Towards proper name generation: A corpus analysis

Thiago Castro Ferreira and **Sander Wubben** and **Emiel Krahmer**
Tilburg center for Cognition and Communication (TiCC)
Tilburg University
The Netherlands
{tcastrof,s.wubben,e.j.krahmer}@tilburguniversity.edu

Abstract

We introduce a corpus for the study of proper name generation. The corpus consists of proper name references to people in webpages, extracted from the Wikilinks corpus. In our analyses, we aim to identify the different ways, in terms of length and form, in which a proper names are produced throughout a text.

1 Introduction

In natural language generation systems, referring expression generation (REG) is the process of producing references to discourse entities. Among the referential forms which can be used to distinguish an entity, proper names are an important and commonly used one. For instance, Ferreira et al. (2016) showed that writers produce a proper name as a first mention to an entity in 91% of the analysed texts.

In generation systems, not only the choice of whether a proper name should be generated is important, but also which *form* the proper name should take. For instance, *Barack Hussein Obama II* is the birth name of the 44th president of United States of America. However, he is also commonly referred to as *Barack Obama*, *Obama*, *President Obama*, etc. How to automatically decide which form to use?

In this paper, we introduce a new corpus of 53,102 proper names referring to people in 15,241 texts[1]. We analyse the corpus in terms of distribution of proper name lengths, intuitively expecting an inversely proportional relation between length of a name and sentence number in a text. We also analyse these references in terms of the presence of the first, middle and last name of the entity; and whether the reference is accompanied by a title or an appositive.

2 Related Studies

Unlike the generation of descriptions (Krahmer and van Deemter, 2012), only a few studies have focussed on the automatic generation of proper names. Reiter and Dale (2000) suggests the use of a full proper name for initial reference, optionally followed by an appositive to indicate properties of the entity important for the discourse. However, their approach does not account for variation in proper name references.

Van Deemter (2014) argues that proper name variants can be generated using standard algorithms for the generation of descriptions. In other words, van Deemter (2014) proposes describing proper names based on a knowledge base of attribute-value pairs. Just like a set of attribute-value pairs $\{(type, cube), (color, blue)\}$ is generated when the target needs to be singled out from differently coloured objects, a proper name like *Frida Kahlo* can be seen to single out one person from a context set. When the set is smaller, generally a shorter name will suffice. Van Deemter, however, does not apply this model in the context of text generation.

Siddharthan et al. (2011) presented a model to (re)generate referring expressions to people in extractive summaries. When generating a proper name, the model chooses between a full name or only a surname. Moreover, it also decides whether

[1] https://ilk.uvt.nl/~tcastrof/regnames

Proceedings of The 9th International Natural Language Generation conference, pages 222–226,
Edinburgh, UK, September 5-8 2016.

to use pre- (role, affiliation and temporal modifiers) or post-modifiers (appositives and relative clauses). As far as we know, this is the only study that introduced a corpus analysis of how humans produce proper names in a discourse. However, it only distinguished proper names among full names and surnames in a small set of 876 news texts.

3 Data Gathering

3.1 Materials

To analyse how proper names are used in text, we analysed webpages from the Wikilinks corpus (Singh et al., 2012). This corpus was originally created to study cross-document coreference and comprises around 40 million mentions to 3 million entities. All the mentions were extracted automatically by finding hyperlinks to Wikipedia pages related to the entities.

To collect our data, we identified the 1,000 most frequently mentioned people in the corpus. To determine which entities are persons, we used DBpedia, a database that provides structured information from Wikipedia (Bizer et al., 2009). From the Wikilinks corpus, we then randomly chose a subset of webpages that contain at least one mention to one of the most frequently mentioned persons. In total, our corpus contains texts from 15,241 webpages.

3.2 Annotation

To annotate the proper name references, we created a knowledge base which describes all variations of a proper name for the studied persons. We also parsed the webpages to identify in which part of the discourse the different proper name references were used. The annotation procedure is explained in more detail below.

Proper Names Knowledge Base We used two ontologies present on DBpedia to extract different proper names for the studied entities. The FOAF (*Friend-of-a-Friend*) ontology was used to extract the name (foaf:name), the given name (foaf:givenName) and the surname (foaf:surname) of a person. From the DBpedia ontology, we extracted the birth name of the entities (dbo:birthName).

