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, which we call communicative functions, associated with linguistic descriptions such as *a story about my speech*, *the story*, *every time I give it*, *this slideshow*. A survey of the literature suggests that definiteness does not express a single communicative function but is a grammaticalization of many such functions, for example, identifiability, familiarity, uniqueness, specificity. Our annotation scheme unifies ideas from previous research on definiteness while attempting to remove redundancy and is easy to annotate. This annotation 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. One of the final goals is to use our semantically annotated corpora to discover how definiteness is grammaticalized in different languages. We release our annotated corpora for English and Hindi, and sample annotations for Hebrew and Russian, together with an annotation manual.

Keywords: definiteness, annotation scheme, communicative functions, anaphoricity, uniqueness, specificity, identifiability

1. Introduction

In this paper we use the word *definiteness* broadly to refer to a range of semantic, pragmatic, and discourse functions relating to the reference of noun phrases. These *communicative functions* include referring to known discourse entities (definites) and introducing new ones (indefinites). We will use the word *article* to refer to non-deictic determiners like English *the* and *a*.

The use of articles in English helps addressees pick appropriate referents from all the possible referents available in the discourse context and helps to maintain common ground between the speaker and hearer about which discourse entities are under discussion. However, not all languages make use of articles to perform these functions. For example, Czech, Hindi, Indonesian, Russsian do not have articles. In article-less languages, the same information can be conveyed using different grammatical strategies. Also within the same language, there may be more than one grammatical strategy to express the range of communicative functions associated with definiteness.

In addition to the use of articles, some other strategies for expressing definiteness are: the use of affixes as markers of definiteness, the use of alternative word orders, the use of special constructions such as differential object marking (Tippets, 2011) or existential constructions (for indefinite subjects), and the use of pronouns (*she*) or possessive descriptions (*John's daughter*) as definite descriptions. Chen (2004) shows that in Chinese, a language without articles, the subject position for transitive verbs is almost exclusively used for definite nouns. Indefinite subjects in Chinese are expressed by the existential construction. Direct objects after the verb may be definite or indefinite, but when expressed with the *ba* construction, they are almost exclusively definite.

Our work is couched in the context of machine translation. Machine translation systems face problems in translating from a language that uses a strategy that is different from the target language. For example, translating from an article-language to an article-less language is problematic (Tsvetkov et al., 2013; Stymne, 2009). Even languages that use the same strategies do not always use them in the same situations. Croft (1991) points out many differences in article usage between English and French depending on the meaning that is being expressed. For machine translation, a target language sentence ideally expresses the same meaning as the source language sentence. 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. The second goal will be the target of future work.

In §2, we discuss the semantic components of definiteness. In §3, we briefly discuss our annotation scheme and in §4, we describe the data and the annotation tool and we report on the inter-annotator agreement for the texts in the English corpus. §5 mentions our ideas about how this annotated corpus can be used to aid machine translation.

Our contributions in this paper, listed below, are publicly released.[1]

1. A unified, comprehensive, and practical annotation scheme
2. A small annotated English corpus (13860 words), a corresponding Hindi corpus for one of the genres annotated for English, and samples of Hebrew and Russian annotations
3. An annotation manual

2. The Notion of Definiteness

A survey of the literature on definiteness suggests notions such as uniqueness, familiarity, identifiability, anaphoricity, specificity, 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 huge range of communicative functions and correlates them with many types of noun phrases and sentential constructions. However, it is baffling to try to combine the collective wisdom because each paper has a different theoretical perspective and covers a different range of data.

In order to make sense out of the many communicative functions related to definiteness, some studies attempt to reduce it to one or two over-arching communicative functions. For example, Kadmon (1987); Evans (1980) explain definiteness in terms of semantic uniqueness. Some noun phrases refer to entities that are unique for all practical purposes such as *the sun* when speaking non-astronomically in our solar system or *the biggest mountain in West Virginia* (Roberts, 2003). It can also be argued that in context other entities become unique 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 situation, and in fact *the book* can become unique in a discourse, referring to the book that we are talking about. However, it can be argued that discourse referents are not always unique. In the famous *donkey sentence* (Every man who owns a donkey beats it), *it* does not have a unique referent. 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*, 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 a category as these accounts have assumed and hence it cannot be reduced to just one or two of the above- mentioned communicative functions. Instead it should be seen as a grammaticalization (Hopper and Traugott, 2003; Chen, 2004) of many such communicative functions. Note even though each of these theories has some advantage over the other in terms of the data they account for, there are overlaps as well. Although we do not claim to have an exhaustive list of communicative functions of definiteness that grammaticalize in each language, in our investigation, we have attempted to make a composite annotation scheme that captures most of the notions covered in the literature and eliminates the redundancies. Henceforth, we will refer to the scheme as *an annotation scheme for the communicative functions of definiteness* (or *the CFD annotations scheme*, or *CFD*). Below we present the CFD annotation scheme and describe how it was operationalized.

