvetko/definiteness_corpus`.

2. The Notion of Definiteness

A survey of the literature on definiteness suggests notions such as uniqueness, familiarity, identifiability, anaphoricity, specificity, and referentiality as being related to definiteness (Birner and Ward, 1994; Condoravdi, 1992; Evans, 1977, 1980; Gundel et al., 1988, 1993; Heim, 1990; Kadmon, 1987, 1990; Lyons, 1999; Prince, 1992; Roberts, 2003; Russell, 1905, inter alia). Collectively, the literature on definiteness covers a wide range of uses of definite descriptions with the communicative functions mentioned above, such as uniqueness, and correlates them with many types of NPs and various constructions. However, different papers may employ different theoretical perspectives and cover slightly different ranges of data.

To formalize the communicative functions related to definiteness, many studies propose to reduce it to one or two overarching communicative functions. For example, Kadmon (1987) and Evans (1980) explain definiteness in terms of semantic *uniqueness*. In that account, definite noun phrases refer to entities that are unique for all practical purposes, such as *the sun* when

[4] Some other grammatical strategies languages use to express definiteness are: differential object marking (Tippets, 2011), existential constructions for indefinite subjects (Chen, 2004), alternative word orders, and special constructions such as the *ba* construction in Chinese (Chen, 2004).

[5] If, for example, an entity is identifiable in one language in a context, it is identifiable in the other language too even if English shows identifiability using the definite article, Hindi may denote it using an accusative postposition, or Chinese by putting the corresponding NP in the subject position of a transitive verb.

speaking non-astronomically in our solar system, or *the biggest mountain in West Virginia* (Roberts, 2003). Other entities become unique in context, by constraining their reference with modifiers and adding facts about them. Books are not unique, but *the blue book on the table* can be unique in a given discourse context, and *the book* can become unique in a discourse, referring to the book that we are talking about. However, definite discourse referents are not always unique. In the famous donkey sentence—*Every man who owns a donkey beats it*—the pronominal *it* does not have a unique referent.[6] It refers to all of the donkeys that are owned.

Roberts (2003) accounts for definiteness in terms of a combination of uniqueness and a presupposition of familiarity. However, it is difficult to apply this approach to possessive definite descriptions (*John's daughter*) and the weak definites (Poesio, 1994) (*My aunt got attacked by the parent of a student whom she had failed*), which are neither unique nor necessarily familiar to the hearer before they are uttered. *John's daughter* can be felicitously uttered when John is known to the addressee. In many discourse contexts his daughter does not need to be previously mentioned and, in fact, he may have more than one daughter. We will return to this type of reference later when we discuss *bridging*, i.e., using a known discourse referent to anchor the reference of another.

Poesio and Vieira (1998) found in their experiments that the two predominant communicative functions, uniqueness and familiarity, could account for only about half of their data. We take such linguistic observations to suggest that definiteness is not as homogeneous category as many previous studies have assumed, and hence it cannot be reduced to just one or two of the aforementioned communicative functions. Instead, it should be seen as a grammaticalization (Hopper and Traugott, 2003; Chen, 2004) of many such communicative functions. In our investigation, we have devised a composite annotation scheme, combining ideas from previous taxonomies, such as Hawkins (1978); Prince (1981, 1992), attempting to capture most of the notions covered in the theoretical literature and eliminating the redundancies. This scheme was revised through a number of iterations by annotating texts from various genres in two languages (English and Hindi). Henceforth, we will refer to the scheme as **an annotation scheme for the communicative functions of definiteness (CFD)**. Below, we present the CFD annotation scheme and describe its operationalization.

[6]Note that personal pronouns such as *I*, *you*, *she*, and *it*; proper names; and NPs with *the*, demonstratives or possessives are treated as definites generally (Prince, 1992).

