SPARCLING: Large-scale Annotation and Alignment of Parallel Corpora for the Investigation of Linguistic Variation

M. Hundt, M. Volk, E. Callegaro, J. Graën

University of Zurich, English Department & Institute of Computational Linguistics

1. Project goals

This collaboration between the departments of English and Computational Linguistics builds on the use of parallel multilingual corpora, which are useful in many fields, such as word sense disambiguation, machine translation, and contrastive (corpus) linguistics.

Computational Linguistics

- Annotation and alignment (both sentence and word level) of large parallel corpora from *Europarl* (Koehn 2005).
- Powerful and highly innovative query language, able to handle these corpora, and to access and view linguistic data in a user-friendly interface.

English Linguistics

- Prove the usefulness of such large annotated and word-aligned corpora for the investigation of linguistic variation.
- Data-driven approach for the analysis of variable article use in English. Methodology: Contexts where one language uses a structure (e.g. an article) are used to retrieve 'zero' contexts (i.e. bare NPs) in other language(s).

2. Collective nouns: contrastive analysis of article use in English and German

- In English, the use of articles with collectives is unpredictable (Poutsma 1904; Christophersen 1939).
- In German, there is no clear rule concerning the use of articles with collective nouns (Dudenredaktion 2005).
- Collective nouns investigated in the case study: *Parliament*, *Council*, *Committee*, and *people*.
- Sample: 1.437 parallel sentences (729 sentences in English originals and 708 sentences of their equivalent translations in German).

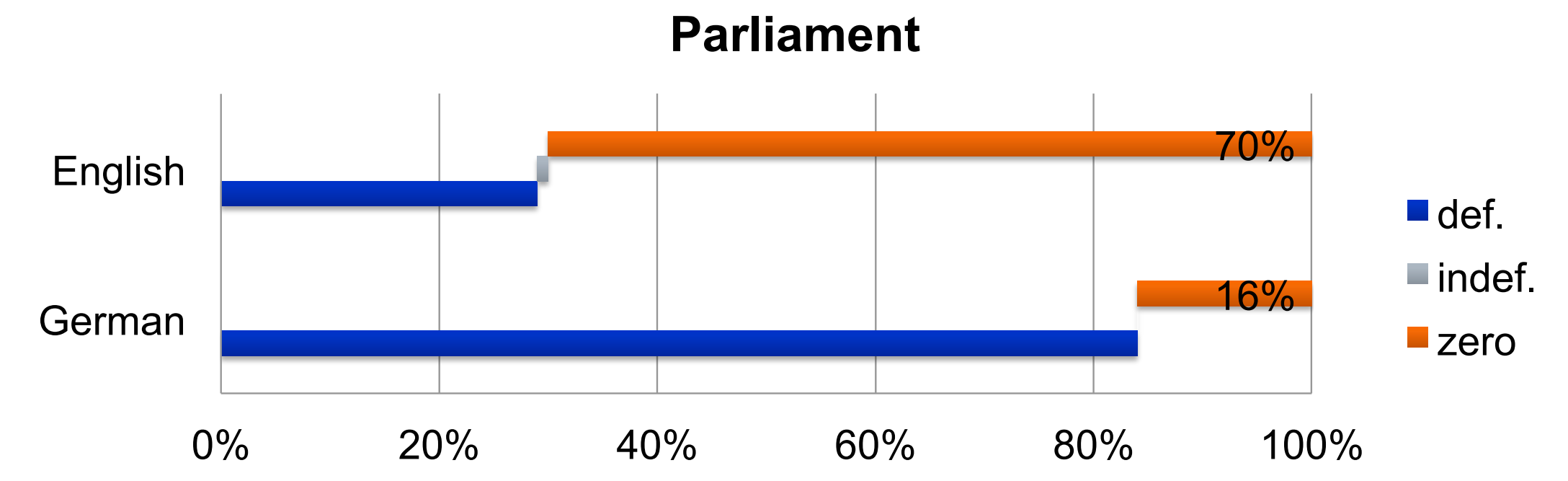


Fig.1: Article distribution for the word *Parliament* in English and German.