[1] http://www.cs.cmu.edu/~ytsvetko/definiteness_corpus.gz

3. The Annotation Scheme for the Communicative Functions of Definiteness

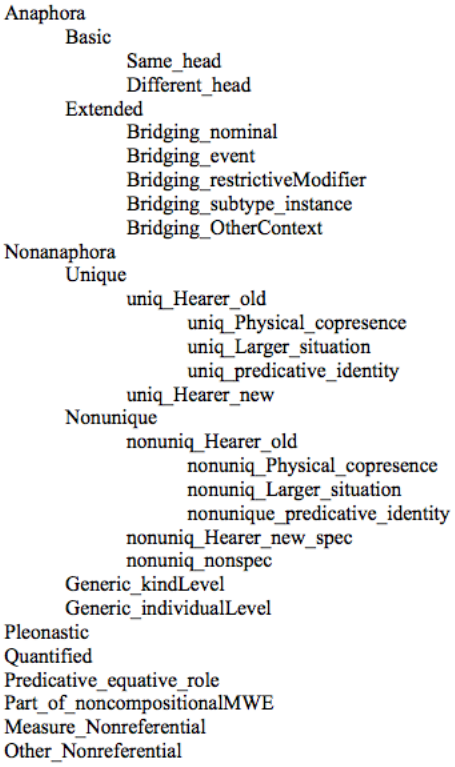
The CFD annotation scheme is a unified compilation of communicative functions related to definiteness studied previously in the literature as well as new functions as were needed while accounting for various types of NPs we came across in our corpora. It was developed by annotating texts from various genres mainly in two languages (English and Hindi), see §4.1. For CFD, the annotatable units are noun phrases (NPs). NPs containing embedded NPs are annotated from the inside out, taking some insights from Discourse Representation Theory in how the discourse representation structure (DRS) is incrementally updated as new information is added to the discourse (van Eijck and Kamp, 1997).

CFD assigns a communicative function label to every NP except for first- person pronouns, second- person pronouns and relative pronouns. The first and second person pronouns are not annotated because they represent entities (the speaker and hearer) that are deictic in the discourse context. Also since they are limited in vocabulary, they can automatically be labelled in a rule-based way if necessary. The relative pronouns, on the other hand, are not annotated because we are not certain at the moment how they fit into our annotation scheme. We will eventually incorporate such cases into the future versions of our scheme. These decisions were taken to reduce the manual annotation effort.

The current version of the CFD annotation scheme (Version 2) is provided in fig. 1. We would like to point out that we are enriching the scheme one part at a time, and we plan to keep revising it until it covers most of the functions definiteness encodes. Currently we are focussing on referential NPs.

The three main communicative functions in CFD are Anaphora vs. Nonanaphora (whether the entity is old in the discourse or not), Hearer-old vs. Hearer-new, and Unique vs. Nonunique (annotated for Nonanaphoric only in the current scheme).

Anaphoric NPs include pronouns and nouns that have been mentioned previously. Previously-mentioned nouns 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 bridging anaphora (in analogy with and extending the notion bridging introduced by (Clark, 1977). These include mentioning *the kitchen* after talking about a house or mentioning *the victims* after using the verb *attack*. A special case of bridging is 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 noun phrase identifiable.



