Coordinate Structures: On the Relationship between Parsing Preferences and Corpus Frequencies

Ilona Steiner

SFB 441 – University of Tübingen, Germany

steiner@sfs.uni-tuebingen.de

1 Introduction

According to the *tuning hypothesis* (Cuetos et al., 1996), initial parsing preferences in syntactically ambigous sentences are determined by people's exposure to similar structures in the past with the result that people prefer the most frequently occurring resolution of an ambiguity. Under this proposal parsing preferences and corpus frequencies should be correlated. The status of this hypothesis is currently under investigation.

Gibson and Schütze (1999) investigated disambiguation preferences in English noun phrase conjunction of the form "NP_1 *Prep* NP_2 *Prep* NP_3 *and* NP_4" and did not find a correlation to corpus frequencies. The authors conclude that "...the sentence comprehension mechanism is not using corpus frequencies in arriving at its preference in this ambiguity and hence the decision principles of sentence comprehension and sentence production must be partially distinct". Another finding that is problematic for the tuning hypothesis comes from the Dutch language. Mitchell and Brysbaert (1998) report a discrepancy between the NP_1 attachment preference they found in reading times (NP_1 *Prep* NP_2 *Relative Clause*) and the corpus findings. On the other hand, Desmet et al. (2002) showed that the experimental stimuli used in Mitchell and Brysbaert (1998) were not representative for the sentences in their corpus. A reanalysis of the Mitchell and Brysbaert (1998) corpus revealed that when the corpus data are controlled carefully to match the experimental sentences (here: if an NP refers to a human entity) the discrepancy between sentence comprehension and sentence production disappears.

In this paper we investigate the relationship between two types of linguistic evidence, namely parsing preferences and corpus frequencies, with respect to coordinate structures in English. We focus on two processing effects that have been found in reading-time studies: 1) the parallel-structure effect and 2) the disambiguation preference in noun phrase vs. sentence coordination. These will be compared to the corresponding corpus data in the English Verbmobil treebank (TÜBA-E, Hinrichs et al. (2000)). The TÜBA-E treebank consists of spoken dialogs in the domain of business appointments and has been annotated manually at the levels of morpho-syntax (parts-of-speech categories), syntactic phrase structure and function-argument structure. We decided to use a corpus of spontaneous speech, i.e., unedited naturally occurring dialogues, in order to avoid that the corpus data do not reflect solely mechanisms of sentence production, but also intervening factors that are due to editing processes. [1]

2 The Parallel-Structure Effect

In a coordinate structure the second conjunct is read faster when it is structurally similar to the first one (Frazier et al., 2000). The noun phrase [*a short poem*] is processed faster in (1a) than in (1b), since the two conjoined NPs in (1a) are structurally identical, which does not hold for the sentence in (1b).

(1) a. *Terry wrote* [$_{NP}$ *a long novel*] *and* [$_{NP}$ ***a short poem***].
    b. *Terry wrote* [$_{NP}$ *a novel*] *and* [$_{NP}$ ***a short poem***].

We examined whether the preference for structural similarity in coordination is also present in corpora. Therefore we analysed a fraction (ca. 3000 sentences) of the TÜBA-E corpus containing 274 occurrences of coordinate structures (with conjunction *and*; within complete sentences). We manually inspected each occurrence to determine the degree of redundancy in the conjuncts (100% meaning both conjuncts having exactly the same structure including the parts-of-speech categories; 0% meaning that not a single syntactic node is redundant). The syntactic annotation in the treebank was used as basis for this process.

Our analysis showed a gradual distribution of redundancy ranging from 0% to 100%. The sentence in (2a) is an example from our dataset with 88% degree of redundancy, whereas the conjuncts in (2b) are having identical structures (100% redundancy).

(2) a. *Monday looks pretty good except for I have* [$_{NP}$ *a early morning meeting*] *and* [$_{NP}$ *a lunch meeting*]
    b. [$_{PP}$ *on the twenty sixth*] *and* [$_{PP}$ *on the twenty seventh*] *I am busy all day*

36% of all occurrences of coordinate structures contained conjuncts that have an identical structure, i.e., 100% redundancy, as in (2b). In order to interpret these figures, we calculated the degree of redundancy that occurs randomly. Therefore, we extracted for each first conjunct of our coordination dataset a random second "conjunct", which is a phrase randomly chosen from the corpus and matched with the original second

[1] The importance of using unedited corpus data for studying language production is emphasised in Gibson et al. (1996b).

conjunct in syntactic category, grammatical function and length. An example of a randomly extracted phrase for the coordinate structure in (2a) is [$_{NP}$ *the twenty fifth*]. It matches with the original second conjunct [$_{NP}$ *a lunch meeting*] in syntactic category (NP), grammatical function (Complement) and length (3 words).

13% of the random dataset (pairs of original first conjunct and random second conjunct) showed exactly the same structure (i.e., 100% redundancy). The difference of the coordination data and the random dataset with respect to structural similarity (36% vs. 13%) is highly significant ($\chi^2(1) = 72.1$; $p < 0.001$). I.e., structural similarity within coordination is significantly more frequent in our corpus than structural similarity of two phrases independent of coordination. These corpus findings match the preference for structural similarity in coordination found in reading-time studies.

Furthermore, a closer inspection of the corpus data revealed a length effect with respect to structural redundancy in the conjuncts: the shorter the conjuncts, the more frequent structural similarity occurs.

