Title, author list not final. Page limit: 5 + 2 bib

A Classifier for Communicative Functions of Definiteness

Abstract

Whether a noun phrase is realized grammatically as definite or not depends on a variety of semantic, pragmatic, and discourse criteria, or *communicative functions*, the interaction of which varies from language to language. This paper reports on a supervised classifier for English that uses lexical, morphological, and syntactic features to predict communicative functions of definiteness. The benefits of this work are twofold: linguistically, the classifier's features and weights model the *grammaticalization* of definiteness in English, not all of which are obvious. Computationally, it presents a framework to predict *semantic and pragmatic* communicative functions of definiteness which, unlike lexical and morphosyntactic features, are preserved in translation. The classifier may also be useful in tracking noun phrase reference for semantic processing tasks.

1 Introduction

Languages display a vast range of variation with respect to the form and meaning of 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). Some languages such as Hausa (Lyons, 1999) have different definite articles for noun phrases that have been previously mentioned in contrast to those that are definite by virtue of the situation (e.g., "the podium" at a conference). Definiteness can also be expressed by affixes as in Arabic. Chen (2004) shows that Chinese, a language without articles, expresses (in)definiteness through 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 the semantic, pragmatic, and discourse–related functions expressed by (in)definites. We will refer to these as *communicative functions*. The literature on definiteness describes functions including 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 one or two communicative functions of definiteness. For example, Kadmon (1987); Evans (1980) propose that semantic uniqueness is the main communicative function of definite NPs. Roberts (2003) proposes that the combination of uniqueness and a presupposition of familiarity underlie all definite descriptions. However, 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.

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.

To build our model, we leverage a cross-lingual definiteness annotation scheme (§2) and annotated English corpus (§3) from prior work (Bhatia et al., 2014). Our classifier, §4, is a supervised log-linear model akin to logistic regression, with features that combine lexical and morphosyntactic information with prespecified groupings of the communicative function labels; the evaluation measures (§5) include one that exploits these label groupings to award partial credit according to relatedness. [$^{NS}_{S}$ TODO: §6 obtain good performance? discover interesting features?]

2 Annotation scheme

We give an overview of the annotation scheme for Communicative Functions of Definiteness (CFD), as described in Bhatia et al. (2014). It is summarized in fig. 1. The hierarchical organization serves to reduce the number of decisions that an annotator needs to make for speed and consistency. The scheme was developed by annotating texts in two languages (English and Hindi) and various genres.

CFD assigns a communicative function label to every noun phrase except for first- and second-person pronouns. The three main communicative functions in the annotation scheme are **Anaphora** vs. **Nonanaphora** (whether the entity is new to to the discourse or not), **Hearer-old** vs. **Hearer-new**, and **Unique** vs. **Nonunique** (annotated for **Nonanaphoric** only in the current scheme). However there are a few twists. Entities that have not been mentioned are considered **Anaphoric** (discourse-old) if they are evoked by a previously mentioned entity. For example, after mentioning a wedding, *the bride*, *the groom*, and *the cake* are considered to be **Anaphoric** (Clark, 1977; Poesio and Vieira, 1998). Entities that are **Non-Anaphoric** can be **Hearer-old** if they are physically present in the speech situation (**Physical-copresence**) or is not co-present but is part of the **Larger-situation** (e.g., spoken on the first day of a conference: “I’m tired. The airplane was noisy.”).[1]

In addition to the three main communicative functions, we have annotations for generic, pleonastic, quantified, predicative, and non-referential noun phrases.

Figure 2 is an excerpt from the “Little Red Riding Hood” annotated with the CFD scheme.

The following examples from Croft (2003, pp. 6–7) illustrate the grammatical expression of three communicative functions in English and French. In the CFD scheme, these should receive labels NONUNIQ_HEARER_NEW_SPEC, OTHER_NONREFERENTIAL, GENERIC_INDIVIDUALLEVEL, and PREDICATIVE_EQUATIVE_ROLE, respectively. The first is expressed the same (with a definite article) in English and French, whereas the other two are expressed differently in the two languages.

[1] Komen (2013) proposed a hierarchy with similar leaf nodes, but different internal structure.

- **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 1: CFD (Communicative Functions of Definiteness) annotation scheme, with number of occurrences in the training data [NS S TODO]. Internal (non-leaf) labels are in bold; these are not annotated or predicted.[NS S 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 $[-A,-B,-G,+O,0P,+R,+S,+U]$.

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]SAME_HEAD gave [her]DIFFERENT_HEAD [a little riding hood of [red velvet]OTHER_NONREFERENTIAL]NONUNIQ_HEARER_NEW_SPEC , which suited [her]SAME_HEAD so well that [she]SAME_HEAD would never wear [anything else]QUANTIFIED; so [she]SAME_HEAD was always called [‘Little Red Riding Hood.’]UNIQ_HEARER_NEW

Figure 2: An annotated sentence from “Little Red Riding Hood.” The previous sentence is shown for context.

