

[^A_B Learning Definiteness across Languages using Classifiers] [^{NS}_S maybe: A Classification and Classifier for Cross-linguistic Functions of Definiteness]

Abstract

Definiteness is a nonhomogeneous category across languages expressing various *communicative functions* (types of semantic, pragmatic and discourse- related information). Languages differ in the grainsize and inventory of communicative functions that are expressed as well as the syntactic means for expressing them. In a previous publication we presented an annotation scheme for the communicative functions of definiteness and released an annotated corpus in English and Hindi. This paper reports on a classifier for English that uses lexical, morphological, and syntactic features to predict communicative functions. This work has twofold benefits: linguistically, the classifier's features and weights model the *grammaticalization* of definiteness in English, some of which is obvious (e.g., *the* and *a*) and some of which is not obvious (include an example from our findings). Computationally, it presents a framework to predict *semantic and pragmatic* communicative functions of definiteness which, unlike lexical and morphosyntactic features, are preserved in translation. This work may also be useful in tracking noun phrase reference for information extraction, entity tracking, and other semantic processing tasks.

1 Introduction

Languages display a vast range of variation with respect to definiteness. For example, while languages like English make use of definite and indefinite articles to distinguish between the discourse status of various entities (*the car* vs. *a car*), many other languages, such as Czech, Hindi, Indonesian, and Russian do not have articles (although

they do have demonstrative determiners). Even the languages that do have articles vary in terms of how these articles are used. For example, English uses only one definite article 'the' for entities irrespective of whether they have been mentioned previously in the discourse or are familiar to the hearer by some other means, but there are languages such as Hausa which have a specialized article for anaphoric definite entities, see example in (1) (borrowed from Lyons (1999)).

- (1) Na kawo keke din-ka
AUX bring bicycle the-2SGM
'I brought your bicycle (previously mentioned).'

There are other means to express (in)definiteness, such as the use of affixes. Arabic, for example, uses a definite prefix *al-* and indefinite suffix *-n*. Chen (2004) shows that Chinese expresses (in)definiteness through the use of various constructions, such as the existential construction for indefinite subjects and the *ba-* construction for definite direct objects. Demonstratives, personal pronouns and possessives (which are found in all languages) are other kinds of definite NPs. Aside from this variation in the form of (in)definite NPs within and across languages, there is also variability in what semantic, pragmatic, discourse- related functions (in)definites express. We will refer to these as *communicative functions*.

The literature on definiteness suggests that it may express communicative functions corresponding to notions such as uniqueness, familiarity, identifiability, anaphoricity, specificity, and referentiality (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*). Reductionist approaches to definiteness try to define a single communicative function of definiteness. For example, Kadmon (1987); Evans (1980) propose that seman-

tic uniqueness is the main communicative function of definite NPs but pronouns do not display uniqueness effects. Roberts (2003) proposes that the combination of uniqueness and a presupposition of familiarity can explain definiteness, but possessive definite descriptions (*John's daughter*) and the weak definites (*the son of Queen Juliana of the Netherlands*) are neither unique nor necessarily familiar to the listener before they are spoken.

We take such linguistic observations to suggest that definiteness is not as homogeneous a category as many accounts have assumed. In contrast to the reductionists, we are following an approach to grammaticalization (?) in which grammar develops over time in such a way that each grammatical construction has some prototypical communicative functions, but also has many non-prototypical communicative functions. Grammaticalization varies quite a lot across languages in both which communicative functions are grammaticalized and in which syntactic constructions are used to grammaticalize them (?).

This paper describes a classifier that predicts communicative function labels for English noun phrases. The features used by the classifier are lexical, morphological, and syntactic. The contribution of our work is in both the output of the classifier and the model that it uses (features and weights). The classifier outputs communicative function labels (see next section, e.g., whether the entities are old or new to the discourse and to the hearer). Communicative function is important because it is usually preserved in translation even when the grammatical mechanisms for expressing it are different. The communicative function labels also represent the discourse status of entities, making them relevant for entity tracking, knowledge base construction, and information extraction.

The model is a form-meaning mapping, consisting of the syntactic, lexical, and morphological features and weights that correlate with—are predictive of—communicative functions. This in itself is linguistically significant in that it shows the grammatical mechanisms beyond the articles *the* and *a* that are used for expressing definiteness in English. In future work we will build such models for languages that do not have articles such as Hindi, Russian, and Chinese. The form-meaning mapping for these languages can be used for machine translation applications. It has been noted

previously that machine translation systems face problems while translating from one language to another when the languages use different grammatical strategies. Tsvetkov et al. (2013); Stymne (2009) mention how translating from an article-language to an article-less language is problematic. If the mapping between different forms and the information they encode is known in both the source language and the target language, this information can be leveraged in improving machine translation across these languages.

