

Abstract

We model reading times relative to pronoun resolution. Instead of using hand-constructed experimental material, we study pronoun resolution in natural text reading. We present a series of models of the first pass reading time on the Dundee Corpus [1]. On the one hand, we show that the models confirm factors of influence described in a large psycholinguistic literature. On the other hand, we present how the models can be used to study new hypotheses about factors of influence on pronoun resolution.

Acknowledgments

Thanks to Yair Haendler and Barbara Hemforth for their advise. We also thank the three anonymous reviewers for their comments. This work was supported by the Labex EFL (ANR-10- LABX-0083).

Eye-movement Corpus

- ▶ the English part of the Dundee Corpus [1]
 - ▷ ~50K words from newspaper articles
 - ▷ read by 10 people
- ▶ APADEC annotation layer [2]
 - ▷ 1109 anaphoric pronouns
 - ▷ every pronoun is annotated with its closest antecedent

Reading Time and Reading Regions

Reading times of pronouns are hard to establish [3, 4], because:

- ▶ pronouns are often not fixated
- ▶ pronouns can be read in parafoveal vision
- ▶ the effects of pronoun resolution might be delayed

Solution:

- ▶ six different regions of interest: the word before the pronoun, the pronoun itself and the four words following it

Statistical Model

- ▶ Each region corresponds to a mixed-effects model that predicts log-transformed first pass reading time
 - ▷ textttlme4 R package [5].
- ▶ random intercepts for pronouns and participants
- ▶ control factors: low-level factors
 - ▷ forward probability [6]
 - ▷ backward probability [6]
 - ▷ length in characters of the reading region
 - ▷ log-frequency in the Dundee Corpus of the word in the region
 - ▷ log-frequency in the BNC of the word in the region
 - ▷ log-frequency in the Dundee Corpus of the word previous to the region
 - ▷ log-frequency in the BNC of the word previous to the region
 - ▷ launch position of the first fixation on the region
 - ▷ the landing position of the first fixation on the region
- ▶ pronoun factors: anaphoric relation (next section)

Pronoun Factors

Our study has two goals:

- ▶ Confirm factors described in the psycholinguistic literature
 - ▷ distance between the pronoun and its antecedent [3]
 - ▷ the frequency of the antecedent [7, 4]
 - ▷ *the subject preference* [8]
 - ▷ *grammatical parallelism* [9]
- ▶ Discover new factors and to study less well known factors
 - ▷ the difference between various gender and number features
 - ▷ the distance from the anaphoric pronoun to the beginning of a text: a larger number of antecedent candidates in a text leads to more processing cost of anaphora [10]
 - ▷ the length of the antecedent in words

Results

- ▶ antecedents with higher frequencies in the corpus lead to less reading time
- ▶ a pronoun in the subject position is read faster
- ▶ antecedents with syntactic roles other than subject and direct object slow down the reading
- ▶ the word before the pronoun is read faster when the distance between the pronoun and the antecedent is longer; the reading is delayed later
- ▶ the neutral gender (*it*) leads to more reading time
 - ▷ *it* can be anaphoric and pleonastic. Hence, the increased reading time can come from the decision readers have to take whether the pronoun is anaphoric or not.
- ▶ *they* is read faster compared to the other pronouns
 - ▷ *they* might be less ambiguous than other pronouns.
- ▶ More distance between the pronoun and the beginning of the text leads to more reading time, suggesting that more discourse referents increase the processing load of anaphora.
 - ▷ In future work we will test this hypothesis.
- ▶ For the length of the antecedent no significant effect was found.

Table 1: Coefficient estimates for the models of first pass reading time for the six regions of interest (from one word before the pronoun to four words after, the pronoun being region 0). Control factors and intercepts are not reported.

Region	-1	0	1	2	3	4
Literature Factors						
distance antecedent & pronoun	-4.1E-03 *	2.2E-03	6.4E-04	1.1E-03	4.2E-03 *	2.4E-03
log freq antecedent Dundee	1.8E-03	2.3E-03	2.8E-03	2.7E-03	-5.5E-03	-8.5E-03 *
log freq antecedent BNC	-2.5E-04	1.4E-03	-4.8E-04	-7.2E-05	3.8E-03	5.6E-03
syntactic role pro subj	6.4E-03	8.1E-03	-7.1E-03	-2.0E-02 *	-7.5E-03	-2.7E-03
syntactic role pro other	9.3E-04	5.3E-03	4.2E-04	-3.3E-03	-1.2E-02	-9.5E-03
syntactic role antecedent subj	7.2E-04	3.7E-03	1.4E-02	4.8E-03	1.1E-02	1.1E-02
syntactic role antecedent other	1.6E-03	-1.3E-03	2.0E-02 *	6.7E-04	1.3E-02	-6.6E-03
parallel function	-5.6E-03	-1.3E-02	-2.6E-03	-1.2E-04	-2.8E-03	-1.1E-02
New Factors						
pronoun = they	1.3E-03	1.5E-02	-1.9E-02 *	-8.5E-03	-1.2E-02	7.4E-03
pronoun = it	-1.0E-02	2.5E-02 *	-3.3E-03	-4.9E-03	1.7E-03	1.2E-02
distance pronoun begin	-3.3E-03	-2.4E-04	4.7E-03 *	7.6E-04	-9.2E-05	-6.3E-04
length antecedent	1.0E-03	7.2E-04	3.6E-04	1.4E-03	-1.1E-03	-2.1E-04

Conclusion

Our study on pronoun resolution is — to our knowledge — the first using data from natural text reading. With our data and method, we were able to confirm the influence of factors known from a vast psycholinguistic literature on anaphora resolution, adding evidence for the robustness of these factors. Furthermore, our method allowed us to study new factors of influence.

References

- [1] Alan Kennedy, Robin Hill, and Joël Pynte. The Dundee Corpus. In *Proceedings of the 12th European conference on eye movement*, 2003.
- [2] Olga Seminck and Pascal Amsili. A gold anaphora annotation layer on an eye movement corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, 2018.
- [3] Kate Ehrlich and Keith Rayner. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22(1):75–87, 1983.
- [4] Roger P. G. Van Gompel and Asifa Majid. Antecedent frequency effects during the processing of pronouns. *Cognition*, 90(3):255–264, 2004.
- [5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.
- [6] Stefan L Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological science*, 22(6):829–834, 2011.
- [7] Joël Pynte and Saveria Colonna. Decoupling syntactic parsing from visual inspection: The case of relative clause attachment in french. *Reading as a perceptual process*, pages 529–547, 2000.
- [8] Rosalind A. Crawley, Rosemary J. Stevenson, and David Kleinman. The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 19(4):245–264, 1990.
- [9] Ron Smyth. Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, 23(3):197–229, 1994.
- [10] Olga Seminck and Pascal Amsili. A computational model of human preferences for pronoun resolution. In *Proceedings of the Student Research Workshop of EACL*, pages 53–63, 2017.