(1) a. He went to **the bank**. (def.)
Il est allé à **la banque**. (def.)

b. He showed **extreme care**. (unmarked)
Il montra **un soin extrême**. (indef.)

c. I love **artichokes** and asparagus. (unmarked)
J’aime **les artichauts** et les asperges. (def.)

d. His brother became **a soldier**. (indef.)
Son frère est devenu **soldat**. (unmarked)

3 Data

We use the English definiteness corpus of Bhatia et al. (2014), which consists of texts from multiple genres annotated with the scheme described in §2. The 16 documents are from prepared speeches (TED talks and a presidential address), published news articles, and fictional narratives. The TED data predominates (about 72% of the corpus); the presidential speech represents 18%, news articles 5%, and fictional narratives 5%. All told, the corpus contains 20,655 words (812 sentences), with 2,950 NPs (the annotatable units). Bhatia et al. (2014) report high inter-annotator agreement, estimating Cohen’s $\kappa = 0.94$ within the TED genre and 0.91 for combined genres.

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, morphological, 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 [$^{\text{NS}}_{\text{S}}$ gold?] noun phrase x as follows:

$$p_{\boldsymbol{\theta}}(y|x) = \log \frac{\exp \boldsymbol{\theta}^{\top}\mathbf{f}(x,y)}{\sum_{y' \in \mathcal{Y}} \exp \boldsymbol{\theta}^{\top}\mathbf{f}(x,y')} \tag{1}$$

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

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

where the percept function $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^c$ produces a vector of real-valued characteristics of the input, and the attribute function $\tilde{\boldsymbol{\omega}} : \mathcal{Y} \to \{0,1\}^a$ 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 $\hat{\boldsymbol{\theta}}$ we choose the y that maximizes this probability:

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

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

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

$$L(\boldsymbol{\theta}, \mathcal{D}) = -\lambda ||\boldsymbol{\theta}||_1 + \sum_{\langle x,y \rangle \in \mathcal{D}} \log \frac{\exp \boldsymbol{\theta}^{\top}\mathbf{f}(x,y)}{\sum_{y' \in \mathcal{Y}} \exp \left(\boldsymbol{\theta}^{\top}\mathbf{f}(x,y')\right)} \tag{5}$$

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

4.2 Percepts

The characteristics of the input that are incorporated in the model, which we call *percepts* to distinguish them from model features linking inputs to outputs,[2] are intended to capture the aspects of English morphosyntax that may be relevant to the semantic and pragmatic functions of definiteness.

After preprocessing the text with a dependency parser and coreference resolver, we extract the several kinds of percepts for each noun phrase (NP).

4.2.1 Basic

Words of interest. These are the *head* within the NP, all of its *dependents*, and its *governor* (external to the NP). We are also interested in the *attached verb*, which is the first verb one encounters when traversing the dependency path upward from the head. For each of these words, we have separate percepts capturing: the token, the part-of-speech (POS) tag, the lemma, the dependency relation, and (for the head only) a binary indicator of plurality (determined from the POS tag). As there may be multiple dependents, we have additional features specific to the first and the last one. Moreover, to better capture tense, aspect and modality, we collect the attached verb's *auxiliaries*. We also make note of *neg* if it is attached to the verb.

Structural. These are: the *path length* from the head up to the root, and to the attached verb. We also have percepts for the number of dependents, and the number of dependency relations that link non-neighbors. Integer values were binarized with thresholding.

Positional. The *token length* of the NP; the NP's *location* in the sentence (first or second half); the *attached verb's position* relative to the head (left or right). 12 additional percept templates record the POS and lemma of the left and right neighbors of the head, governor, and attached verb.

4.2.2 Contextual NPs

When extracting features for a given NP (call it the "target"), we also consider NPs in the following relationship with the target NP: its *immediate parent*, which is the smallest NP whose span fully subsumes that of the target; the *immediate child*, which is the largest NP subsumed within the target; the *immediate precedent* and *immediate successor* within the sentence; and the *nearest preceding coreferent mention*.

For each of these related NPs, we include all of their basic percepts conjoined with the nature of the relation to the target.

[2] See above.

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. 1). In order to capture these similarities in the model's features, we define the attribute vector function $\boldsymbol{\omega}(y) =$

$$[y, A(y), B(y), G(y), O(y), P(y), R(y), S(y), U(y)]^{\top}$$

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 $\tilde{\boldsymbol{\omega}}(y)$; 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 the proportion of attributes whose (categorical) values match: $|\boldsymbol{\omega}(y) \cap \boldsymbol{\omega}(y')|/9$.
- **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 $2^{\left(\sum_{\langle x,y\rangle\in\mathcal{E}} \log_2 p_{\hat{\boldsymbol{\theta}}}(y|x)\right)/|\mathcal{E}|}$.

