Premodifying –ing participles in the parsed BNC

Turo Vartiainen[1] & Jefrey Lijffijt[2]

[1]Department of English,
VARIENG, University of Helsinki, Finland
[2]Department of Information and Computer Science,
ALGODAN, Aalto University, Finland

DAMMOC
DATA MINING TOOLS
for changing modalities of communication

Premodifying –ing participles

- Participles of the type:
  - *An amusing story*
  - *The running men*
- A theoretically debated class.
  - Verbs or adjectives? Both?
  - How to annotate the participles?

Other –ing forms

- An additional challenge: there are other kinds of (nominal) premodifying –ing forms, as in:
  - *A parking attendant* ‘traffic warden’
  - *An eating contest* ‘a contest in eating’
- Compare:
  - *The parking man* ‘a man who is parking’
  - *The eating man* ‘a man who is eating’

Premodifying –ing participles

- Furthermore, the premodifying –ing participle is a very infrequent item in English.
  - Large datasets need to be analysed.
  - The British National Corpus
- Dependency information required for accurate and efficient retrieval of the –ing participles.
  - The parsed BNC

The parsed BNC

- Parsed with RASP (Briscoe et al. 2006; Andersen et al. 2008)
- Based on BNC-XML
  - Does not modify corpus, just adds information
    - Word level: new POS tags, lemmatization
    - Phrase & sentence level: grammatical relations
- Grammatical relations
  - Relation (head word, dependent word)
  - *An amusing story* → *ncmod* (story, amusing)