In §2, we review the annotation scheme we have used to determine the mapping between the form and meaning it encodes with respect to (in)definiteness. In §3, we provide information about the English data we have annotated using this scheme. §4 is an overview of the classification framework we have used. In §5, we describe the measures we have used to evaluate the predictions against the gold standard for the held-out data. In §6, we describe our experiments and discuss the results. In §7, we summarize our findings and briefly discuss the future tasks we are undertaking or envision. [\[^Y_T I think that detailed description of definiteness in linguistics in the beginning of this section deserves a section by itself; I'd split the Introduction section into Introduction \(shorter, only what we do, motivation, contribution\) and then Definiteness.\]](#)

2 Annotation scheme

Based on the literature on definiteness, we compiled a list of semantic, pragmatic, and discourse-related features which are relevant for the form of (in)definite NPs in English (and many other languages). [\[^Y_T ref to the LREC paper?\]](#) These features were put together in a hierarchical structure to ease the annotation effort for the annotators. At each step, the annotators make a decision and thus reduce the number of labels they have to consider before annotating an NP. This approach also increases the consistency in annotations due to the reduced effort on considering the options among which to select the correct label for the NP. The annotation scheme was revised through many iterations after attempts at annotating natural language data from different genres using the scheme. The current scheme is presented in Figure 2.

The first distinction annotators make is whether the NP at hand is anaphoric (discourse old) or non anaphoric (discourse new). We do not annotate

(2) I went to a wedding last weekend. *The bride* was a friend of mine. She baked *the cake* herself.

for uniqueness, hearer old/new distinction, specificity or genericity for the anaphoric NPs as those distinctions were deemed inconsequential for English anaphoric NPs¹. [Y For someone who is not familiar with the scheme, this would be very hard to understand without looking at the definiteness scheme. I think the scheme should be on this page, and description should always point to the place in the scheme. May be we can write all scheme-related terms using the same font as they appear in the scheme?] The anaphoric NPs' category is further subdivided into basic anaphora and extended anaphora. The basic anaphora is used for entities which have been mentioned previously in the discourse. The extended anaphora is inspired by the bridging references Clark (1977) had mentioned. These are used for entities which are not discourse old or even hearer old, but at the same time, they are not entirely new either. See example in Figure ?? (borrowed from Poesio and Vieira (1998)).

Within the non anaphoric category, the annotators decide whether the NP denotes a unique NP, a non unique NP or a generic NP. For the non-generic cases, further decisions are made based on whether the entity is hearer old or new, and if it is old, what kind of situation makes it old. For generics also, there are two categories depending on whether they appear with kind-level or individual-level predicates resulting in different properties of these NPs. Besides these, there are a few non-referential categories as well.

Besides the annotation categories, a few other decisions were also taken regarding annotations. For example, corresponding to the basic anaphora cases, the anaphoric links can be established with antecedent NPs in previous sentences as well as with NPs within the same sentence. Also to determine whether an NP gets a Same_head or a Different_head label, we refer to the closest antecedent even if it appears within the same sentence. Regarding annotations of nested NPs, we assume that these NPs are interpreted inside out. This helps determine the cases of Bridging_RestrictiveModifier category. If an NP consists of a modifier that restricts its reference enough to make the referent identifiable to the addressees, the embedding NP is

¹However, in future, we may include these features for anaphoric NPs as well if contrary evidence were found.

annotated with the Bridging_RestrictiveModifier label.

Figure 1 is an excerpt from Little Red Riding Hood demonstrating the above-mentioned decisions regarding annotations and illustrating some of the categories from our annotation scheme.

[S^{NS} we should probably have a name for the scheme—something like “Functions of Definiteness Across Languages (FDAL)” and a name for the categories (I have been calling them semantic functions).][B^A I actually like the name “Functions of Definiteness Across Languages” but am not sure how to incorporate it here.] [S^{NS} suggestion: lead with a discourse excerpt, ideally illustrating a few of the categories, a nested NP, and an anaphoric link across sentence boundaries. I can help format it.]

[S^{NS} make sure to cite related computational stuff such as (Reiter and Frank, 2010)]

3 Data

We annotated data from a few genres, namely TED talks, news articles, speech (Presidential inauguration speech) and stories in English. However, currently the majority of our annotated data is from the TED talks corpus. For future work, we will be annotating other genres more as well. We annotated a total of 1783 sentences [B^A that is all our data, some is not yet annotated, am working on it today, hopefully we will be able to use most of it, otherwise we will change the numbers here] (a total of 28574 words), which consist of 10476 NPs (i.e. the annotatable units).