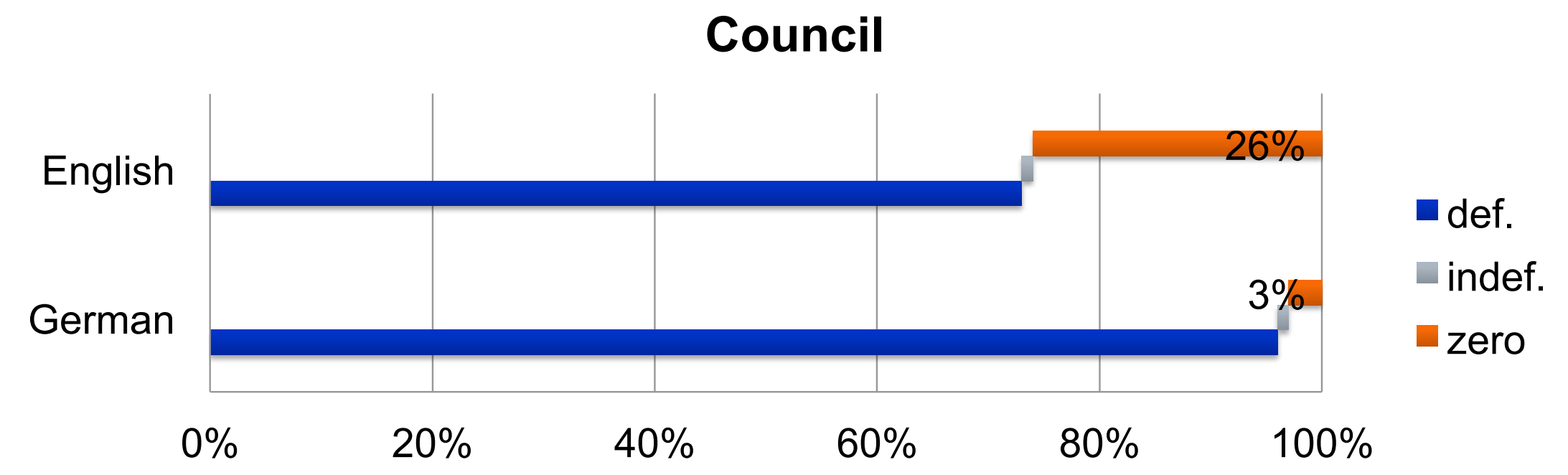


Fig.2: Article distribution for the word *Council* in English and German.

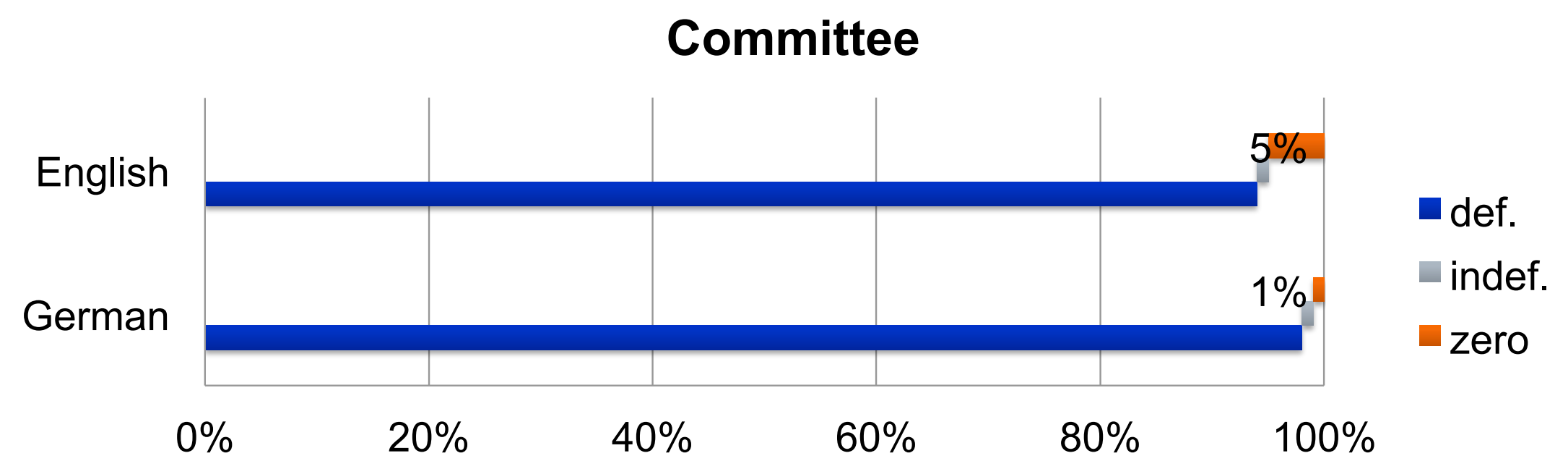


Fig.3: Article distribution for the word *Committee* in English and German.

