Individual Variation in the Choice of Referential Form

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

Abstract

This study aims to measure the variation between writers in their choices of referential form by collecting and analysing a new and publicly available corpus of referring expressions. The corpus is composed of referring expressions produced by different participants in identical situations. Results, measured in terms of normalized entropy, reveal substantial individual variation. We discuss the problems and prospects of this finding for automatic text generation applications.

1 Introduction

Automatic text generation is the process of automatically converting data into coherent text - practical applications range from weather reports (Goldberg et al., 1994) to neonatal intensive care reports (Portet et al., 2009). One important way to achieve coherence in texts is by generating appropriate referring expressions throughout the text (Krahmer and van Deemter, 2012). In this generation process, the choice of referential form is a crucial task (Reiter and Dale, 2000): when referring to a person or object in a text, should the system use a proper name ("Phillip Anschutz"), a definite description ("the American entrepreneur") or a pronoun ("he")?

Despite the large amount of algorithms developed for deciding upon the form of a referring expression (Callaway and Lester, 2002; Greenbacker and McCoy, 2009; Gupta and Bandopadhyay, 2009; Orăsan and Dornescu, 2009; Greenbacker et al., 2010), it is difficult to know how well these algorithms actually perform. Typically, such algorithms are evaluated against a corpus of human written texts, predicting what form each reference should have in a given context. Now consider a situation in which the algorithm predicts that a reference should be a description, while this same reference is a pronoun in the corpus text. Should this count as an error? The answer is: it depends. The use of a pronoun does not necessarily mean that the use of a description is incorrect. In fact, other writers might have used a description as well.

In general, corpora of referring expressions have only *one* gold standard referential form for each situation, while different writers may conceivably vary in the referential form they would use. This complicates the development and evaluation of text generation algorithms, since these will typically attempt to predict the corpus gold standard, which may not always be representative of the choices of different writers. Although recent work in text generation has explored individual variation in the content determination of definite descriptions (Viethen and Dale, 2010; Ferreira and Paraboni, 2014), to the best of our knowledge this has not been systematically explored for choosing referential forms.

In this paper, we collect and analyze a new corpus to address this issue. In the collection, we presented different writers with texts in which all references to the main topic of the text have been replaced with gaps. The task of the participants was to fill each of those gaps with a reference to the topic. In the analysis, we estimated to what extent different writers agree with each other in terms of normalized entropy. In addition, we study whether this variation depends on the text genre, compar-

Proceedings of NAACL-HLT 2016, pages 423–427,
San Diego, California, June 12-17, 2016.

ing encyclopedic texts with news and product reports. Moreover, we discuss the implications of our findings for automatic text generation, exploring whether factors such as syntactic structure, referential status and recency affect the variation between the writers' choices. The annotated corpus is made publicly available[1].

2 Data Gathering

2.1 Material

For our study, we used 36 English texts, equally distributed over three different genres: news texts, reviews of commercial products and encyclopedic texts. The encyclopedic texts were selected from the GREC corpus (Belz et al., 2010), which is a standard corpus for testing and evaluating models for choice of referential form. The news and review texts were selected from the AQUAINT-2 corpus[2] and the SFU Review corpus (Konstantinova et al., 2012), respectively.

Note that, depending on the genre, texts may address different kinds of topics. For instance, the news texts usually are about a person, a company or a group; the product reviews may be about a book, a movie or a phone; and the encyclopedic texts about a mountain, a river or a country. In all texts, all expressions referring to the topic were replaced with gaps, which the participants should fill in.

2.2 Participants

Participants were recruited through CrowdFlower[3]. 78 participants completed the survey. 53 were female and 25 were male. Their average age was 37 years old. Most were native speakers (73 participants) or fluent in English (5 participants).

2.3 Procedure

The participants were first presented with an introduction to the experiment, explaining the procedure and asking their consent. Next, they were asked for their age, demographic information and English language proficiency. After this, participants were randomly assigned to a list, containing 9 texts (3 per genre).

