

Forms of Anaphoric Reference to Organisational Named Entities: Hoping to widen appeal, they diversified

Christian Hardmeier¹ Luca Bevacqua² Sharid Loáiciga³ Hannah Rohde²

¹Department of Linguistics and Philology, Uppsala University

²Department of Linguistics and English Language, University of Edinburgh

³CLASP, University of Gothenburg

christian.hardmeier@lingfil.uu.se lbevacqu@ed.ac.uk

sharid.loaiciga@gu.se hannah.rohde@ed.ac.uk

Abstract

Proper names of organisations are a special case of collective nouns. Their meaning can be conceptualised as a collective unit or as a plurality of persons, permitting different morphological marking of anaphoric pronouns. This paper explores the variability of references to organisation names with 1) a corpus analysis and 2) two crowd-sourced story continuation experiments. The first shows the bias for singular vs. plural conceptualisation depends on the level of formality of a text. In the second, we observe a strong preference for plural *they* typical of informal speech. This preference is reduced for edited corpus data compared with constructed sentences.

1 Introduction

The names of organisations such as political bodies or companies are often made-up words (e. g., “Intel”, “Novartis”) or acronyms (e. g., “EU”, “Unesco”). They differ from other noun phrases in that they offer very little information about their grammatical properties such as number or, in languages where this is relevant, gender. Such names are a special case of the broader category of collective nouns, which also includes common nouns such as “team” or “committee”, and they can be conceptualised in different ways by focusing on the collective as a singular unit or on the plurality of people which the organisation is comprised of. When they occur as antecedents of referring expressions, names of organisations are a challenge for natural language processing (NLP) because they can trigger different types of morphological marking on the anaphoric elements. Moreover, the preference for certain types of agreement varies across different genres and, we expect, different languages. The

experiments presented here address English only and serve as a pilot study for an investigation of reference to organisations across multiple languages.

Via a corpus analysis of the OntoNotes corpus (Pradhan et al., 2013) and two crowd-sourced story continuation experiments, we study how organisational named entities are referenced after their introduction in a discourse. Specifically, we consider anaphoric expressions coreferent with the proper name of an organisation that are separated from their antecedent by a sentence boundary, but no intervening mentions belonging to the same coreference chain. The expressions are categorised into four classes: repetition of the proper name (*name*), paraphrastic noun phrases with a common noun such as “the company” (*noun*), and forms of the pronouns *it* and *they*. The pronominal case is informative to speakers’ choice between a conceptualisation as singular (*it*) or plural (*they*).

2 Related literature

Morphological agreement with collective nouns has received some attention in English linguistics, but most research focuses on the agreement of verbs rather than pronouns, and – to an even larger extent – on collective common nouns such as “team”, which are formally singular but can trigger plural agreement, rather than proper names.

There is broad agreement that American English prefers singular verb agreement with collective nouns, whereas notional concord with plural forms is not uncommon in British English (Fries, 1988; Bock et al., 2006; Hundt, 2009). Other varieties of English range in between (Hundt, 2006). Shift towards singular agreement is considered to be an ongoing diachronic process (Hundt, 2009), but the extent to which plural verb agreement with collectives is disappearing among younger speakers of British English is disputed (Fries, 1988).

| | it | they | name | noun | other | total |
|----|-----|------|------|------|-------|-------|
| bc | 8 | 15 | 59 | 10 | 13 | 105 |
| bn | 11 | 12 | 146 | 44 | 12 | 225 |
| mz | 17 | 11 | 91 | 24 | 4 | 147 |
| nw | 76 | 11 | 926 | 193 | 36 | 1242 |
| tc | 2 | 3 | 7 | 0 | 0 | 12 |
| wb | 6 | 4 | 52 | 8 | 4 | 74 |
| | 120 | 56 | 1281 | 279 | 69 | 1805 |

Table 1: Reference types per genre in OntoNotes

The situation for pronouns is different. Pronouns following collective nouns are more likely to receive plural marking than verbs (Hundt, 2006, 2009), and there is psycholinguistic evidence of processing differences favouring syntactic (singular) agreement for verbs and notional (plural) concord for pronouns (Bock et al., 2006). In corpus data, there are also instances where singular and plural agreement co-occur with the same mention (“mixed concord”), typically involving a singular verb and a plural pronoun (Hundt, 2009).