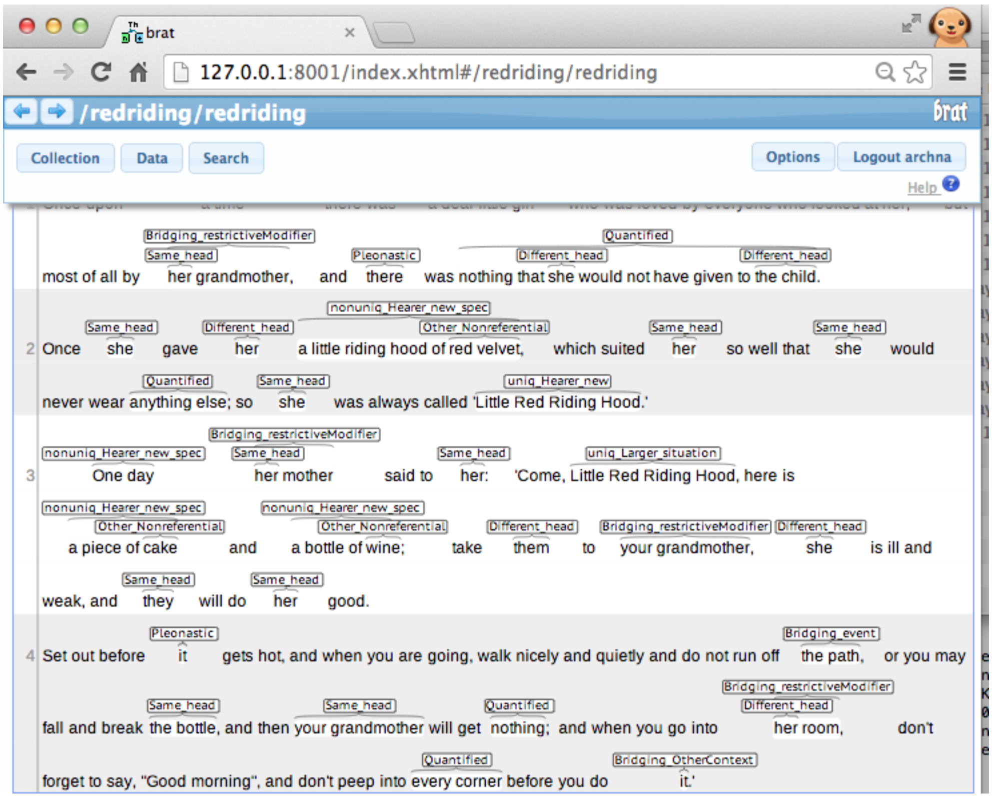
- Anaphora
  - Basic
    - Same_head
    - Different_head
  - Extended
    - Bridging_nominal
    - Bridging_event
    - Bridging_restrictiveModifier
    - Bridging_subtype_instance
    - Bridging_OtherContext
- Nonanaphora
  - Unique
    - uniq_Hearer_old
      - uniq_Physical_copresence
      - uniq_Larger_situation
      - uniq_predicative_identity
    - uniq_Hearer_new
  - Nonunique
    - nonuniq_Hearer_old
      - nonuniq_Physical_copresence
      - nonuniq_Larger_situation
      - nonunique_predicative_identity
    - nonuniq_Hearer_new_spec
    - nonuniq_nonspec
  - Generic_kindLevel
  - Generic_individualLevel
- Pleonastic
- Quantified
- Predicative_equative_role
- Part_of_noncompositionalMWE
- Measure_Nonreferential
- Other_Nonreferential

Figure 1: Version 2 of the Communicative Functions of Definiteness (CFD) Annotation Scheme.

Non-anaphoric NPs are those that have not been mentioned or evoked by something that was mentioned. They can be specific (*She wants to marry an Irishman. His name is Paul.*) or non-specific (*She wants to marry an Irishman. She should go and find one*). Some non-anaphoric nominals are known to the addressee because they are physically present or because of the situation that the speaker and hearer are in. For example, you can talk about *the hotel* or *the program chair* at a conference even when those things have not been previously mentioned. Non-anaphoric NPs also include those with unique, common-knowledge referents such as *the Empire State Building*, *Barack Obama*. Aside from anaphoric and non-anaphoric nominals, other categories are pleonastic, quantified, predicative, non-referential, and part of non-compositional multi-word expression.

A feature of CFD is that it is hierarchical in nature. This hierarchical organization of CFD serves to reduce the number of decisions that an annotator needs to make for speed and consistency.

Although the annotation scheme is somewhat stable,



most of all by her grandmother, and there was nothing that she would not have given to the child.

2 Once she gave her a little riding hood of red velvet, which suited her so well that she would never wear anything else; so she was always called 'Little Red Riding Hood.'

3 One day her mother said to her: 'Come, Little Red Riding Hood, here is a piece of cake and a bottle of wine; take them to your grandmother, she is ill and weak, and they will do her good.