A total of 6 annotators contributed to the annotations. Two of the annotators are linguists while others are computer scientists working on languages and an undergraduate student of Linguistics. Two annotators annotated most of the data, the inter annotator agreement between them was very high with Cohen's Kappa score of 0.94 within the TED corpus and 0.91 for combined genres. However the other 4 annotators also annotated some of the data, but one of the two annotators with most experience took these data as inputs and revised annotations according to their annotations. This two-step process of annotation was taken to increase the amount of data within a short period of time and also to achieve consistency within annotations. All annotators were trained using the annotation guidelines and example annotated texts. Then the annotators were asked to

Once upon a time there was a dear little girl who was loved by everyone who looked at her, but most of all by her grandmother, and there was nothing that she would not have given to the child.

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.'

SAME_HEAD DIFFERENT_HEAD OTHER_NONREFERENTIAL SAME_HEAD
NONUNIQ_HEARER_NEW_SPEC
SAME_HEAD QUANTIFIED SAME_HEAD UNIQ_HEARER_NEW

Figure 1: An annotated sentence from “Little Red Riding Hood.” The previous sentence is shown for context.

- **NONANAPHORA** $[-A, -B]$ (00)
 - **UNIQUE** $[+U]$ (00)
 - * **UNIQ_HEARER_OLD** $[-G, +O, +S]$ (00)
 - UNIQ_PHYSICAL_COPRESENCE $[+R]$ (00)
 - UNIQ_LARGER_SITUATION $[+R]$ (00)
 - UNIQ_PREDICATIVE_IDENTITY $[+P]$ (00)
 - * UNIQ_HEARER_NEW $[-O]$
 - **NONUNIQUE** $[-U]$
 - * **NONUNIQ_HEARER_OLD** $[+O]$
 - NONUNIQ_PHYSICAL_COPRESENCE $[-G, +R, +S]$
 - NONUNIQ_LARGER_SITUATION $[-G, +R, +S]$
 - NONUNIQ_PREDICATIVE_IDENTITY $[+P]$
 - * NONUNIQ_HEARER_NEW_SPEC $[-G, -O, +R, +S]$
 - * NONUNIQ_NONSPEC $[-G, -S]$
 - **GENERIC** $[+G, -R]$
 - * GENERIC_KINDLEVEL
 - * GENERIC_INDIVIDUALLEVEL
- **ANAPHORA** $[+A]$
 - **BASIC** $[+O, -B]$
 - * SAME_HEAD
 - * DIFFERENT_HEAD
 - **EXTENDED** $[+B]$
 - * BRIDGING_NOMINAL $[-G, +R, +S]$
 - * BRIDGING_EVENT $[+R, +S]$
 - * BRIDGING_RESTRICTIVEMODIFIER $[-G, +S]$
 - * BRIDGING_SUBTYPE_INSTANCE $[-G]$
 - * BRIDGING_OTHERCONTEXT $[+O]$
- **MISCELLANEOUS** $[-R]$
 - PLEONASTIC $[-B, -P]$
 - QUANTIFIED
 - PREDICATIVE_EQUATIVE_ROLE $[-B, +P]$
 - PART_OF_NONCOMPOSITIONAL_MWE
 - MEASURE_NONREFERENTIAL
 - OTHER_NONREFERENTIAL

Figure 2: Taxonomy of definiteness functions, with number of occurrences in the training data [_S^{NS} TODO]. Internal (non-leaf) labels are in bold; these are not annotated or predicted. [_S^{NS} TODO: normalize capitalization] +/- values are shown for ternary attributes Anaphoric, Bridging, Generic, Hearer-Old, Predicative, Referential, Specific, and Unique; these are inherited from supercategories, but otherwise default to 0. Thus, for example, the full attribute specification for UNIQ_PHYSICAL_COPRESENCE is $[-A, -B, -G, +O, 0P, +R, +S, +U]$.

annotate some data, these annotations were discussed to reach at consensus. The annotations used for discussions to reach at consensus were not included while calculating the inter annotator agreement.