3 Corpus analysis

3.1 Corpus and extraction

The OntoNotes corpus (Pradhan et al., 2013) contains about 1.7 million words of annotated English text predominantly of American origin from different genres, or data sources: *newswire* (nw), *broadcast news* (bn), *broadcast conversation* (bc), *magazine* (mz), *telephone conversation* (tc), *web data* (wb) and *pivot text* (pt).¹ We extract examples using the gold-standard annotations of coreference and named entity type. Each example is a pair of mentions belonging to the same coreference chain. To ensure that the corpus analysis is comparable with the continuation studies described in Section 5, we only extract pairs of mentions in adjacent sentences, excluding both pairs of mentions in the same sentence and pairs with intervening sentences. A pair of mentions is extracted if the two mentions are neighbouring members of the same coreference chain (i. e., no mentions of the same chain occur in between) and the first mention is annotated as a named entity of type ORG.

3.2 Overview

Table 1 and the first six bars of Figure 1 show the distribution of reference types for the different

¹The pt subcorpus contains excerpts of the Bible and is not used in this paper.

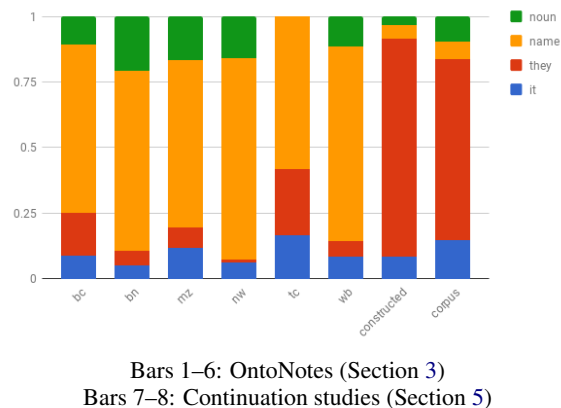


Figure 1: Proportions of reference types

OntoNotes genres. The size of the individual subcorpora varies substantially and so does the number of examples that can be extracted from each. The smallest non-empty sample ($N = 12$) is from the *telephone conversations* (tc) subcorpus, the largest ($N = 1242$) is from *Newswire*.

The most common type of reference, making up 58–75% of the examples in all subcorpora, is a repetition of the name. Paraphrasing noun phrases are more common in *broadcast news* (19.6%), *magazine* (16.3%) and *newswire* (15.5%) than in *web data* (10.8%) and *broadcast conversation* (9.5%). Many examples in the *other* category are instances of the first-person pronoun *we* that occur when a representative of the organisation is quoted or speaking. The relative frequency of pronominal references (*it* and *they*) varies considerably between genres. It is greatest in *telephone conversations*, where 5 out of 12 references are of this type. In *newswire* (7%) and *broadcast news* (10.2%), pronominal references are much less common. *Web data* (13.5%), *magazine* (19.0%) and *broadcast conversation* (21.9%) are in between. Among the pronominal references, we observe large differences in the preference for *it* vs. *they* across subcorpora, with numbers ranging from 34.8% *it* in *broadcast conversation* to 87.4% in *newswire*.

4 The effect of formality

In this section, we examine the hypothesis that the cross-corpus variation in the conceptualisation of organisations as singular or plural can be explained by the different levels of *formality* of the texts.