4 Set out before it gets hot, and when you are going, walk nicely and quietly and do not run off the path, or you may fall and break the bottle, and then your grandmother will get nothing; and when you go into her room, don't forget to say, "Good morning", and don't peep into every corner before you do it.'

Figure 2: A snapshot of the Brat annotation tool with annotated data.

there are many issues about which we are still deliberating. For example, we have not yet finalized our approach to coordinate structures: whether to annotate each conjunct within a coordinate structure or whether to assign a single tag to the whole coordinate structure. Currently we do not analyze coordinated phrases internally. *Shoes or boots* is taken as one annotatable unit instead of analyzing it as consisting of 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*).

Figure 2 shows a part of the text from *Little Red Riding Hood* annotated with the CFD scheme.

4. Definiteness Corpus

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

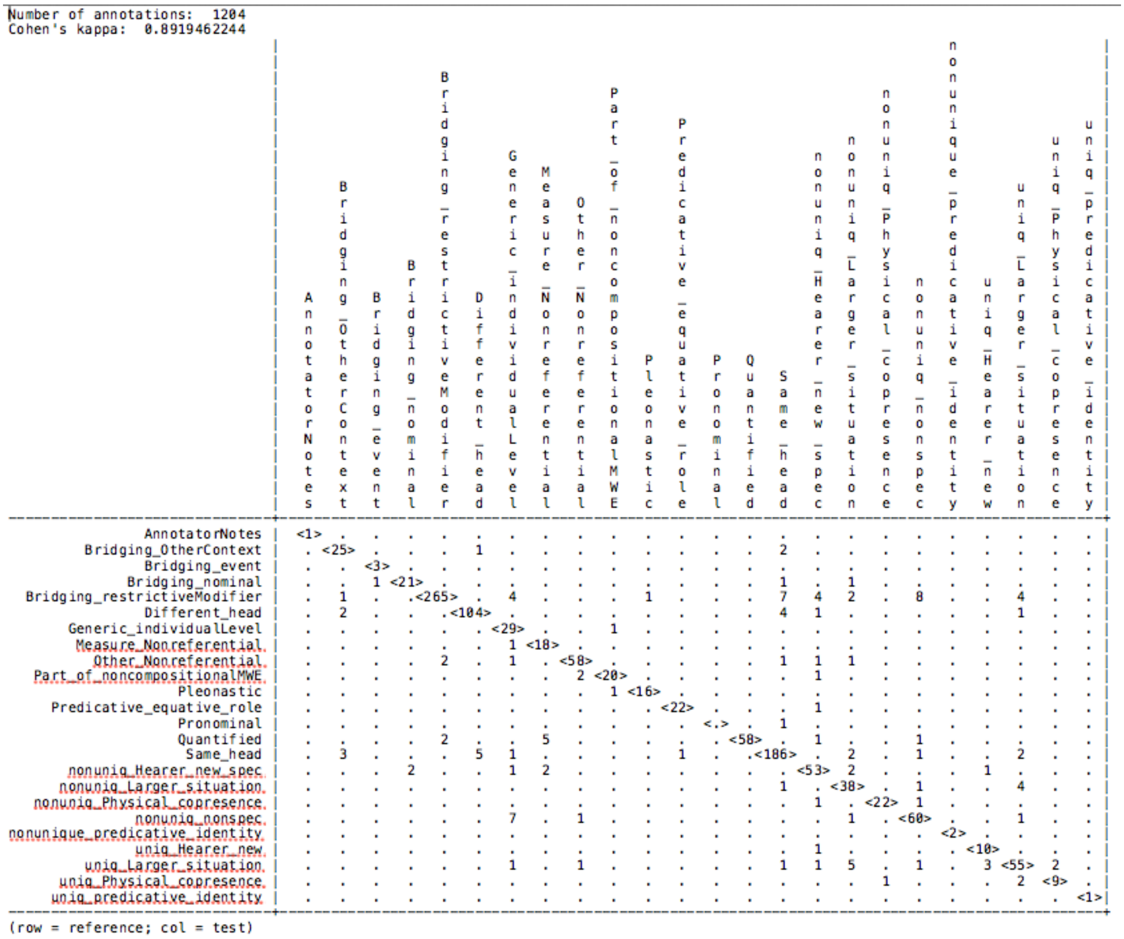
4.1. Data

We are annotating data from two languages currently, English and Hindi, most of which is parallel data (described below). 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. We have parallel data from the TED talks corpus in all four languages.[2] There are sixteen TED talks, which have been 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%. The Hindi corpus contains the same 16 TED talks that were annotated for English.