[_B^A description of data, IAA, other statistics-n(annotations), n(annotations/ category) classifiers for English- baseline, accuracy etc, confusion matrix- classifier predictions and gold standard may be a brief discussion of feature ablation study- which features are most helpful at least in these languages corresponding to individual categories (a clue to grammaticalization) Principle component analysis, clustering of features]

4 Classification framework

To model the relationship between the grammar of definiteness and its semantic functions in a data-driven fashion, we work within the supervised framework of feature-rich discriminative classification, treating the functional categories from §2 as output labels y and various lexical, morpholog-

ical, and syntactic characteristics of the language as features of the input x . Specifically, we learn a probabilistic log-linear model similar to multiclass logistic regression, but deviating in that logistic regression treats each output label (response) as atomic, whereas we decompose each into *attributes* based on their linguistic definitions, enabling commonalities between related labels to be recognized. Each weight in the model corresponds to a feature that mediates between *percepts* (characteristics of the input noun phrase) and attributes (characteristics of the label). This is aimed at attaining better predictive accuracy as well as feature weights that better describe the form-function interactions we are interested in recovering.

Our setup is formalized below, where we discuss the mathematical model and linguistically motivated features.

4.1 Model

At test time, we model the probability of semantic label y conditional on a [_S^{NS} gold?] noun phrase x

as follows:

$$p_{\theta}(y|x) = \log \frac{\exp \theta^T \mathbf{f}(x,y)}{\sum_{y' \in \mathcal{Y}} \exp \theta^T \mathbf{f}(x,y')} \quad (1)$$

where $\theta \in \mathbb{R}^d$ is a vector of parameters (feature weights), and $\mathbf{f} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$ is the feature function over input-label pairs. The feature function is defined as follows:

$$\mathbf{f}(x,y) = \phi(x) \times \tilde{\omega}(y) \quad (2)$$

where the percept function $\phi : \mathcal{X} \rightarrow \mathbb{R}^c$ produces a vector of real-valued characteristics of the input, and the attribute function $\tilde{\omega} : \mathcal{Y} \rightarrow \{0,1\}^a$ encodes characteristics of each label. There is a feature for every percept-attribute pairing: so $d = c \cdot a$ and $f_{(i-1)a+j}(x,y) = \phi_i(x) \tilde{\omega}_j(y)$, $1 \leq i \leq c$, $1 \leq j \leq a$. The contents of the percept and attribute functions are detailed in §§4.3 and 4.2.

For prediction, having learned weights $\hat{\theta}$ we choose the y that maximizes this probability:

$$\hat{y} \leftarrow \arg \max_{y' \in \mathcal{Y}} p_{\hat{\theta}}(y'|x) \quad (3)$$

Training optimizes $\hat{\theta}$ so as to maximize a convex L_1 -regularized learning objective over the training data \mathcal{D} :

$$\hat{\theta} = \arg \max_{\theta} L(\theta, \mathcal{D}) \quad (4)$$

$$L(\theta, \mathcal{D}) = -\lambda \|\theta\|_1 + \sum_{(x,y) \in \mathcal{D}} \log \frac{\exp \theta^T \mathbf{f}(x,y)}{\sum_{y' \in \mathcal{Y}} \exp(\theta^T \mathbf{f}(x,y'))} \quad (5)$$

With $\tilde{\omega}(y) = \text{the identity of the label}$, this reduces to standard logistic regression.

4.2 Features

Features of words of interest The text is parsed with a dependency parser. In an NP parse there are certain words that we are mostly interested in. They include its *head*, *dependents*, and *outer heads* (which are words that the heads depend on). We are also interested in the *attached verb*, which is the first verb one encounters when going upward the dependency path from the head. We extract their part-of-speech, lemmas, and grammatical information (of which we only collect plurality right now). Since there may be multiple dependents, we have special features for the first and the last one. Moreover, to better

capture tense, aspect and modality, we collect the attached verb's *aux* words. We also make note of *neg* if it is attached to the verb. Of these words we collect their lemmas, POS's and dependency relation. Since there may be many dependents, we also specifically mark the first and the last word.

Parse structure related features These features include length of the upward path from the head to the root, and to the attached verb. We also have features for the number of dependents, and the number of dependency relations that link non-neighbors.

Position related features Including the length of the NP, NP position in the sentence (upper half or lower half), relative position of the verb to the head (left or right). We also collect information of the words of interest' left and right neighbors.

Other NPs These include features in annotated NPs such as *immediate parent*, which is the smallest NP that fully covers this NP, and similarly *immediate child*. We also have *immediate precedent* and *immediate successor*. Using a coreference resolution program, we identify mentions, and extract their features as well.