The task of the participants was to fill in each gap with a reference to the topic of the text. To inform the participants about the entities, a short description - extracted from the Wikipedia page about the topic - was provided before each text.

Participants were encouraged to fill in the gaps according to their preferences, so that they felt the texts would be easy to understand. We made sure that participants did not fill all the gaps in a text with only one referring expression (to avoid copy/paste behaviour). Participants could also not leave any gap empty (they were instructed to use the "-" symbol for empty references).

2.4 Annotation

The first author of this study annotated the referring expressions produced by participants for referential form, syntactic position, referential status, and recency. Coding was straightforward, and the few difficult cases were resolved in discussions between the co-authors.

The referring expressions were assigned to one of five forms: **proper names** ("*Philip Anschutz*, 66, will have no trouble keeping busy."); **pronouns** ("*It* is the highest peak [...]", "Huffman, *who* spoke at the sentencing phase [...]"); **definite descriptions** ("[...] *the Russian President* defended the country's contribution [...]"); **demonstratives** ("You'll probably have screaming kids who want to see *this movie*."); and **empty references** ("He rarely grants on-the-record media interviews and __ seldom allows himself to be photographed.").

Following the GREC Project scheme (Belz et al., 2010), referring expressions were annotated for three syntactic positions: subject noun phrases, object noun phrases, and genitive noun phrases that function as determiners (***Google's*** *stock*). Referential status refers to whether a referring expression is a first mention to the topic (new) or not (old). We annotated this at the level of the text, paragraph and sentence, so that a reference can be new in paragraph, but old in the text. Recency, finally, is the distance between a given referring expression and the last, previous reference to the same topic, measured in terms of number of words within a paragraph. If the referring expression was the first mention to the topic in the paragraph, its recency is set to 0.

In total, 10,977 referring expressions were col-

[1] http://ilk.uvt.nl/~tcastrof/vareg

[2] http://catalog.ldc.upenn.edu/LDC2008T25

[3] http://www.crowdflower.com/

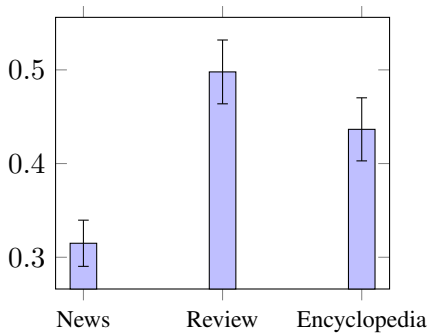


Figure 1: Average entropy per gap as a function of text genre. The error bars represent the 95% confidence intervals.

lected in 563 referential gaps. 3,682 were annotated as proper names, 4,662 as pronouns, 768 as definite descriptions, 318 as demonstratives and 158 as empty references. The remaining 1,389 were ruled out of the corpus, since they did not consist of a reference to the target entity or changed the meaning of the original sentence.

2.5 Analysis

We measured variation between participants' choices for each gap, using the normalized entropy measure, defined in Equation 1, where X corresponds to the references in a given gap, and $n = 5$ the number of referential forms annotated.

$$H(X) = -\sum_{i=1}^{n=5} \frac{p(x_i)\log(p(x_i))}{\log(n)} \quad (1)$$

The measure ranges from 0 to 1, where 0 indicates the complete agreement among the participants for a particular referential form, and 1 indicates the complete variation among their choices.

3 Results

Figure 1 presents the main result, depicting the amount of individual variation in referential forms, measured in terms of entropy, as a function of text genre. The averaged entropies are significantly higher than 0 for all three genres according to a Wilcoxon signed-rank test (News: $V = 20,910.0$, $p < .001$; Reviews: $V = 11,476.0$, $p < .001$; and Encyclopedic texts: $V = 10,153.0$, $p < .001$). This clearly shows that different writers can vary substantially in their choices for a referential form. Comparing the three different genres, we find that writers' choices of referential form varied most in review texts and least in news texts, with encyclopedic texts sandwiched in between (Kruskal-Wallis $H = 70.73$, $p < .001$).