6 Experiments

6.1 Experimental Setup

The annotated corpus of Bhatia et al. (2014) (§3) contains 16 documents in 3 genres: 12 prepared speeches (mostly TED talks), 2 newspaper articles, and 2 fictional narratives. We arbitrarily choose some documents to hold out from each genre; the resulting test set consists of 2 TED talks ("Alisa_News", "RobertHammond_park"), 1 newspaper article ("crime1_iPad_E"), and 1 narrative ("Little Red Riding Hood"). The test set then contains 3,558 tokens (110 sentences), in which there are 492 annotated NPs; while the training set contains 2,458 NPs among 17,097 tokens (702 sentences). Gold NP boundaries are assumed throughout our experiments.

We use an in-house implementation of supervised learning with L_1-regularized AdaGrad (Duchi et al., 2011). Hyperparameters are tuned on a dev set formed by holding out every tenth instance from the training set (test set experiments use the full training set).[$^{\text{NS}}_{\text{S}}$ early stopping? what exactly is tuned? optimize soft match acc?] Automatic dependency parses and coreference information were obtained with the parser and coreference resolution system in Stanford CoreNLP v. 3.3.0 (Socher et al., 2013; Recasens et al., 2013) for use in features (§4.2).

6.2 Results

[$^{\text{NS}}_{\text{S}}$ English: ±cost function, ±non-identity attributes, ±predicting intermediate labels]

[$^{\text{NS}}_{\text{S}}$ maybe: which attribute groupings produce the best classifier, if we want to force a hierarchy]

[$^{\text{NS}}_{\text{S}}$ feature/attribute ablations]

[$^{\text{NS}}_{\text{S}}$ Hindi?]

7 Related Work

[$^{\text{NS}}_{\text{S}}$ computational approaches to things related to definiteness, e.g. in MT. also would be good to mention Bresnan's work on predicting syntactic alternations with logistic regression (here we want to predict the hidden information so that the classifier is useful for applications!).]

Automated prediction of (in)definite articles has been found beneficial in a variety of applications, including postediting of MT output (Knight and Chander, 1994), text generation (Elhadad, 1993; Minnen et al., 2000), and identification and correction of ESL errors (Han et al., 2006; De Felice and Pulman, 2008; Gamon et al., 2008; Rozovskaya and Roth, 2010). More recently, Tsvetkov et al. (2013) trained a classifier to predict where English articles might plausibly be added or removed in a phrase, and used this classifier to improve the

Condition	$\lvert\boldsymbol{\theta}\rvert$	Exact Match Accuracy	Soft Match Accuracy	Perplexity
Majority baseline	—			
Log-linear classifier, no grouping by attributes				
Full log-linear classifier				

Table 1: Classifier versus baselines.

quality of statistical machine translation. To the best of our knowledge, no studies have been conducted on automatic prediction of semantic and pragmatic properties of definiteness.

8 Conclusion

References

Archna Bhatia, Mandy Simons, Lori Levin, Yulia Tsvetkov, Chris Dyer, and Jordan Bender. 2014. A unified annotation scheme for the semantic/pragmatic components of definiteness. In *Proc. of LREC*. Reykjavík, Iceland.

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.

William Croft. 2003. *Typology and Universals*. Cambridge University Press.

Rachele De Felice and Stephen G Pulman. 2008. A classifier-based approach to preposition and determiner error correction in l2 english. In *Proc. of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 169–176. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Michael Elhadad. 1993. Generating argumentative judgment determiners. In *Proc. of AAAI*, pages 344–349.

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

C. Evans. 1980. Pronouns. *Linguistic Inquiry*, 11.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for esl error correction. In *Proc. of IJCNLP*, volume 8, pages 449–456.

J.K. Gundel, N. Hedberg, and R. Zacharski. 1988. The generation and interpretation of demonstrative expressions. In *Proc. of XIIth 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.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers.

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.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *Proc. of the National Conference on Artificial Intelligence*, pages 779–779. Seattle, WA.

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

C. Lyons. 1999. *Definiteness*. Cambridge University Press. URL `http://books.google.com/books/about/Definiteness.html?id=1MeX8240YTUC`.

Guido Minnen, Francis Bond, and Ann Copestake. 2000. Memory-based learning for article generation. In *Proc. of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 43–48. Association for Computational Linguistics.

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.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: identifying singleton mentions. In *Proc. of NAACL-HLT*, pages 627–633. Atlanta, Georgia, USA.

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

Alla Rozovskaya and Dan Roth. 2010. Training paradigms for correcting errors in grammar and usage. In *Proc. of NAACL-HLT*, pages 154–162. Association for Computational Linguistics.

B. Russell. 1905. On denoting. *Mind, New Series*, 14:479–493.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with compositional vector grammars. In *Proc. of ACL*, pages 455–465. Sofia, Bulgaria.

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 Archna Bhatia. 2013. Generating English determiners in phrase-based translation with synthetic translation options. In *Proc. of WMT*. Association for Computational Linguistics.