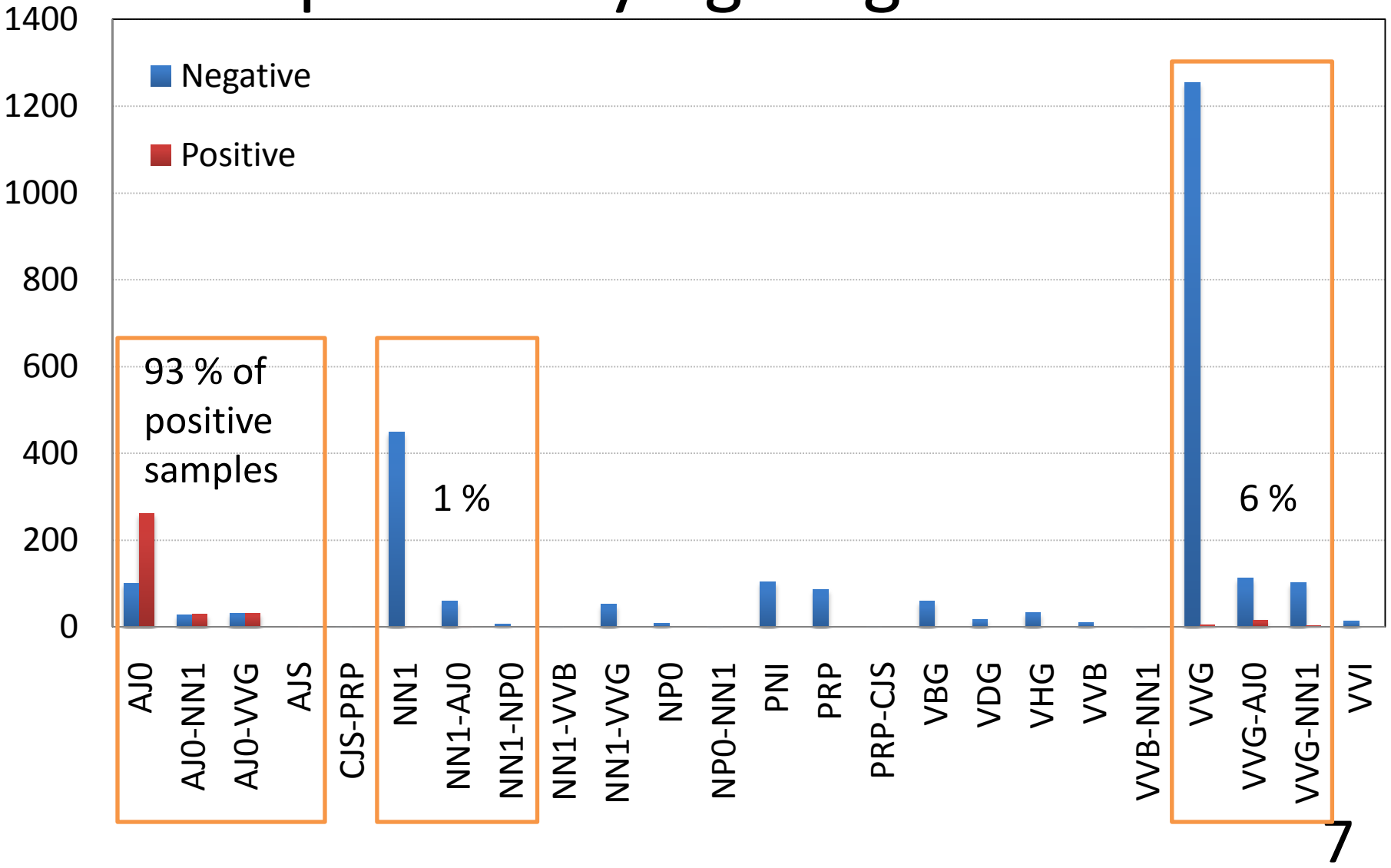
Mining premodifying –ing participles

- Constructed training set with ground truth
  - Three randomly selected texts
  - Approx. 3000 –ing forms