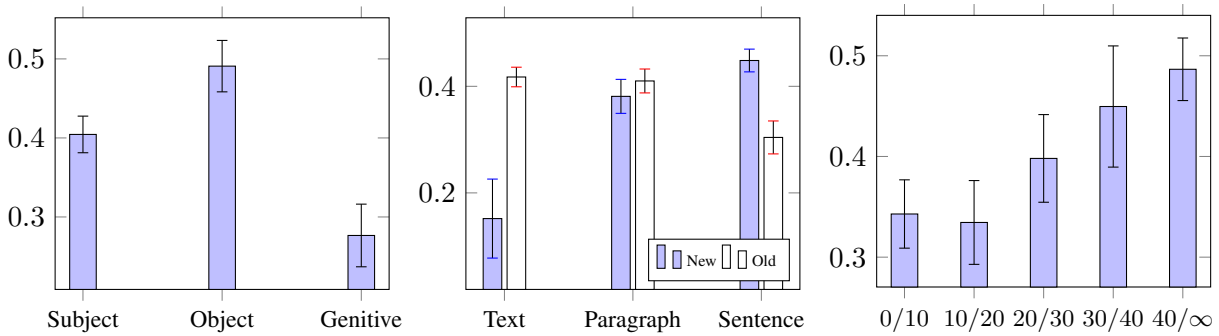
In comparison with the original texts, 44% of the referring expressions produced by the writers differ from the original ones in a same referential gap. Furthermore, the form of the original referring expressions differs from the major choice of the writers in 38% of the referential gaps.

To get a better understanding of factors potentially influencing individual variation, we investigate the effects of three linguistic factors: syntactic position, referential status and recency. Figure 2 depicts the average entropies for each of these.

Comparing the three syntactic positions, Figure 2a suggests that the highest variation is found when writers need to choose referential forms in the object position of a sentence, whereas the lowest variation is found for references that function as a genitive noun phrase determiner (Kruskal-Wallis $H = 52.53$, $p < .001$).

Figure 2b depicts individual variation in the choice of referential form for old and new references in the text, paragraph and sentence. The data suggests a higher amount of individual variation when writers need to refer to a topic already mentioned in the text rather than a first mention (Mann-Whitney $U = 3,916.0$, $p < .001$), presumably because for a topic which is new in the text, writers were more likely to agree to use proper names (91% of the choices). Looking at old and new references within paragraphs reveals no significant differences in individual variation (Mann-Whitney $U = 32,669.5$, $p < .094$). At the sentence level, finally, there is more individual variation for references to a new topic than for references to a previously mentioned one (Mann-Whitney $U = 21,873.0$, $p < .001$). When writers referred to a previously mentioned referent in the sentence, they tended to agree on the use of a pronoun (76% of the choices).

Figure 2c shows the individual variation in referential form as a function of recency. Except for the relatively nearby intervals (between 0 and 10 words, and between 11 and 20 words), the data suggests that when the distance between two consecutive references gets larger, the variation among writ-



(a) Syntactic position (b) Referential status (c) Recency

Figure 2: Average entropy per gap as a function of: (2a) syntactic position, (2b) referential status, (2c) recency. Error bars represent 95% confidence intervals. In Figure 2c, the bars represent the average entropies for the group of references where the most recent prior reference is 10 or less words away, between 11 and 20 words, between 21 and 30 words, between 31 and 40 words and more than 40 words away.

ers' choices increases (Kruskal-Wallis $H = 35.31$, $p < .001$).

4 Discussion

In this paper, we studied individual variation in the choice of referential form by collecting a new (and publicly available) dataset in which different participants (writers) were asked to refer to the same referent throughout a text. This was done for different genres (news, product review and encyclopedic texts) by measuring the variation between participants in terms of normalized entropy. If participants would all use the same referential form in the same gap, we would expect entropy values of 0 (no individual variation), but instead we found a clearly different pattern in all three text genres. Moreover, we also saw a considerably difference in form among the original referring expressions and the ones generated by the participants. This reveals that substantial individual variation between writers exists in terms of referential form.

