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

Christian Hardmeier\textsuperscript{1}  Luca Bevacqua\textsuperscript{2}  
Sharid Loáiciga\textsuperscript{3}  Hannah Rohde\textsuperscript{2}  

\textit{presented by Joakim Nivre}\textsuperscript{1}

\textsuperscript{1}Department of Linguistics and Philology, Uppsala University  
\textsuperscript{2}Department of Linguistics and English Language, University of Edinburgh  
\textsuperscript{3}CLASP, University of Gothenburg
Organisational Named Entities

- Names of organisations: Companies, political bodies, sport teams, music bands, etc.
- Often made-up words (*Intel, Novartis*) or acronyms (*EU, Unesco*)
- Little information about number or gender
- Different conceptualisation
  - Singular: collective as a unit
  - Plural: individuals within organisation
Names of Organisations as Collective Nouns

- Special case of *collective nouns* such as *team*, *family*, etc.
- Studied in English linguistics, especially for verb agreement
- Can be used with singulaires (*syntactic agreement*) or plurals (*notional concord*) in English
- American English: often singular verbs but plural pronouns
- Singular and plural agreement can co-occur (*mixed concord*)
Our Study

**Research question:** What forms are possible and preferred when re-mentioning named entities?

- Current study on English – multilingual extension planned
- Two types of experiments:
  - Corpus study on OntoNotes
  - Story continuation experiments on Mechanical Turk
We consider four types of references to organisations:

name    noun    it    they

**Name: Repetition of the proper name**

Since the introduction of the first MacBook, Apple grew bigger and bigger.

Last year, Apple sold the most MacBooks in its history.
We consider four types of references to organisations:

- name
- noun
- it
- they

**Noun: Paraphrastic noun phrases**

AC/DC achieved international success in 1976.

In the next forty years, the band continued to attract more loyal fans.
We consider four types of references to organisations:

name       noun       it       they

It: Pronoun with singular conceptualisation

Since the introduction of the first MacBook, Apple grew bigger and bigger.

Last year, it had record sales.
Four Types of References

We consider four types of references to organisations:

name  noun  it  they

They: Pronoun with plural conceptualisation

Google entered the search machine business in 1998.

Ten years later, they were still in business.
Example Extraction

- OntoNotes: ~1.7 million words of American English text
- Gold-standard coreference and named entity annotations
- Subcorpora:
  - **bc** broadcast conversation
  - **bn** broadcast news
  - **mz** magazines
  - **nw** newswire
  - **tc** telephone conversations
  - **wb** web data
- Each example:
  - a pair of mentions belonging to the same coreference chain
  - occurring in adjacent sentences
  - with no intervening mentions from the same chain
Reference Types per Genre

bc
bn
mz
nw
tc
wb

green: noun
orange: name
red: they
blue: it
## Reference types per genre

<table>
<thead>
<tr>
<th></th>
<th>it</th>
<th>they</th>
<th>name</th>
<th>noun</th>
<th>other</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>bc</td>
<td>8</td>
<td>15</td>
<td>59</td>
<td>10</td>
<td>13</td>
<td>105</td>
</tr>
<tr>
<td>bn</td>
<td>11</td>
<td>12</td>
<td>146</td>
<td>44</td>
<td>12</td>
<td>225</td>
</tr>
<tr>
<td>mz</td>
<td>17</td>
<td>11</td>
<td>91</td>
<td>24</td>
<td>4</td>
<td>147</td>
</tr>
<tr>
<td>nw</td>
<td>76</td>
<td>11</td>
<td>926</td>
<td>193</td>
<td>36</td>
<td>1242</td>
</tr>
<tr>
<td>tc</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>wb</td>
<td>6</td>
<td>4</td>
<td>52</td>
<td>8</td>
<td>4</td>
<td>74</td>
</tr>
</tbody>
</table>

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>120</td>
<td>56</td>
<td>1281</td>
<td>279</td>
<td>69</td>
<td>1805</td>
</tr>
</tbody>
</table>
Formality and Use of *it*

- **Hypothesis:** *Singular conceptualisation is more likely in more formal text genres.*
- Suggested for general collective nouns (Hundt, 2009)
- Measure: proportion of *it* among pronominal references:

\[
\frac{N(\text{it})}{N(\text{it}) + N(\text{they})}
\]
Measuring Formality

- Metric of text formality (Heylighen & Dewaele, 2002)
- Assumption: Formality is reflected in the use of certain parts of speech.
- *Formal* vocabulary: nouns, adjectives, prepositions, articles
- *Deictic* vocabulary: pronouns, verbs, adverbs, interjections
- Score calculation:

\[ F = 100 \cdot \frac{N_{\text{formal}} - N_{\text{deictic}}}{2N} + 50 \]
Formality and Use of *it*

![Graph showing the relationship between formality score and proportion of 'it' among pronominal references. The x-axis represents the proportion of 'it' among pronominal references, ranging from 0.4 to 0.8. The y-axis represents formality score, ranging from 40 to 70. The graph shows a positive correlation with data points labeled bc, bn, mz, wb, and nw.]
Correlations between formality and singular conceptualisation confirmed in OntoNotes.

- Rank correlation is significant ($\rho = 0.886; p < 0.05$).
- Linear correlation is not ($r = 0.67; p = 0.146$).

Modality also seems to play a role:
Strongest preference for *they* in the spoken subcorpora.
Continuation Experiments

- Two crowdsourcing experiments on Amazon Mechanical Turk
- Participants saw 16 target items + 48 fillers
- Each item was a pair of sentences:
  - Sentence #1: introduced a named entity in subject position
  - Sentence #2: adverbial prompt to elicit a reference to the named entity
- Instructions: complete sentence #2
Two Studies

Study 1: Constructed stimuli

- 27 mturk participants (restricted to US IP addresses)
- Prompt sentences constructed by the authors
- Four types of named entities: Companies, publishers, sport teams and music bands

Last week, Intel announced the shutdown of the factory. In the press release, ____________________________
Two Studies

Study 2: Corpus stimuli

- 19 mturk participants (same US IP address restriction)
- Prompt sentences extracted from OntoNotes and simplified
- Continuations constructed to increase chances of eliciting a reference to the named entity
- Generally longer and more complex than Study 1 stimuli
- Unrelated filler items likewise from corpus data

To distinguish itself, CNN is also expanding international coverage and adding a second global-news program. At the annual press conference, ____________________________
Continuation Studies: Results

<table>
<thead>
<tr>
<th></th>
<th>constructed</th>
<th>corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>it</td>
<td>32</td>
<td>24</td>
</tr>
<tr>
<td>they</td>
<td>307</td>
<td>113</td>
</tr>
<tr>
<td>name</td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>noun</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>total</td>
<td>370</td>
<td>164</td>
</tr>
</tbody>
</table>
All Results
Conclusions

- Very high proportion of *they* in continuation study.
- More varied responses with corpus stimuli, but *they* is still dominant.
- In OntoNotes, *they* use is negatively correlated with formality.
- Results of continuation study are more representative of informal and spoken language, even though the task was done in writing.
- Results will be used as a baseline in a multilingual experiment on English, German, French, Italian and Spanish.
Questions

Further questions can be addressed to:
- Christian Hardmeier: christian.hardmeier@lingfil.uu.se
- Luca Bevacqua: lbevacqu@ed.ac.uk
- Sharid Loáiciga: sharid.loaiciga@gu.se
- Hannah Rohde: hannah.rohde@ed.ac.uk