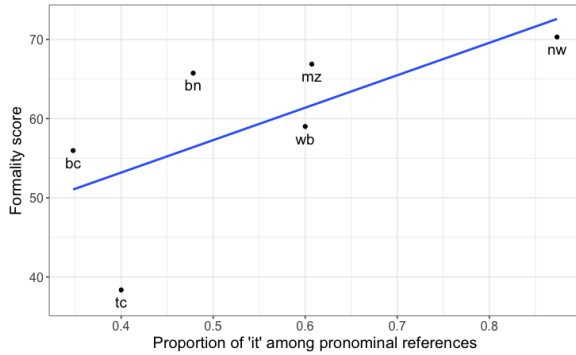


Figure 2: Formality score and prevalence of *it* in OntoNotes subcorpora

4.1 Measuring formality

To measure the formality of discourse, we use an automatic metric proposed by Heylighen and Dewaele (2002). The metric is called *F-score* by the original authors, but we use *Formality score* to avoid confusion with the entirely unrelated F-score derived from precision and recall. The fundamental assumption of Heylighen and Dewaele (2002) is an opposition between *formality* and *contextuality*, with the claim that more formal texts prefer more absolute and less context-dependent forms of expression, which is reflected in lexical choice. The authors identify two (non-exhaustive) subsets of the lexicon that they call *formal* or *non-deictic* and *deictic*, respectively. This distinction is then operationalised via part-of-speech (POS) categories with nouns, adjectives, prepositions and articles taken to be non-deictic, and pronouns, verbs, adverbs and interjections as deictic. The score is calculated as:

$$F = 100 \cdot \frac{N_{\text{formal}} - N_{\text{deictic}}}{2N} + 50 \quad (1)$$

where N_{formal} and N_{deictic} are the counts of formal and deictic tokens and N is the total corpus size.

4.2 Choice of referring expression

Since the number of pronouns in a corpus enters the computation of the Formality score through the N_{deictic} term, we must exercise care when we measure referential preferences so that we do not use a metric that is correlated by construction with the Formality score. The preference among pronominal references between the conceptualisation of organisation as singular versus collective entities can be measured as the proportion of references with *it* among third-person pronominal references, i. e., $N_{it}/(N_{it} + N_{they})$. As both *it* and *they* are counted as

pronouns in the Formality score, their proportion can be measured independently from the score.

Figure 2 plots the proportion of *it* among pronominal references (x-axis) against the Formality score (y-axis). The ranking predicted by the Formality score seems intuitively reasonable: The *newswire*, *magazine* (two edited written genres) and *broadcast news* (prepared speech in a very formal setting) genres are identified as most formal, whereas *telephone conversations* are predicted to be least formal. The Pearson correlation coefficient between the Formality score and the proportion of *it* vs. *they* is 0.67, which fails to reach significance ($p = 0.146$). However, the Spearman rank correlation reaches a value of $\rho = 0.886$, which is significantly different from 0 ($p < 0.05$). This confirms that formality is a relevant factor to explain the language-internal variation in the number marking of pronouns with organisational antecedents, as was suggested for collective nouns more generally by Hundt (2009).

5 Continuation experiments

Two story-continuation studies presented participants with prompts to elicit entity coreference.

5.1 Study 1: Constructed stimuli

Materials The 16 experimental items consisted of a context sentence and a prompt. The first sentence introduced a named entity in the subject position of the matrix clause with some additional contextual information, followed by a prompt with a discourse adverbial or other connective (e.g., *In the following years*, *Because of this*). The named entities belonged to four categories: names of companies, publishers, sport teams and music bands.

The experimental items were interleaved with 48 filler items. They included 20 fillers composed of a sentence introducing two animate or inanimate entities, followed by an adverbial prompt, 24 items for an unrelated production experiment involving the coreference of the pronouns *it* and *this*, and 4 catch trials with a straightforward correct response, which were mentioned in the instructions.

Participants Twenty-seven monolingual American English speakers aged 19–63 (mean age 36, $\sigma = 11.2$; 15 male) were recruited from Amazon’s Mechanical Turk (Munro et al., 2010; Gibson et al., 2011). All had US IP addresses and received \$4 for an estimated 30-minute task.

| | it | they | name | noun | total |
|---------|----|------|------|------|-------|
| Study 1 | 32 | 307 | 19 | 12 | 370 |
| Study 2 | 24 | 113 | 11 | 16 | 164 |
| | 56 | 420 | 30 | 28 | 534 |

Table 2: Reference types in the continuations

Procedure The continuations were collected via a web-based interface that participants accessed directly from Amazon’s Mechanical Turk. The website displayed a background questionnaire, a consent form and an instructions page, and then proceeded to display one item at a time with a text box for participants to write their continuations.

Analysis The continuations were annotated for referent type, using the same labels as in the corpus analysis, plus *other* for continuations in which the named entity was not mentioned in any way. One of the authors of the paper annotated the whole set of continuations, and two others labelled half of it. The annotations did not present any real case of disagreement among the authors.