To get a better understanding of which factors influence individual variation, we analysed to what extent three linguistic factors had an impact on the entropy scores: syntactic position, referential status and recency. We found a higher amount of individual variation when writers had to choose referential forms in the direct object position, referring to previously mentioned topics in the text and first mentioned ones in the sentence, and references that were relatively distant from the most recent antecedent reference to the same topic.

These findings can be related to theories of reference involving the salience of a referent (Gundel et al., 1993; Grosz et al., 1995, among others). Brennan (1995), for example, argued that references in the role of the subject of a sentence are more likely to be salient than references in the role of the object. Chafe (1994), to give a second example, pointed out that references to previously mentioned referents in the discourse and ones that are close to their antecedent are more likely to be salient than references to new referents or ones that are distant from their antecedents. Note, incidentally, that none of these earlier studies address the issue of individual variation in referential form.

Arguably, the amount of individual variation is even larger than the data reported here suggest. To illustrate this, consider, for instance, that different participants referred to *Phillip Frederick Anschutz* - the main topic of one of the texts used - as *Phillip Frederick Anschutz*, *Mr. Phillip Frederick Anschutz*, *Anschutz*, *Mr. Anschutz* and *Phillip Anschutz*. Even though these all have the same referential form (proper names), there is also a lot of variation *within* this category. Indeed, it would be interesting in future research to explore which factors account for this within-form variation.

The current findings are important for automatic text generation algorithms in two ways. First, they are beneficial for developers of text generation systems, since they allow for a better understanding of

the range of variation that is possible in referring expression generation. Second, they allow for a more principled evaluation of algorithms predicting referential form. In fact, the collected corpus paves the way for developing models which predict frequency distributions over referential forms, rather than merely predicting a single form in particular context (as current models do).

Acknowledgments

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

We would also like to thank the members of the JUGO group from TiCC by their insightful comments on the manuscript.

References

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2010. Empirical methods in natural language generation. chapter Generating Referring Expressions in Context: The GREC Task Evaluation Challenges, pages 294–327. Springer-Verlag, Berlin, Heidelberg.

Susan E. Brennan. 1995. Centering attention in discourse. *Language and Cognitive Processes*, 10(2):137–167.

Charles B. Callaway and James C. Lester. 2002. Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 88–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 and Ivandré Paraboni. 2014. Referring expression generation: Taking speakers preferences into account. In Petr Sojka, Aleš Horák, Ivan KopeČek, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 8655 of *Lecture Notes in Computer Science*, pages 539–546. Springer International Publishing.

Eli Goldberg, Norbert Driedger, and Richard I. Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert: Intelligent Systems and Their Applications*, 9(2):45–53, April.

Charles F Greenbacker and Kathleen F McCoy. 2009. Feature selection for reference generation as informed by psycholinguistic research. In *Proceedings of the CogSci 2009 Workshop on Production of Referring Expressions (PRE-Cogsci 2009)*.

Charles F. Greenbacker, Nicole L. Sparks, Kathleen F. McCoy, and Che-Yu Kuo. 2010. Udel: Refining a method of named entity generation. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG '10, pages 239–240, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Comput. Linguist.*, 21(2):203–225.

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

Samir Gupta and Sivaji Bandopadhyay. 2009. Junlg-msr: A machine learning approach of main subject reference selection with rule based improvement. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, UCNLG+Sum '09, pages 103–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Natalia Konstantinova, Sheila C. M. de Sousa, Noa P. Cruz Díaz, Manuel J. Maña López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 3190–3195.

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

Constantin Orăsan and Iustin Dornescu. 2009. Wlv: A confidence-based machine learning method for the grec-neg'09 task. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, UCNLG+Sum '09, pages 107–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franois Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(78):789 – 816.

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

Jette Viethen and Robert Dale. 2010. Speaker-dependent variation in content selection for referring expression generation. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 81–89, Melbourne, Australia, December.