Based on the proper names collected in DBpedia, we created a knowledge base by identifying 3 proper name attributes: **first name**, **middle name** and **last name**. First names consist of the first token from the name, given name and birth name, whereas last names consist of the token from the surname and the last tokens from the name and birth name. Middle names are all the tokens which are not the first token in the given and birth names and last token in the name and birth name. For instance, *Charles Bukowski* has *Charles*, *Bukowski*, *Charles Bukowski* and *Heinrich Karl Bukowski* as his given name, surname, name and birth name in DBpedia, respectively. Based on this information, the knowledge base for this entity would consist of *Charles* and *Heinrich* as first names; *Karl* as middle name; and *Bukowski* as last name.

Discourse Annotation The webpages were parsed using the Stanford CoreNLP software (Manning et al., 2014). Using this tool, we performed part-of-speech tagging, lemmatization, named entity recognition, dependency parsing, syntactic parsing, sentiment analysis and coreference resolution.

To improve the coreference resolution we performed a post hoc sanity check, to see whether references which were labelled as being to the same entity were correct. For each entity distinguished by the software, we checked the proper nouns of each proper name reference. If at least the proper nouns of one proper name were values present in the knowledge base of the target entity, all the references of the entity distinguished by the software were considered references to the target entity.

Once the references to the target entity were distinguished, we annotated their syntactic positions based on the output of the dependency parser and their referential statuses in the text and in the sentence - whether a reference is a first or an old mention to an entity. We also checked for the presence of a title or an appositive in the proper name references. These features were extracted based on the named entity recognition and dependency parser, respectively. In total, 53,102 proper name references were annotated in this way (an average of 3 per text).

3.3 Analyses

To analyse how proper names referring to people are distributed over a text, we checked the length of these references in terms of tokens. We also anal-

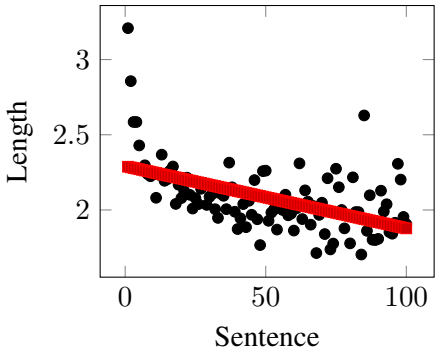


Figure 1: Average length of the proper names in tokens by sentence.

Title	2.4%
First Name	59.3%
Middle Name	7.1%
Last Name	89%
Appositive	1.7%

Table 1: Percentage of the proper name attributes

ysed the possible variations of a proper name by checking the presence of the first, middle and last name of the entity, and whether the proper name was accompanied by a title or an appositive.

4 Results

Figure 1 depicts the average length of proper name references in the first 100 sentences of the texts. A linear regression clearly shows that the length of a proper name decreases along the text, as predicted. Table 1 summarized the percentage of proper name attributes. It reveals that the last name is the most used one, followed by first name. The others occur less frequently.

Figure 2 shows the average length of proper name references as a function of syntactic position and referential status. Proper names in the object role of a sentence are generally longer than those in subject position (a); proper names that are new in the text are longer than those that have been mentioned in the text before, and vice versa when looking at new/old references per sentence (b).

Table 2 depicts frequency of various attribute sets, as a function of syntactic position and referential status in the text and sentence. Proper names consisting of both first and last name are the most common in the corpus. This proper name form is the most common one in the subject role of a sentence and as a mention to a new entity in the discourse. On the other hand, in the object role of a sentence and as mention to an old entity in the text, the use of only the last name is most common.