3 Disambiguation Preference: NP vs. S Coordination

Frazier (1979) compared reading times of sentences containing two conjoined noun phrases, as in (3a), with sentences that contain the coordination of two sentences (3b). When encountering the conjunction in (3a,b) the parser is faced with a local ambiguity that cannot be resolved prior to the last word.

(3) a. *Peter kissed* [$_{NP}$ *Mary*] *and* [$_{NP}$ *her sister*] *too.*
    b. [$_S$ *Peter kissed Mary*] *and* [$_S$ *her sister laughed*].

The results show a garden-path effect, i.e. significantly longer reading times, at the last word *laughed* in (3b) compared to the last word *too* in (3a). These results indicate that the parser has the preference to interpret the noun phrase *her sister* as part of a conjoined noun phrase and not as the beginning of a new sentence.

We examined whether the preference for NP coordination (compared to sentence coordination) is also present in the corpus data. Therefore we extracted all occurrences from our coordination dataset that have the form "*...*NP_{Subj}*...Verb... *NP_{Obj} *and...*" and that allow (semantically) the continuation with a noun phrase or a sentence. Example (4a) shows a relevant construction from our dataset that continues with a noun phrase, whereas the one in (4b) continues with a sentence.

(4) a. *and I will bring* [$_{NP}$ *the doughnuts*] *and* [$_{NP}$ *coffee*] *I guess*
    b. *and* [$_S$ *Friday I have a nine to ten meeting*] *and* [$_S$ *I also have a meeting in the early afternoon*]

Our analysis showed that in 64% of the relevant constructions, the sentence continued with a noun phrase (NP coordination), only 32% continued with a sentence (S coordination). Thus we found another case of corpus frequencies matching the disambiguation preference (here: NP vs. S coordination) in reading-time studies.

4 Conclusion

We have shown that for both processing effects (the parallel-structure effect and the disambiguation preference NP vs. S coordination) a correlation between parsing preferences and corpus frequencies can be established. There are two general possibilities to explain these correlations. First, as stated in the tuning hypothesis, the human parser might be sensitive to the statistical patterns occurring in natural language, as, for example, the relative frequencies in corpus data. Here, the statistics in language production is seen as the cause for the development of parsing preferences.

There is however another possibility to account for the correlations, namely a common source for language production and comprehension. There are several structure-based accounts that are able to explain the parsing preferences discussed above. The parallel-structure effect, for example, can be explained by a recycling mechanism, which exploits the structural redundancy in the conjuncts (Steiner, 2005). In this account the human parser, when processing the second conjunct, reuses the structure that was already built up for the first conjunct. The preference for noun phrase conjunction as opposed to sentence coordination can either be derived from the *Minimal Attachment Principle* (Frazier, 1979; Frazier and Clifton, 1996), which favours the structure that requires fewer syntactic nodes, or it might also be due to a general preference for recency (see, e.g, Gibson et al. (1996a)). In these structure-based accounts the preferred construction is usually more economical and therefore easier and faster to process. If the same mechanisms are also active during language production, the constructions that are easier to understand would also be easier to produce and are presumably produced more often. A common mechanism for language production and comprehension would lead to faster reading times during parsing and to higher frequencies during production.

With the present study we are not able to differentiate between the two possibilities, but we could show that production and comprehension preferences are closer to each other than expected. And if it can be shown that the correlation between these two types of linguistic evidence holds in general, an interesting consequence would arise, namely that corpus data can be used to evaluate (at least qualitatively) models of sentence processing.

References

Cuetos, F., D. C. Mitchell, and M. M. B. Corley (1996). Parsing in different languages. In M. Carreiras, J. E. Garcia-Albea, and N. Sebastian-Galles, eds., Language Processing in Spanish, pp. 145–187. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Desmet, T., M. Brysbaert, and C. De Baecke (2002). The correspondence between sentence production and corpus frequencies in modifier attachment. The Quarterly Journal of Experimental Psychology, **55A**(3):879–896.

Frazier, L. (1979). On Comprehending Sentences: Syntactic Parsing Strategies. Ph.D. thesis, University of Connecticut, Storrs.

Frazier, L. and C. Clifton (1996). Construal. MA: MIT Press.

Frazier, L., A. Munn, and C. Clifton (2000). Processing coordinate structures. Journal of Psycholinguistic Research, **29**(4):343–370.

Gibson, E., N. Pearlmutter, E. Canseco-Gonzalez, and G. Hickok (1996a). Recency preference in the human sentence processing mechanism. Cognition, **59**:23–59.

Gibson, E. and C. T. Schütze (1999). Disambiguation preferences in noun phrase conjunction do not mirror corpus frequency. Journal of Memory and Language, **40**:263–279.

Gibson, E., C. T. Schütze, and A. Salomon (1996b). The relationship between the frequency and processing complexity of linguistic structure. Journal of Psycholinguistic Research, **25**(1):59–92.

Hinrichs, E. W., J. Bartels, Y. Kawata, V. Kordoni, and H. Telljohann (2000). The VERBMOBIL treebanks. In Proceedings of KONVENS 2000.

Mitchell, D. C. and M. Brysbaert (1998). Challenges to recent theories of crosslinguistic variation in parsing: Evidence from Dutch. In D. Hillert, ed., Syntax and semantics: A crosslinguistic perspective, pp. 313–335. San Diego, CA: Academic Press.

Steiner, I. (2005). On the syntax of DP-coordination: Combining evidence from reading-time studies and agrammatic comprehension. In S. Kepser and M. Reis, eds., Linguistic Evidence, pp. 507–527. Mouton de Gruyter.