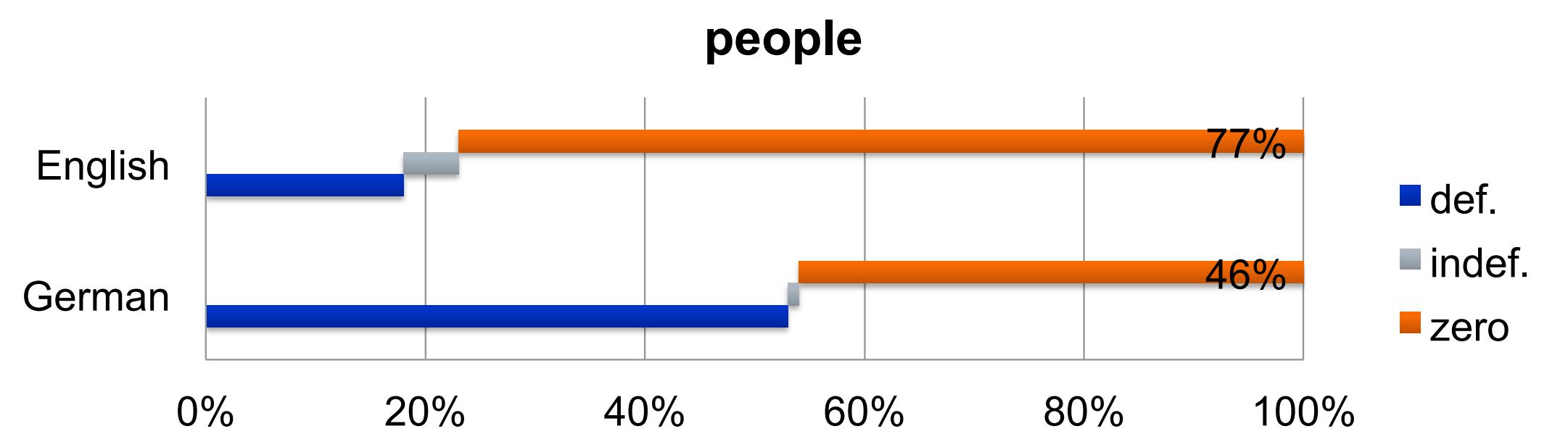


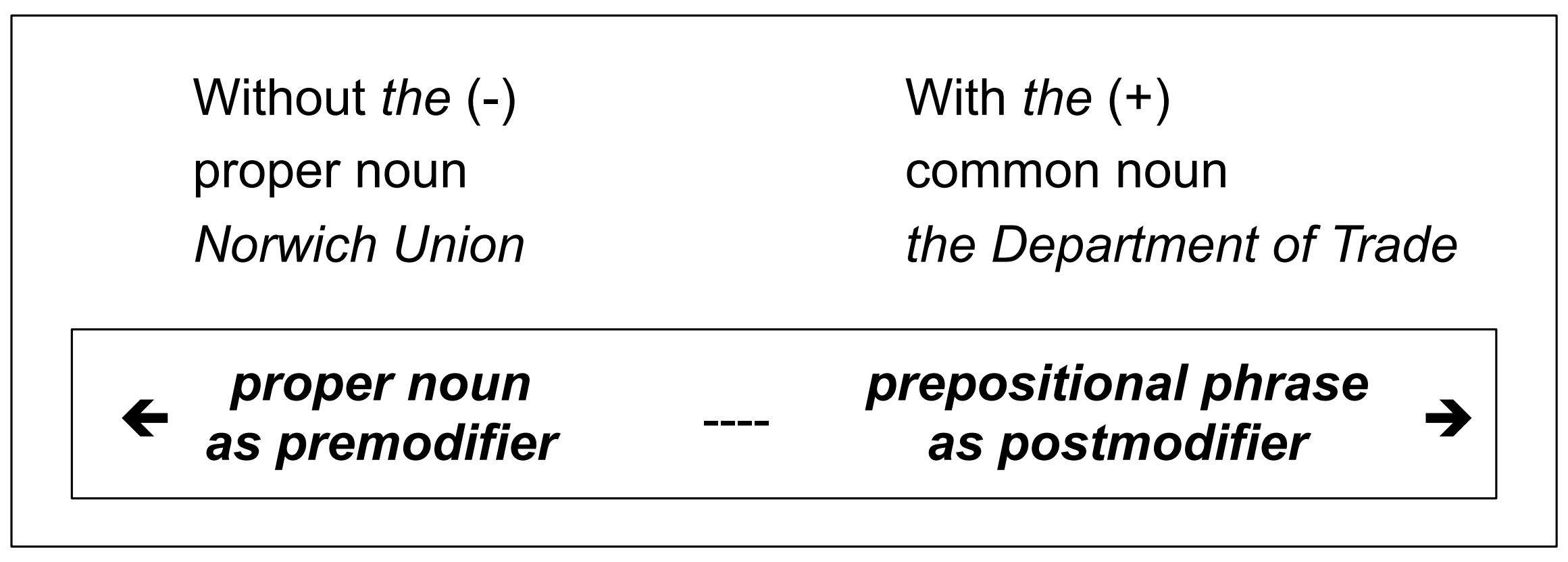
Fig.4: Article distribution for the word *people* in English and German.