3. The Annotation Scheme for the Communicative Functions of Definiteness

The CFD annotation scheme, organized around the taxonomy in fig. 1, is a unified compilation of communicative functions related to definiteness studied previously in the literature, as well as some new functions.[7] The CFD annotation scheme annotates for the communicative functions associated with Anaphoricity, (Basic and Extended/Bridging), Genericity, Familiarity (i.e. being Hearer-Old), Predicativity, Referentiality, Specificity, and Uniqueness. Below we traverse through the annotation scheme and describe the distinctions that it makes.[8]

The first main distinction the CFD annotation scheme makes is with respect to **Anaphoricity**: whether an NP is Anaphoric or Nonanaphoric—i.e., whether the entity is old in the discourse or not.

The anaphoric NPs include pronominal and nominal descriptions that have been mentioned previously. These previously-mentioned NPs do not need to be identical in form to their antecedents, e.g., *the child* can be an anaphoric reference to *a girl*. NPs whose existence is evoked by previous NPs or events are also treated as anaphoric with the subheading of EXTENDED_ANAPHORA, in analogy with and extending the notion *bridging* introduced by (Clark, 1977). These include mentioning *the kitchen* after talking about a house (BRIDGING_NOMINAL) or mentioning *the victims* after using the verb *attack* (BRIDGING_EVENT). A special case of bridging (BRIDGING_RESTRICTIVE_MODIFIER) applies to NPs that contain a modifier that evokes them as in *the woman who lives next door*, which can be used in a conversation where the woman has not been previously mentioned. *Next door* is used deictically relative to the speaker, making the referent of the whole NP identifiable. The label BRIDGING_SUBTYPE_INSTANCE is used for NPs in a type-subtype or type-instance relationship with an already mentioned entity, e.g., a mention of *steel pennies* after mentioning *coins* earlier in the discourse. Finally, the label BRIDGING_OTHER_CONTEXT is used to mark the cases where the NP refers not to an

[7]The current version of the scheme is focusing mainly on the communicative functions associated with referential NPs; however, it identifies some nonreferential NPs as well, most of which appear under the categories of MISCELLANEOUS.

[8]For further details, refer to the CFD annotation manual at `http://www.cs.cmu.edu/~ytsvetko/definiteness_corpus`.

- **ANAPHORA** [+A]
  - **BASIC_ANAPHORA** [−B,+F]
    - SAME_HEAD
    - DIFFERENT_HEAD
  - **EXTENDED_ANAPHORA** [+B]
    - BRIDGING_NOMINAL [−G,+R,+S]
    - BRIDGING_EVENT [+R,+S]
    - BRIDGING_RESTRICTIVE_MODIFIER [−G,+S]
    - BRIDGING_SUBTYPE_INSTANCE [−G]
    - BRIDGING_OTHER_CONTEXT [+F]
- **NONANAPHORA** [−A,−B]
  - **UNIQUE** [+U]
    - **UNIQUE_HEARER_OLD** [+F,−G,+S]
      - UNIQUE_PHYSICAL_COPRESENCE [+R]
      - UNIQUE_LARGER_SITUATION [+R]
      - UNIQUE_PREDICATIVE_IDENTITY [+P]
    - UNIQUE_HEARER_NEW [−F]
  - **NONUNIQUE** [−U]
    - **NONUNIQUE_HEARER_OLD** [+F]
      - NONUNIQUE_PHYSICAL_COPRESENCE [−G,+R,+S]
      - NONUNIQUE_LARGER_SITUATION [−G,+R,+S]
      - NONUNIQUE_PREDICATIVE_IDENTITY [+P]
    - NONUNIQUE_HEARER_NEW_SPEC [⌝F,−G,+R,+S]
    - NONUNIQUE_NONSPEC [−G,−S]
  - **GENERIC** [+G,−R]
    - GENERIC_KIND_LEVEL
    - GENERIC_INDIVIDUAL_LEVEL
- **MISCELLANEOUS** [−R]
  - PLEONASTIC [−B,−P]
  - QUANTIFIED
  - PREDICATIVE_EQUATIVE_ROLE [−B,+P]
  - PART_OF_NONCOMPOSITIONAL_MWE
  - MEASURE_NONREFERENTIAL
  - OTHER_NONREFERENTIAL