  - 351 premodifying –ing participles
  - 12 ambiguous cases discarded

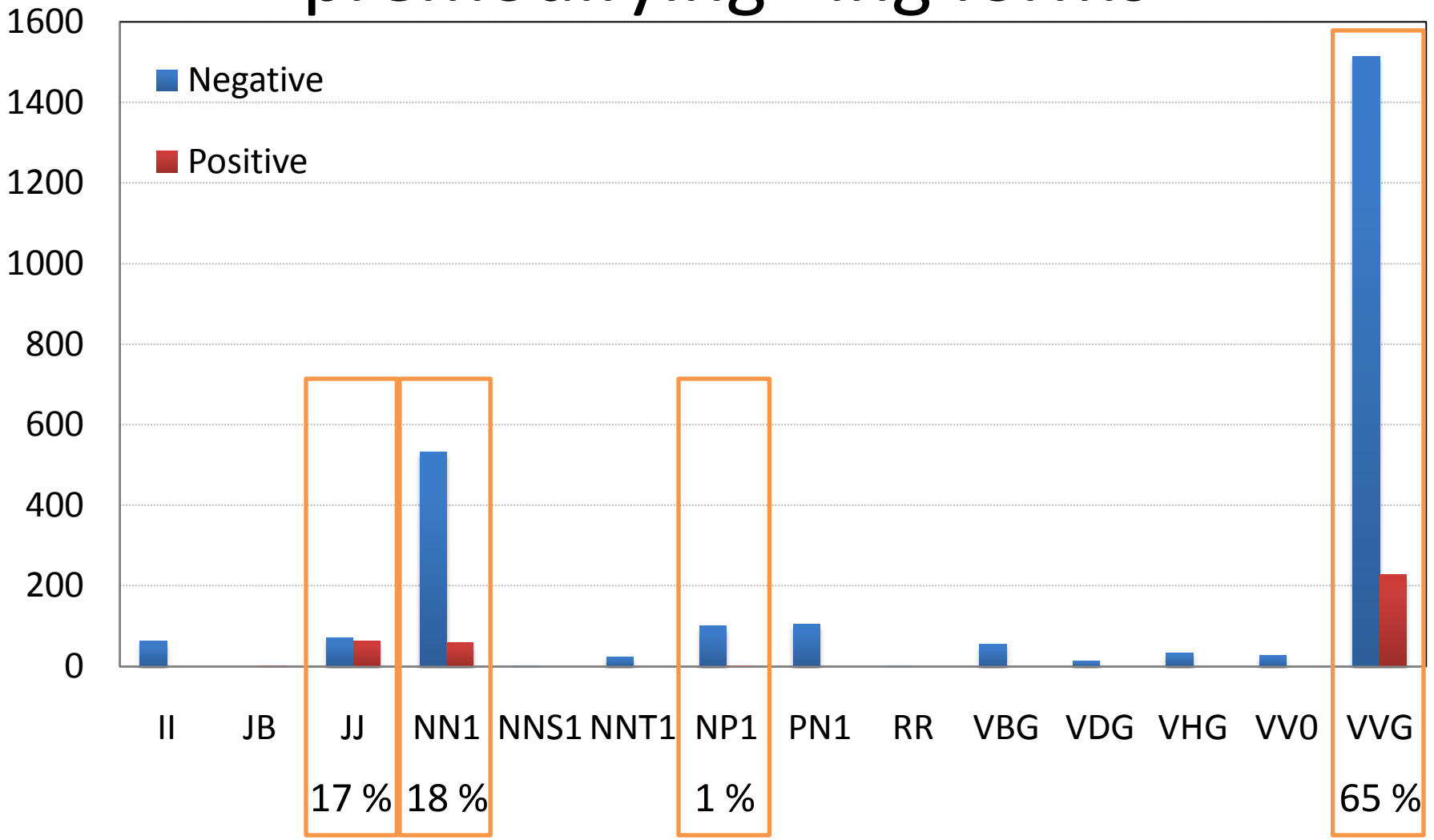
- Q1: How have these been annotated?
  - Did the POS taggers produce the same annotation?