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



```
Number of annotations:  1204
Cohen's kappa:  0.8919462244
                               |                                                                                               n                     |
                               |                                                                                               o                     |
                               |                   B                                                                           n                     |
                               |                   r                        P                                            n     u                     |
                               |                   i                        a                                            o     n                     |
                               |                   d                        r          P                                 n     i                   u |
                               |                   g                        t          r                           n     u     q              u    n |
                               |                   i         G              _          e                      n    o     n     u              n    i |
                               |                   n         e    M         o          d                      o    n     i     e              i    q |
                               |        B          g         n    e         f          i                      n    u     q     _         u    q    _ |
                               |        r          _         e    a    O    _          c                      u    n     _     p         n    _    p |
                               |        i          r         r    s    t    n          a                      n    i     P     r         i    P    r |
                               |        d          e         i    u    h    o          t                      i    q     h     e         q    h    e |
                               |        g          s         c    r    e    n          i                      q    _     y     d         _    y    d |
                               |        i         B    t         _    e    r    c          v                      _    L     s     i         L    s    i |
                               |        n         r    r         i    _    _    o          e                      H    a     i    n    c    u    a    i    c |
                               |   A    g    B    i    i    D    n    N    N    m         _                   e    r    c    o    a    n    r    c    a |
                               |   n    _    r    d    c    i    d    o    o    p         e                   a    g    a    n    t    i    g    a    t |
                               |   n    O    i    g    t    f    i    n    n    o         q                   r    e    l    u    i    q    e    l    i |
                               |   o    t    d    i    i    f    v    r    r    s         u                   e    r    _    n    v    _    r    _    v |
                               |   t    h    g    n    v    e    i    e    e    i    P    a    P    Q         r    _    c    i    e    H    _    c    e |
                               |   a    e    i    g    e    r    d    f    f    t    l    t    r    u    S    _    s    o    q    _    e    s    o    _ |
                               |   t    r    n    _    M    e    u    e    e    i    e    i    o    a    a    n    i    p    _    i    a    i    p    i |
                               |   o    C    g    n    o    n    a    r    r    o    o    v    n    n    m    e    t    r    n    d    r    t    r    d |
                               |   r    o    _    o    d    t    l    e    e    n    n    e    o    t    e    w    u    e    o    e    e    s    e    e |
                               |   N    n    e    m    i    _    L    n    n    a    a    _    m    i    _    _    a    s    n    n    r    i    s    n |
                               |   o    t    v    i    f    h    e    t    t    l    s    r    i    f    h    s    t    e    s    t    _    t    e    t |
                               |   t    e    e    n    i    e    v    i    i    M    t    o    n    i    e    p    i    n    p    i    n    i    n    i |
                               |   e    x    n    a    e    a    e    a    a    W    i    l    a    e    a    e    o    c    e    t    e    o    c    t |
                               |   s    t    t    l    r    d    l    l    l    E    c    e    l    d    d    c    n    e    c    y    w    n    e    y |
-------------------------------+------------------------------------------------------------------------------------------------------------------------+
                AnnotatorNotes |  <1>   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    . |
         Bridging_OtherContext |   .  <25>   .    .    .    1    .    .    .    .    .    .    .    .    2    .    .    .    .    .    .    .    .    . |
                Bridging_event |   .    .   <3>   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    . |
              Bridging_nominal |   .    .    1  <21>   .    .    .    .    .    .    .    .    .    .    1    .    1    .    .    .    .    .    .    . |
  Bridging_restrictiveModifier |   .    1    .    . <265>   .    4    .    .    .    1    .    .    .    7    4    2    .    8    .    .    4    .    . |
                Different_head |   .    2    .    .    . <104>   .    .    .    .    .    .    .    .    4    1    .    .    .    .    .    1    .    . |
       Generic_individualLevel |   .    .    .    .    .    .  <29>   .    .    1    .    .    .    .    .    .    .    .    .    .    .    .    .    . |
        Measure_Nonreferential |   .    .    .    .    .    .    1  <18>   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    . |
          Other_Nonreferential |   .    .    .    .    2    .    1    .  <58>   .    .    .    .    .    1    1    1    .    .    .    .    .    .    . |
   Part_of_noncompositionalMWE |   .    .    .    .    .    .    .    .    2  <20>   .    .    .    .    .    1    .    .    .    .    .    .    .    . |
                    Pleonastic |   .    .    .    .    .    .    .    .    .    1  <16>   .    .    .    .    .    .    .    .    .    .    .    .    . |
     Predicative_equative_role |   .    .    .    .    .    .    .    .    .    .    .  <22>   .    .    .    1    .    .    .    .    .    .    .    . |
                    Pronominal |   .    .    .    .    .    .    .    .    .    .    .    .   <.>   .    1    .    .    .    .    .    .    .    .    . |
                    Quantified |   .    .    .    .    2    .    .    5    .    .    .    .    .  <58>   .    1    .    .    1    .    .    .    .    . |
                     Same_head |   .    3    .    .    .    5    1    .    .    .    .    1    .    . <186>   .    2    .    1    .    .    2    .    . |
       nonuniq_Hearer_new_spec |   .    .    .    2    .    .    1    2    .    .    .    .    .    .    .  <53>   2    .    .    .    1    .    .    . |
      nonuniq_Larger_situation |   .    .    .    .    .    .    .    .    .    .    .    .    .    .    1    .  <38>   .    1    .    .    4    .    . |
   nonuniq_Physical_copresence |   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    1    .  <22>   1    .    .    .    .    . |
               nonuniq_nonspec |   .    .    .    .    .    .    7    .    1    .    .    .    .    .    .    .    1    .  <60>   .    .    1    .    . |
nonunique_predicative_identity |   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .   <2>   .    .    .    . |
               uniq_Hearer_new |   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    1    .    .    .    .  <10>   .    .    . |
         uniq_Larger_situation |   .    .    .    .    .    .    1    .    1    .    .    .    .    .    1    1    5    .    1    .    3  <55>   2    . |
      uniq_Physical_copresence |   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    1    .    .    .    2   <9>   . |
     uniq_predicative_identity |   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .   <1>|
-------------------------------+------------------------------------------------------------------------------------------------------------------------+
(row = reference; col = test)
```