Figure 1: CFD (Communicative Functions of Definiteness) annotation scheme. The non-leaf labels are in bold. +/− values are shown for ternary attributes Anaphoric, Bridging, Familiar, Generic, Predicative, Referential, Specific, and Unique; these are inherited from supercategories, but otherwise default to 0. Thus, for example, the full attribute specification for UNIQUE_PHYSICAL_COPRESENCE is $[-A,-B,+F,-G,0P,+R,+S,+U]$.

entity in the real world or in the discourse representation structure (as defined in Discourse Representation Theory), but to a part of the prior discourse. For example, *This* in a sentence *This is entirely true.*, referring to the story a speaker has just narrated to his addressee, is annotated with this label.

Nonanaphoric NPs are entities that have not been mentioned or are not evoked by something that was mentioned. The next main distinction CFD makes is within this class of Nonanaphoric NPs, namely **Uniqueness**. Some Nonanaphoric nominals get a unique interpretation semantically as there is only one referent corresponding to that description, e.g., *the sun*; others receive a unique interpretation in a context where the corresponding referent is the most salient entity associated with that description, e.g., *the table* when uttered in a physical context where there is just one table present, even though there may be millions of tables present in the world. Proper names are treated as Unique entities as well. Nonunique NPs are the descriptions which do not get a unique referent semantically or situationally.

For the Unique and Nonunique NPs, another distinction is made with respect to **Familiarity**: whether the NP is hearer-old (familiar) or hearer-new (unfamiliar). An NP may be considered hearer-old due to (a) its referent being perceptually present (UNIQUE_PHYSICAL_COPRESENCE, or NONUNIQUE_PHYSICAL_COPRESENCE), or (b) the discourse participants having common knowledge about the larger situation as a result of being part of a community (UNIQUE_LARGER_SITUATION,

or NONUNIQUE_LARGER_SITUATION). For UNIQUE_LARGER_SITUATION, let's consider the following examples. When the NP *the President* is uttered in USA contemporarily, the discourse participants associate it with Barack Obama as the referent, but when it is uttered in India, the discourse participants take Pranab Mukherjee as the referent. Other examples for NPs with common-knowledge referents are *the Empire State Building*, *Taj Mahal*, *Barack Obama*, and *Gandhi*. For NONUNIQUE_LARGER_SITUATION, consider the following situation. One can talk about *the hotel* or *the program chair* at a conference even when those things have not been previously mentioned. This is possible due to the conference participants having common knowledge about conferences. UNIQUE_PREDICATIVE_IDENTITY and NONUNIQUE_PREDICATIVE_IDENTITY are also considered hearer-old. The hearer-new NPs are further distinguished for **specificity**. Thus, we may get a specific interpretation (*She wants to marry an Irishman. His name is Paul.*) or a non-specific interpretation (*She wants to marry an Irishman. She should go and find one.*).

Another distinction is made for the Nonanaphoric NPs with respect to **Genericity**: whether an NP is generic or not. The generic NPs are the NPs that appear with a predicate or in a context where they refer to the kind or the whole class, rather than a selected set of members from that class. For example, *Dinosaurs* in *Dinosaurs are extinct.* (GENERIC_KIND_LEVEL) and *Cats* in *Cats have fur.* (GENERIC_INDIVIDUAL_LEVEL).

Additionally, a number of miscellaneous cases are also identified, which mostly represent various types of Nonreferential NPs (**Referentiality**). For some of these cases, if it is their first mention, they are labelled as the Miscellaneous category, e.g., QUANTIFIED, and MEASURE_NONREFERENTIAL. However, if they have appeared earlier in the discourse, they are marked as anaphoric as the subsequent mentions can be taken to refer to the first mention itself. Others, however, are marked as Miscellaneous category only, such as PLEONASTIC.