- Q2: Can we retrieve *only* the premodifying –ing forms?
  - Does the parser give us the necessary information to query the corpus?

CLAWS 5: Annotation of premodifying –ing forms



RASP POS: Annotation of premodifying –ing forms



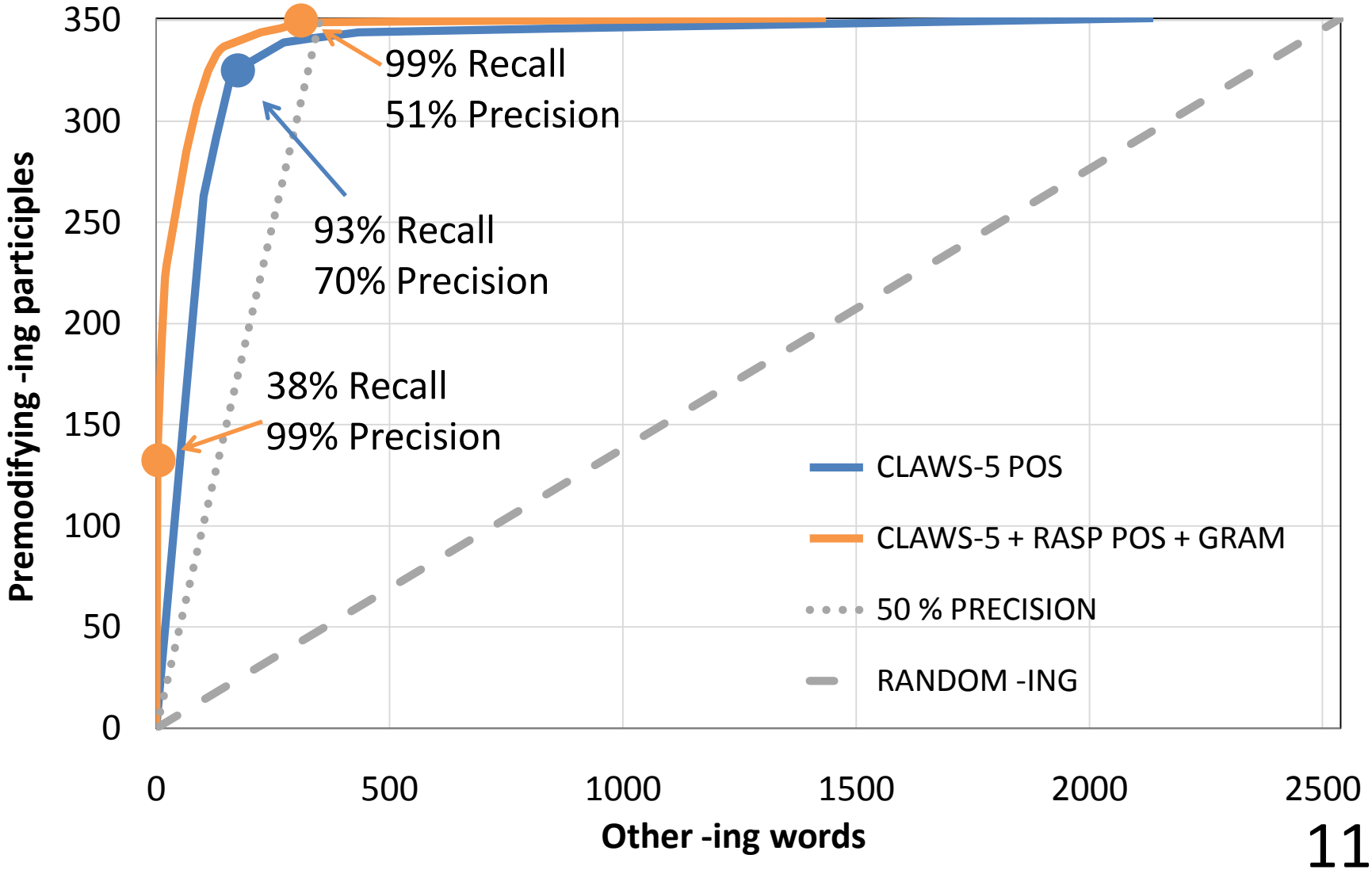
Querying the parsed BNC (1/2)