Figure 3: Confusion matrix for inter-annotator agreement on annotations of 4 texts from different genres combined together.

4.2. Annotation Tool

For annotating the data, we are using an existing annotation tool, the brat rapid annotation tool (`http://brat.nlplab.org`). A screenshot of the tool with annotated data is shown in fig. 2. The annotator selects the span (an NP) that is to be annotated. This opens a dialogue box with the labeling options from 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 annotation, containing the following information: an ID that is given to the annotation, the annotation label, the location of the annotated unit in the document in terms of character offsets, and the annotated NP itself.

4.3. Inter-annotator Agreement

For inter annotator agreement, we looked at the agreement in annotation labels assigned *given an NP as the annotatable unit*. Two English annotators independently annotated the same two texts from different genres (1 talk from the TED 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 and updated the annotation scheme. Once they reached consensus, they annotated four new texts without discussion and measured inter-annotator agreement. We found almost perfect agreement with Cohen's $\kappa = 0.89$ over a total of 1204 annotated units from the four texts. The confusion matrix is presented in fig. 3. On individual genres, the inter-annotator agreement scores are: Cohen's $\kappa = 0.89$ for TED (504 annotated NPs), 0.87 for presidential speech (492 annotated NPs), 0.95 for narrative (129 annotated NPs), 0.85 for news articles (79 annotated NPs).

5. Future Work and Conclusions

In this 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.

This corpus could be used for building classifiers for prediction of various grammaticalization strategies based on the CFD annotations. The main idea is that 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) can help improve MT across languages. For example, an option is to transform the source language to look like the target language in terms of grammatical expression of the annotated semantic features. The target-like source language can then be used for training a machine translation system. Another option is to provide in the training data the possible structures as complex structural tags which have been identified as corresponding to the annotated semantic features.

This paper is an attempt at the first step in the process—building a corpus that lets us determine the relevance of these semantic and pragmatic notions (communicative functions of definiteness) and the grammaticalization strategies. In our future work, we plan to use the current version of the CFD annotation scheme and the definiteness corpus described in §4.1 to build classifiers for predicting 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 more non-referential categories. Finally, we plan on using the classifiers we build to aid machine translation systems and coreference resolution 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.

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.

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. (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.

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.