An important property of the CFD scheme is that it is hierarchical in nature. This hierarchical organization does not necessarily represent how these classes are grouped in natural languages.[9] Instead, it serves to reduce the complexity of decisions (selecting out of 24 classes vs. out of only a few classes) that an annotator needs to make for better speed and consistency. We, however, map our categories to eight attributes, such as *Anaphoric*, *Familiar*, and *Unique*, which represent formal groupings based on relevant semantic/ pragmatic notions. As shown in fig. 1, we have identified values for each of our classes corresponding to these attributes. The + value suggests the presence of an attribute (e.g. NONUNIQUE_HEARER_NEW_SPECIFIC is + *Referential*), the - value suggests the absence (e.g. NONUNIQUE_HEARER_NEW_SPECIFIC is - *Familiar*, and - *Unique*), and 0 value is used as a default when the presence or absence of an attribute is not associated with/ specifiable for a class. This information enables us to manipulate the annotated data in formal groupings which are not expressed in the hierarchy (CFD scheme) necessarily but may turn out to be useful for an understanding of definiteness.

4. Definiteness Corpus

In this section, we briefly describe our data, and the annotation procedure. We also provide an evaluation of the annotations in terms of inter-annotator agreement.

4.1. Data

We have primarily annotated data from two languages: English and Hindi. However, a few sample annotations using the CFD annotation scheme are also provided for Hebrew and Russian. It should be noted that all of these four languages have different grammatical expressions of definiteness. English has both definite and indefinite articles. Hebrew has an explicit marker for definiteness but not for indefiniteness. Hindi, on the other hand, uses *ek* (one) which sometimes can act as an indefinite article. It does not have any unmarked definite article. Russian does not have definite or indefinite articles.

We have selected four genres: TED talks, presidential addresses, published news articles, and fictional narratives. The parallel data from the TED talks corpus is in all four languages.[10] There are sixteen TED talks annotated in both English and Hindi, two of which also have been annotated for Hebrew and Russian. In addition, for English, we have one presidential address, two newspaper articles, and two fictional narratives.

The size of the English corpus is 13,860 words, containing 868 sentences, which contain 3422 noun phrases. The breakdown of the corpus is as follows: the TED talk genre represents about 75% of the corpus; the presidential address represents about 16%; fictional narratives about 5%; and news articles 4%.

[9] Komen (2013) has also proposed a hierarchy with similar leaf nodes, but different internal organization.

[10] These TED talks were obtained from a large parallel corpus, `http://www.ted.com/talks/`.

The Hindi corpus contains the same 16 TED talks that were annotated for English.

4.2. Annotation Procedure

The annotatable units for CFD are noun phrases (NPs).[11] NPs containing embedded NPs are annotated from the inside out, borrowing insights from Discourse Representation Theory on how the discourse representation structure is incrementally updated as new information is added to the discourse (van Eijck and Kamp, 1997).

Annotators use the browser-based brat annotation software (Stenetorp et al., 2012)[12] as follows: The annotator selects the span (an NP) that is to be annotated. This opens a dialog box with the hierarchical label inventory of fig. 1, from which the annotator selects the appropriate label. The annotations are stored as a text file. Each line in the annotation file contains information about one NP annotation: a unique ID, the annotation label, the character offsets of the annotated unit in the document, and the annotated NP itself.

Figure 2 is an excerpt from a news article "Cops: Burglar leaves his car, iPad at scene of the crime" annotated with the CFD scheme.