Results 50 out of 420 continuations were excluded because they were labelled as *other*. This left 370 labelled annotation for the analysis. The results are shown in Table 2 and Figure 1. It is striking that the participants produced an extremely high number of pronominal continuations, most of them with *they* (accounting for 83% of the referential types vs. only 3.1% in the OntoNotes data). By contrast, the *name* category occurred only infrequently (5.1% of types vs. 71% in OntoNotes).

5.2 Study 2: Corpus stimuli

Materials The 24 target passages were extracted from the data described in Section 3.1. They were interleaved with 76 filler items. 24 of these were extracted from the ParCorFull corpus (Lapshinova-Koltunski et al., 2018). These fillers mentioned a range of referents; the sentence continuation prompt was an adverbial expression (e.g., *Eventually*). 48 additional fillers were items of the aforementioned unrelated production experiment, and a final 4 fillers repeated the catch trials from Study 1.

Participants Nineteen monolingual English-speaking participants aged 23–44 (mean age 30, $\sigma = 6.5$; 13 male) were recruited as in Study 1, and received \$7 for an estimated 50-minute task.

Procedure and Analysis Identical to Study 1.

Results 43 out of 207 continuations were labelled *other* and excluded from the data set. This left 164 continuations labelled for referential type. The results are shown in Table 2 and Figure 1. While the continuations produced in this study still contain a much larger number of pronominal references than the OntoNotes examples, the proportion of pronouns (66.2%) is considerably lower than in Study 1 (80.7%), and the proportion of *it* among pronominal references is higher (17.5% vs. 9.4%). The difference between the distributions observed in Studies 1 and 2 is statistically significant ($\chi^2 = 145.71$; $p < 0.001$) in a χ^2 test with Monte Carlo simulation (Hope, 1968).

6 Conclusions

Focusing on pronouns referring to proper names, our study confirms a number of results suggested by earlier research concentrating primarily on collective common nouns and verb agreement (Hundt, 2009). There is significant language-internal variation in English in how speakers and writers refer to organisational named entities. In the OntoNotes corpus data, the most frequent way of referring to an organisation is by repeating its name. The number of pronominal references and their distribution among *it* and *they* varies greatly across genres. As suggested by Hundt (2009), we find a correlation between the level of formality of a text and the prevalence of singular pronominal references.

In the story continuation studies, we observe a distribution of reference types that is more extreme in its preference for *they* than even the most informal OntoNotes genres. This suggests that the patterns we obtain in this type of study are more representative of informal and spoken language than of more formal written genres, despite the written modality of the task. As a result, we cannot automatically generalise the findings from these studies across different genres. However, the combination of crowd-sourced continuation study and corpus analysis provides us with a useful baseline in terms of both methods and results for a planned cross-lingual study of reference to named entities.

Acknowledgments

Christian Hardmeier was supported by the Swedish Research Council under grant 2017-930, *Neural Pronoun Models for Machine Translation*. Hannah Rohde was supported via a Leverhulme Trust Prize in Languages & Literatures.

References

- Kathryn Bock, Sally Butterfield, Anne Cutler, J. Cooper Cutting, Kathleen M. Eberhard, and Karin R. Humphreys. 2006. Number agreement in British and American English: Disagreeing to agree collectively. *Language*, 82(1):64–113.
- Udo Fries. 1988. The crew have abandoned the ship: Concord with collective nouns revisited. *AAA: Arbeiten aus Anglistik und Amerikanistik*, 13(2):99–104.
- Edward Gibson, Steve Piantadosi, and Kristina Fedorenko. 2011. Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, 5(8):509–524.
- Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7:293–340.
- Adey C. A. Hope. 1968. A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(3):582–598.
- Marianne Hundt. 2006. The committee has/have decided... On concord patterns with collective nouns in inner- and outer-circle varieties of English. *Journal of English Linguistics*, 34(3):206–232.
- Marianne Hundt. 2009. Concord with collective nouns in Australian and New Zealand English. In Pam Peters et al., editors, *Comparative studies in Australian and New Zealand English: Grammar and beyond*, pages 207–224. John Benjamins.
- Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of 11th Language Resources and Evaluation Conference*, pages 00–00, Miyazaki, Japan. European Language Resources Association (ELRA). To appear.
- R. Munro, S. Bethard, V. Kuperman, V.T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. [Towards robust linguistic analysis using ontonotes](#). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.