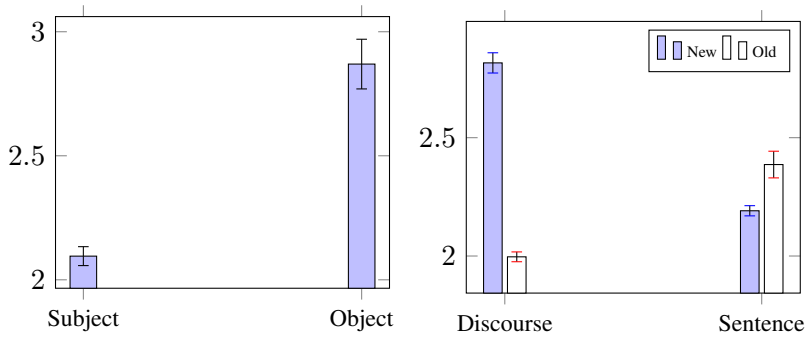
In general, proper names described by the first and last names, and by the first, middle and last names occur more often in the subject role of a sentence as a mention to a new entity in the text. The combination of first and last names is also more likely as a mention to old entities in the sentence. Proper names described by just one proper name attribute reveal the opposite behaviour, occurring more in the object role of a sentence as a mention to an old entity in the text or new in the sentence.

5 Discussion

This study introduced a corpus for the study of proper name generation. We analysed the different forms in which proper name references occur in text by checking their length as well as the occurrence of different proper name attributes including the first, middle, last names of the mentioned entity, as well as possible modifiers, such as titles or appositives.

Analyses revealed that longer proper names - in terms of number of tokens and proper name attributes - are more likely to be generated early in the text, in the object role of a sentence, and as the reference to a new entity in the text or an old in the sentence. Concerning referential status in text, our results are broadly in line with Siddharthan et al. (2011), which shows that a new entity in the text is more likely to be referred to the full name, whereas only the surname is used for an old entity. Concerning referential status in the sentence, the fact that a proper name reference to an old entity in the text is more likely to be longer than one to a new entity was somewhat unexpected, since some referential theories argue that a reference to previously mentioned entities tend to be shorter (Chafe, 1994). A possible explanation could be the presence of cataphora, as in *Unlike **his** peers, **Harold Camping** does not pack a positive punch.*

As future work, we aim to develop a computational model for proper name generation based on the reported findings. Besides the variation between proper name forms in different parts of



(a) Syntactic position (b) Referential status

Figure 2: Average length of the proper names as a function of: (2a) syntactic position and (2b) referential status. Error bars represent 95% confidence intervals.

	Syntax		Text		Sentence		General
	Subject	Object	New	Old	New	Old	
First+Last	57.41%	38.74%	69.52%	36.53%	44.19%	57.16%	46.2%
Last	24.45%	37.17%	10.60%	44.26%	35.93%	26.61%	34.9%
First	6.15%	11.98%	4.33%	10.12%	8.58%	7.78%	8.5%
Middle+Last	3.39%	3.38%	4.62%	2.02%	2.91%	1.76%	2.8%
First+Middle+Last	2.92%	2.79%	4.72%	1.36%	2.44%	1.53%	2.3%
Middle	1.06%	1.88%	0.78%	1.74%	1.57%	0.80%	1.5%
Others	4.62%	4.06%	5.43%	3.97%	4.38%	4.36%	3.8%

Table 2: Percentage of the attribute sets in the proper name references

a text, this model should be able to address the proper name preferences for each entity. For instance, it should account that *Winston Churchill* is typically mentioned by his surname (*Churchill*), whereas *Napoleon Bonaparte* is by his first name (*Napoleon*). We will address this by training individual models combining the a priori probability of a particular proper name for a particular individual with contextual factors. Additionally, we plan to annotate the proper name references to all the entities present in the texts of our corpus, and not only the references to the 1,000 people studied here. We think this expansion will give a broader view of the generation of proper names, since we will be able to study the process as a function of other discourse conditions, as topicality.

Acknowledgments

This work has been supported by the National Council of Scientific and Technological Development from Brazil (CNPq).

References

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165. The Web of Data.

Wallace L. Chafe. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Individual variation in the choice of referential form. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California. Association for Computational Linguistics.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Comput. Linguist.*, 38(1):173–218, March.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge University Press, New York, NY, USA.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015.

Kees van Deemter. 2014. Referability. In Amanda Stent and Srinivas Bangalore, editors, *Natural Language Generation in Interactive Systems*, chapter 5, pages 101–103. Cambridge University Press, New York, NY, USA.