A bungling burglar (NONUNIQUE_HEARER_NEW_SPECIFIC) left footprints (BRIDGING_NOMINAL), fingerprints (BRIDGING_NOMINAL), and even photos of himself (DIFFERENT_HEAD) (BRIDGING_RESTRICTIVE_MODIFIER) at a Los Angeles home (NONUNIQUE_HEARER_NEW_SPECIFIC), police (NONUNIQUE_LARGER_SITUATION) said. Lt. Paul Vernon (UNIQUE_HEARER_HEARER_NEW) said the man (DIFFERENT_HEAD) kicked in the front door of a house in Arleta (UNIQUE_HEARER_NEW) (NONUNIQUE_HEARER_NEW_SPECIFIC) (BRIDGING_RESTRICTIVE_MODIFIER) on Nov. 13 (NONUNIQUE_LARGER_SITUATION) and started to ransack a bedroom (BRIDGING_NOMINAL).

Figure 2: An excerpt from a news article ''Cops: Burglar leaves his car, iPad at scene of the crime" annotated using the CFD annotation scheme.

[11] CFD assigns a communicative function label to every NP except for first person pronouns, second person pronouns, and relative pronouns. These were excluded to reduce the manual annotation effort. The first and second person pronouns do not need to be manually annotated because they represent entities (the speaker and hearer) that are deictic in the discourse context. Because they are limited in vocabulary, they can automatically be labelled in a rule-based way if necessary. The identification of the communicative functions associated with relative pronouns is left for future work.

[12] The tool is available at `http://brat.nlplab.org`.

4.3. Inter-annotator Agreement

To measure inter-annotator agreement, we looked at the agreement in annotation labels assigned *given an NP as the annotatable unit*.[13] Two English annotators independently annotated the same two texts from different genres (one text from the TED talks genre and one text from the fictional narratives) as part of their training on the annotation scheme. These annotators then began rounds of reconciling their annotations in order to reach consensus, updating the annotation scheme as necessary. Once they reached consensus, they annotated four new texts without discussion and measured inter-annotator agreement. We found strong agreement with Cohen's $\kappa = 0.88$ over a total of 1202 annotated units from the four texts. The confusion matrix is presented in table 1. On individual genres, the inter-annotator agreement scores are: Cohen's $\kappa = 0.89$ for TED (502 annotated NPs), 0.85 for the presidential speech (492 annotated NPs), 0.95 for narratives (129 annotated NPs), 0.83 for news articles (79 annotated NPs).

5. Conclusions and Future Work

We have developed an annotation scheme for the communicative functions of definiteness, and have used this scheme to create a definiteness corpus which encodes semantic, pragmatic, and discourse functions of definiteness. We are releasing this corpus for English and Hindi, along with sample annotations using the same scheme for two other languages, Hebrew and Russian. We also provide an annotation manual to help other annotators produce similar corpora for definiteness.

The annotation scheme is somewhat stable, however, there are issues about which we are still deliberating. For example, our approach to coordinate structures is not finalized. Currently we do not analyze coordinated phrases internally. *Shoes or boots* is taken as one annotatable unit instead of analyzing it as two separate annotatable units *shoes* and *boots*. However, sometimes internal analysis of conjuncts is needed, for example if both conjuncts repeat some modifier (*end-use electricity and end-use of all energy*).

Looking at the table 1, the UNIQUE_LARGER_SITUATION seems to be a class that is slightly confusing between the annotators. For example, both annotators confused this class with NONUNIQUE_LARGER_SITUATION

[13] We had also looked at the boundary disagreements between the annotators, we found there were very few cases, less than 1%, where the two annotators did not agree on what constituted an NP.