Conclusion I

- In this context, German uses articles more frequently, whereas in English articles are more variable.
- Article use seems to be influenced by the meaning of the collective noun itself and not by the entire category of collectives (see *Parliament*).

3. Multi-word organization names: contrastive study of article use in English and German

- Tse (2003) investigates the grammatical factors that influence the presence and omission of the definite article in front of multi-word organization names in British English newspapers. Article use seems to vary according to different kinds of modification (**Fig.5**).
- The question is whether our data, which come from parliamentary debates, show a similar tendency for English.



Without *the* (-)
proper noun
Norwich Union

With *the* (+)
common noun
the Department of Trade

← ***proper noun as premodifier*** ---- ***prepositional phrase as postmodifier*** →

Fig.5: Proper-noun-common-noun scale (based on Tse, 2003: 299).

English		German	
def. art.	*bare NP*	*def. art.*	*bare NP*
97.5% (*N* 1662)	**2.5 %** (*N* 42)	**99.2%** (*N* 1537)	**0.8%** (*N* 12)

Fig.6: Article distribution with multi-word organization names in English and German.

English		German	
With *the* (+) ↑			
factor	*p value*	*factor*	*p value*
AJ	0.9813	***NP***	0.9935
GN	0.9823	**AJ**	0.9943
PN	0.9826	**PP**	0.0015 ••
SP	0.9840	**PG**	0.0556 .
CN	0.9841	**NPH**	0.2338
NP	0.9842	**SP**	0.9999
PO	0.1426	**PN**	0.9999
NA	0.9997	**CN**	0.9999
PP	0.0744 .	**AC**	0.9999
NCH	0.0039 ••	**PO**	0.9956
CNH	0.9793	***CNH***	0.9925
Without *the* (-) ↓			

Fig.7: Proper-noun-common-noun scale (based on SPARCLING data)

Conclusion II

- Our results differ slightly from Tse's (2003).
- Pronounced tendency for English multi-word organization names to occur with the definite article. In German, this preference is even stronger than in English.
- Political discourse might be more specific and restricted than newspaper language, which seems to be more advanced concerning variable article use.

Contacts

Prof. Dr. Marianne Hundt
English Department
m.hundt@es.uzh.ch

Prof. Dr. Martin Volk
Institute of Computational Linguistics
volk@cl.uzh.ch

Elena Callegaro
English Department
elena.callegaro@es.uzh.ch

Johannes Graën
Institute of Computational Linguistics
graen@cl.uzh.ch

References

1. Christophersen, P. 1939. *The articles: a study of their theory and use in English*. Copenhagen: E. Munksgaard.
2. Dudenredaktion (ed.). 2005. *Duden – die Grammatik: unentbehrlich für richtiges Deutsch*, 7th ed., vol. 4. Mannheim: Dudenverlag.
3. Koehn, P. (2005), *Europarl: A parallel corpus for statistical machine translation*. In *MT summit*, Vol. 5, pp. 79-86.
4. Poutsma, H. 1904. *A grammar of late modern English, for the use of continental, especially Dutch students*. Part II The parts of speech. P. Noordhoff.
5. Tse, Grace Y. W. 2003. 'Validating the Logistic Model of Article Usage Preceding Multi-Word Organization Names with the Aid of Computer Corpora.' *Literary and Linguistic Computing* 18(3): 287-313.

FNSNF
Schweizerischer Nationalfonds
zur Förderung der wissenschaftlichen Forschung