4.3 Attributes

As noted above, though labels are organized into a tree hierarchy, there are actually several dimensions of commonality that suggest different groupings. These attributes are encoded as ternary characteristics; for each label (including internal labels), every one of the 8 attributes is assigned a value of +, −, or 0 (refer to fig. 2). In order to capture these similarities in the model's features, we define the attribute vector function $\omega(y) =$

$$[y, A(y), B(y), G(y), O(y), P(y), R(y), S(y), U(y)]^T$$

where $A : \mathcal{Y} \rightarrow \{+, -, 0\}$ returns the value for Anaphoric, $B(y)$ for Bridging, etc. The identity of the label is also included in the vector so that different labels are always recognized as different by the attribute function. The categorical components of this vector are then binarized to form $\tilde{\omega}(y)$; however, instead of a binary component that fires for the 0 value of each ternary attribute, there is a component that fires for *any* value of

the attribute—a sort of bias term. The weights assigned to features incorporating + or − attribute values, then, are easily interpreted as deviations relative to the bias.

5 Evaluation

The following measures will be used to evaluate our predictor against the gold standard for the held-out evaluation (dev or test) set \mathcal{E} :

- **Exact match:** This accuracy measure gives credit only where the predicted and gold labels are identical.
- **By leaf label:** We also compute precision and recall of each leaf label to determine which categories are reliably predicted.
- **Soft match:** This accuracy measure gives partial credit where the predicted and gold labels are related. It is computed as in §5.
- **Perplexity:** This determines how “surprised” our model is by the gold labels in the test set; the greater the probability mass assigned to the true labels, the higher the score. It is computed as $2^{(\sum_{(x,y) \in \mathcal{E}} \log_2 p_{\hat{\theta}}(y|x)) / |\mathcal{E}|}$.

6 Experiments

$[\text{S}^{\text{NS}} \text{ English: } \pm \text{cost function, } \pm \text{non-identity attributes, } \pm \text{predicting intermediate labels}]$
 $[\text{S}^{\text{NS}} \text{ maybe: which attribute groupings produce the best classifier, if we want to force a hierarchy}]$
 $[\text{S}^{\text{NS}} \text{ feature/attribute ablations}]$
 $[\text{S}^{\text{NS}} \text{ Hindi?}]$

7 Conclusion

References

- B. Birner and G. Ward. 1994. Uniqueness, familiarity and the definite article in english. In *Proc. of the Twentieth Annual Meeting of the Berkeley Linguistics Society*, pages 93–102.
- P. Chen. 2004. Identifiability and definiteness in chinese. *Linguistics*, 42:1129–1184.
- H.H. Clark. 1977. Bridging. In P.N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press.
- C. Condoravdi. 1992. Strong and weak novelty and familiarity. In *Proc. of SALT II*, pages 17–37.
- C. Evans. 1977. Pronouns, quantifiers and relative clauses. *Canadian Journal of Philosophy*, 7(3):46.
- C. Evans. 1980. Pronouns. *Linguistic Inquiry*, 11.
- J.K. Gundel, N. Hedberg, and R. Zacharski. 1988. The generation and interpretation of demonstrative expressions. In *Proc. of XIIIth International Conference on Computational Linguistics*, pages 216–221.
- J.K. Gundel, N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.
- I. Heim. 1990. E-type pronouns and donkey anaphora. *Linguistics and Philosophy*, 13:137–177.
- N. Kadmon. 1987. *On unique and non-unique reference and asymmetric quantification*. Ph.D. thesis, University of Massachusetts.
- N. Kadmon. 1990. Uniqueness. *Linguistics and Philosophy*, 13:273–324.
- C. Lyons. 1999. *Definiteness*. Cambridge University Press. URL <http://books.google.com/books/about/Definiteness.html?id=1MeX8240YTUC>.
- M. Poesio and R. Vieira. 1998. A corpus based investigation of definite description use. *Computational Linguistics*, 24:183–216.
- E.F. Prince. 1992. The zpg letter: Subjects, definiteness and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins.
- Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proc. of ACL*, pages 40–49. Uppsala, Sweden.
- C. Roberts. 2003. Uniqueness in definite noun phrases. *Linguistics and Philosophy*, 26:287–350.
- B. Russell. 1905. On denoting. *Mind, New Series*, 14:479–493.
- S. Stymne. 2009. Definite noun phrases in statistical machine translation into danish. In *Proc. of Workshop on Extracting and Using Constructions in NLP*, pages 4–9.
- Yulia Tsvetkov, Chris Dyer, Lori Levi, and Archana Bhatia. 2013. Generating English determiners in phrase-based translation with synthetic translation options. In *Proc. of WMT*. Association for Computational Linguistics.