Semantic Function Labels	Bridging_Other_Context	Bridging_Event	Bridging_Nominal	Bridging_Restrictive_Modifier	Different_Head	Generic_Individual_Level	Measure_Nonreferential	Other_Nonreferential	Part_of_Noncompositional_MWE	Pleonastic	Predicative_Equative_Role	Quantified	Same_Head	Nonunique_Hearer_New_Specific	Nonunique_Larger_Situation	Nonunique_Physical_Copresence	Nonunique_Nonspecific	Nonunique_Predicative_Identity	Unique_Hearer_New	Unique_Larger_Situation	Nonunique_Physical_Copresence	Unique_Predicative_Identity
Bridging_Other_Context	25	0	0	1	2	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
Bridging_Event	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bridging_Nominal	0	0	21	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
Bridging_Restrictive_Modifier	0	0	0	265	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0
Different_Head	1	0	0	0	104	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
Generic_Individual_Level	0	0	0	4	0	29	1	1	0	0	0	0	1	1	0	0	7	0	0	1	0	0
Measure_Nonreferential	0	0	0	0	0	0	18	0	0	0	0	5	0	2	0	0	0	0	0	0	0	0
Other_Nonreferential	0	0	0	0	0	0	0	58	2	0	0	0	0	0	0	0	1	0	0	1	0	0
Part_of_Noncompositional_MWE	0	0	0	0	0	1	0	0	20	1	0	0	0	0	0	0	0	0	0	0	0	0
Pleonastic	0	0	0	1	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0
Predicative_Equative_Role	0	0	0	0	0	0	0	0	0	0	22	0	1	0	0	0	0	0	0	0	0	0
Quantified	0	0	0	0	0	0	0	0	0	0	0	58	0	0	0	0	0	0	0	0	0	0
Same_Head	2	0	1	7	4	0	0	1	0	0	0	0	186	0	1	0	0	0	0	1	0	0
Nonunique_Hearer_New_Specific	0	0	0	4	1	0	0	1	1	0	1	1	0	53	0	1	0	0	1	1	0	0
Nonunique_Larger_Situation	0	0	1	2	0	0	0	1	0	0	0	0	2	2	38	0	1	0	0	5	0	0
Nonunique_Physical_Copresence	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	1	0
Nonunique_Nonspecific	0	0	0	8	0	0	0	0	0	0	0	1	1	0	1	1	60	0	0	1	0	0
Nonunique_Predicative_Identity	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
Unique_Hearer_New	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	10	3	0	0
Unique_Larger_Situation	0	0	0	4	1	0	0	0	0	0	0	0	2	0	4	0	1	0	0	55	2	0
Nonunique_Physical_Copresence	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	2	9	0
Unique_Predicative_Identity	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 1: Confusion matrix for inter-annotator agreement on annotations of 4 texts from different genres combined together. (Number of annotations: 1202, Cohen's $\kappa = 0.88$)

for about 6-7% of the cases. This suggests that we need to have better criteria to determine whether an entity is unique or nonunique. Similarly, for GENERIC_INDIVIDUAL_LEVEL and NONUNIQUE_LARGER_SITUATION, it would be helpful to have clearer tests to identify NPs with these communicative function labels.

This corpus could be used for building classifiers to learn grammaticalization strategies based on the CFD annotations. In principle, knowledge about the shared semantic and pragmatic information across languages and the knowledge of how that information is expressed in each of the languages (source and target) should help improve MT across languages. For example, the source language could be transformed to mirror the target language in terms of grammatical expression of the annotated semantic features. The target-like source language can then be used to train a machine translation system. Another option is to provide CFD labels as auxiliary information to the MT system.

This paper is an attempt at the first step in the process—building a corpus that lets us determine the relevance of the communicative functions of definiteness and the corresponding grammaticalization strategies. In our future work, we plan to use the CFD annotation scheme and the definiteness corpus described in §4.1 to build classifiers for predicting the communicative functions and for learning grammaticalization corresponding to the communicative functions. We plan on expanding the corpus further and in all the four languages, and revising the annotation scheme further to cover more communicative functions as well as additional Nonreferential categories. Finally, we plan on using the classifiers we build to aid machine translation systems in ways described above. The classifier output can also be used for other downstream applications, for example to aid the coreference resolution systems, and the information retrieval systems.

6. Acknowledgements

This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533. We would also like to thank Shay Cohen for providing us with the Hebrew annotations, and Nathan Schneider for his continuous help and support.

References

Birner, Betty and Ward, Gregory (1994). Uniqueness, familiarity and the definite article in English. In *Proc. of the Twentieth Annual Meeting of the Berkeley Linguistics Society*, pages 93–102.