- Construct rule / query
  - Word x relevant iff
  - “C5 (x) = ADJ” or “RASP (x) = JJ” (POS rule), and
  - “ncmod (y, x)” where y > x (premodifier rule)
- Decision tree classifier seems suitable
- Gives too simple a model
  - Many negative, few positive examples
    - Favours negative
  - Many tags with only few examples
    - Favours not using the attribute at all

Querying the parsed BNC (2/2)

- Solution:
  - Cross-tabulate all possible rules
  - Incrementally select rules using precision
- Simple model works fine!
  - "C5 = ADJ" (BNC-XML)
    - 70 % precision
    - 93 % recall
  - Rule with BNC-XML and RASP
    - 71 % precision
    - 96 % recall
    - Or very high precision / recall
    - Still room for improvement

Trade-off curve for different features



Pilot study

- Preliminary comparison of two genres:
  1. Academic and non-academic scientific texts (natural sciences; social sciences)
  2. Imaginary prose (novels)

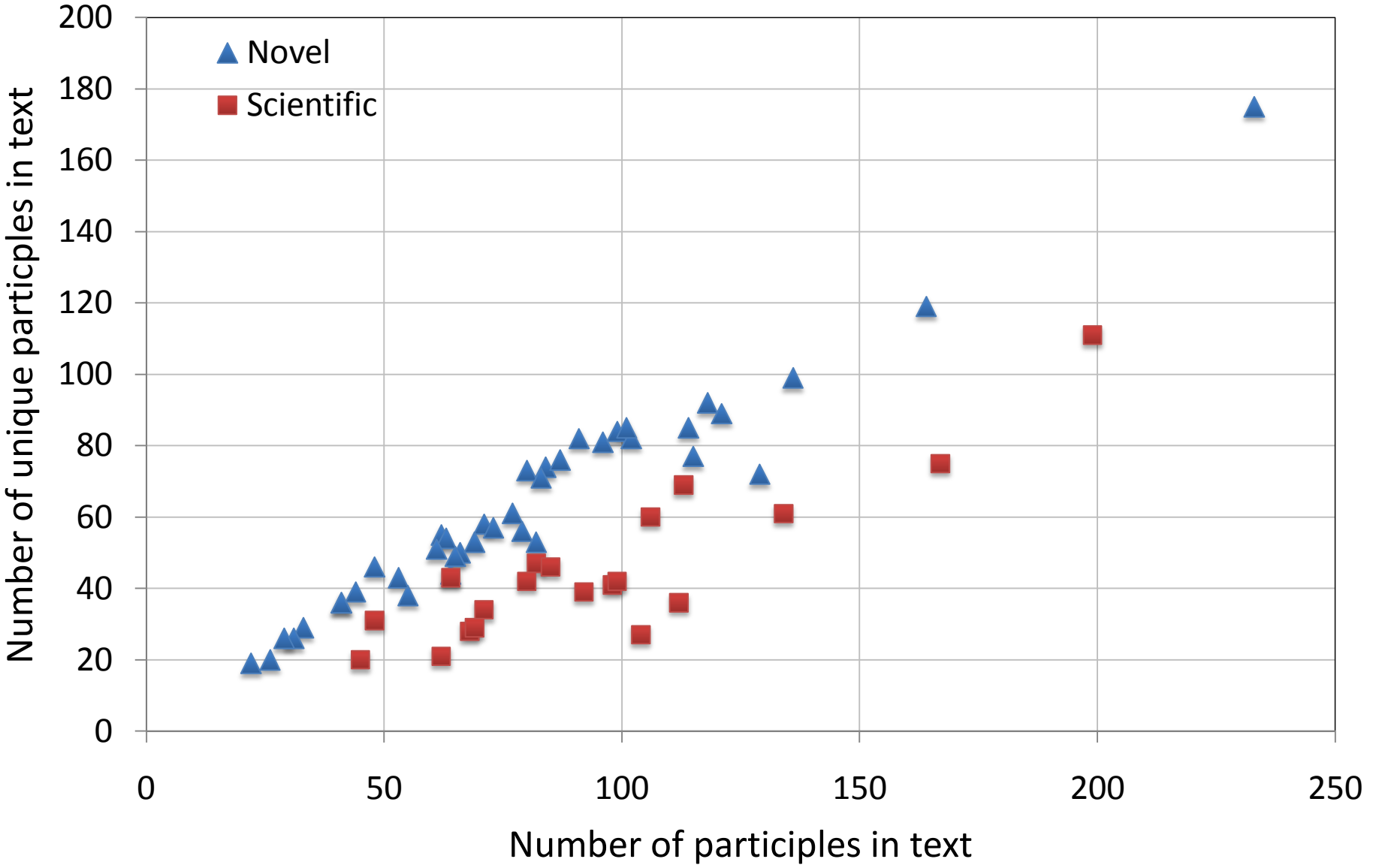
- 50 files from the parsed BNC
- 2,304,371 words
- 5,106 premodifying –ing participles

DAMMOC
DATA MINING TOOLS
for changing modalities of communication

Pilot study

- The average frequency of –ing participle tokens is high in the scientific domain
  - However, the number of participle types is consistently lower in scientific texts than in imaginary prose.

Type/token ratio per genre



Explaining the differences

- Scientific texts:
  - Topical words (e.g. the leading stars)
  - Cohesive words (e.g. following, preceding, foregoing, succeeding...)
- Imaginary prose:
  - Cohesive participles rare
  - More variation in the use of –ing participles in general

Conclusion

- We can efficiently find premodifying –ing participles using information both from the BNC-XML and the parsed BNC.
- The pilot study will provide the basis for a detailed study of –ing participles in the BNC.

References

- Briscoe, E., J. Carroll and R. Watson (2006) The Second Release of the RASP System. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia.
- Andersen, O., J. Nioche, E. Briscoe and J. Carroll (2008) `The BNC parsed with RASP4UIMA'. In Proceedings of the Sixth Language Resources and Evaluation Conference (LREC), Marrakech, Morocco.