Chen, Ping (2004). Identifiability and definiteness in Chinese. *Linguistics*, 42:1129–1184.

Clark, Herbert H. (1977). Bridging. In Johnson-Laird, P.N. and Wason, P.C., editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press.

Condoravdi, Cleo (1992). Strong and weak novelty and familiarity. In *Proc. of SALT II*, pages 17–37.

Croft, William (1991). *Syntactic Categories and Grammatical Relations: The Cognitive Organization and Information*. The University of Chicago Press.

Evans, Gareth (1977). Pronouns, quantifiers and relative clauses. *Canadian Journal of Philosophy*, 7(3):46.

Evans, Gareth (1980). Pronouns. *Linguistic Inquiry*, 11.

Gundel, Jeanette K., Hedberg, Nancy, and Zacharski, Ron (1988). The generation and interpretation of demonstrative expressions. In *Proc. of XIIth International Conference on Computational Linguistics*, pages 216–221.

Gundel, Jeanette K., Hedberg, Nancy, and Zacharski, Ron (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.

Hawkins, John A. (1978). *Definiteness and Indefiniteness*. Croom Helm.

Heim, Irene (1990). E-type pronouns and donkey anaphora. *Linguistics and Philosophy*, 13:137–177.

Hopper, Paul J. and Traugott, Elizabeth Closs (2003). *Grammaticalization*. Cambridge University Press.

Kadmon, Nirit (1987). *On unique and non-unique reference and asymmetric quantification*. Ph.D. thesis, University of Massachusetts.

Kadmon, Nirit (1990). Uniqueness. *Linguistics and Philosophy*, 13:273–324.

Komen, Erwin Ronald (2013). *Finding focus: a study of the historical development of focus in English*. LOT, Utrecht.

Lyons, Christopher (1999). *Definiteness*. Cambridge University Press.

Poesio, Massimo (1994). Definite descriptions, focus shift, and a theory of discourse interpretation. In *in Proceedings of the Conference in Focus on Natural Language*.

Poesio, Massimo and Vieira, Renata (1998). A corpus based investigation of definite description use. *Computational Linguistics*, 24:183–216.

Prince, Ellen F. (1981). Toward a taxonomy of given-new information. In Cole, P., editor, *Radical Pragmatics*, pages 223–256. Academic Press.

Prince, Ellen F. (1992). The ZPG letter: Subjects, definiteness and information status. In Thompson, S. and Mann, W., editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins.

Roberts, Craig (2003). Uniqueness in definite noun phrases. *Linguistics and Philosophy*, 26:287–350.

Russell, Bertrand (1905). On denoting. *Mind, New Series*, 14:479–493.

Schwarz, Florian (2013). Different types of definites crosslinguistically. *Language and Linguistics Compass*, 7:534–559.

Stenetorp, Pontus, Pyysalo, Sampo, Topić, Goran, Ohta, Tomoko, Ananiadou, Sophia, and Tsujii, Jun'ichi (2012). brat: a web-based tool for nlp-assisted text annotation. In *Proc. of the Demonstrations Session at EACL*.

Stymne, Sara (2009). Definite noun phrases in statistical machine translation into Danish. In *Proc. of Workshop on Extracting and Using Constructions in NLP*, pages 4–9.

Tippets, Ian (2011). Differential object marking: Quantitative evidence for underlying hierarchical constraints across spanish dialects. In *Selected Proc. of the 13th Hispanic Linguistics Symposium*, pages 107–117.

Tsvetkov, Yulia, Dyer, Chris, Levin, Lori, and Bhatia, Archna (2013). Generating English determiners in phrase-based translation with synthetic translation options. In *Proc. of WMT*. Association for Computational Linguistics.

van Eijck, Jan and Kamp, Hans (1997). Representing discourse in context. In *Handbook of Logic and Language*, pages 179–237. Elsevier.