Prosodic cue weighting in disambiguation: Case ambiguity in German

Anja Gollrad, Esther Sommerfeld, Frank Kügler

Department of Linguistics, Potsdam University, Germany

agollrad@googlemail.com, esther.sommerfeld@gmail.com, kuegler@uni.potsdam.de

Abstract

Previous work has shown that speakers and listeners efficiently exploit prosodic information to make the meaning of syntactically ambiguous sentences explicit. However, quantifiable phonetic properties of prosody in speech production (segmental duration, pause duration and fundamental frequency (f0)) stand in a complex relationship to the percept they invoke in the auditory domain. Not all measurable prosodic differences are actually used in sentence parsing. This study investigates the prosodic cues used by speakers to disambiguate a German case ambiguity in order to examine to which degree the individual cues contribute to disambiguation in perception. In a series of perception experiments sentences were consecutively manipulated to verify whether segmental duration, pause duration or pitch was one of the cues used by listeners in assigning a syntactic structure. Our findings show that durational cues are sufficient for listeners to identify the reading speakers assigned to the structures, whereas solely f0 information does not allow listeners to disambiguate the structures.

Index Terms: prosodic disambiguation, speech production, speech comprehension, ambiguity, prosodic manipulation

1. Introduction

One of the central issues in theories of parsing has been what kinds of linguistic information are considered during sentence processing. Much research has been devoted to that issue which revealed that not only syntactic information is used to guide the initial hypothesis about a sentence's syntactic structure but also non-syntactic information (e.g., prosody) can influence the initial decision of the parser [3],[7] among others. For many ambiguities prosody is the only information available to make the meaning of syntactically ambiguous sentences explicit. For example, when the sentence ´It's Johns turn to pay Anna`, is spoken as a single phrase with a prosodic boundary at the end of the word ´Anna`, it is usually understood to mean that John will pay Anna. However, when the same string of words is spoken as two phrases, with prosodic boundaries at the end of the words ´pay` and ´Anna`, the verb is interpreted to be intransitive, and the sentence is understood to be addressed to Anna.

Previous research has shown that listeners exploit such prosodic information in speech comprehension to determine the speaker's intended meaning for ambiguous utterances. Lehiste [3] for example conducted a combined production and perception experiment and selected a set of 15 ambiguous English sentences which were recorded by four naïve speakers. After pointing out the possible meanings to the speakers, the sentences were recorded twice asking the speaker to make a conscious effort to convey each of the sentence's meanings. All three productions of each sentence by each speaker were presented to listeners who were asked to identify the meaning intended by the speakers. The results indicated that 10 out of 15 sentences were disambiguated. The acoustic analysis revealed that word duration appeared to be the primary cue speakers used to disambiguate the sentences. The duration of words increased directly preceding stronger boundaries, i.e., the word "men" in (1a) precedes a stronger boundary due to the closure of the prosodic phrase "old men" and is thus lengthened as compared to "men" in (1b) which appears to be phrase initially (the "||" diacritic stands for a strong boundary). Other correlates of strong boundaries found in Lehiste [3] were laryngealization and insertion of pauses, but they happened to be less systematic.

(1) a. the (old men) (and woman)
the old | men || and woman

b. the old (men and woman)
the old || men | and woman

To investigate which of these phonetic cues are exploited by listeners in speech comprehension Lehiste and colleagues [4] carried out a second experiment with manipulated sentences in order to verify whether segmental duration (as opposed to pitch) was one of the cues used by listeners in assigning the syntactic structure. Pitch was set to a constant value of 100 Hertz for the sentences, and duration was manipulated. The results showed that durational cues indeed determined the reading listeners assigned to the structures.

Independently of prosodic cues provided in the signal, sentence processing theory is interested in the question whether one of two ambiguous readings is preferred in the process of language comprehension. For example it has been claimed that a reading is preferred when it is syntactically simpler in that it contains fewer syntactic nodes (cf. Garden-Path-Model with its parsing principle Minimal Attachment (MA) [6]. According to such an approach, in our data an advantage for (2b) over (2a) is predicted since the genitive DP in (2a) is more expanded and thus structurally more complex [10]. Additionally, it has been claimed that when the parser faces optionality, it preferably interprets incoming elements as arguments (like in 2b) and not as adjuncts (like in 2a) (see [11], [12]).

In the present study, we tested whether German listeners exploit durational cues and/or effects of intonation to resolve German case ambiguities as in (2). A production and a series of three perception experiments with controlled temporal as well as intonational manipulation of sentence` ambiguous constituents had been carried out to disentangle if prosodic information provided in the signal contributes to German ambiguity resolution differently.

2. Experiments

The speech production experiment was conducted to investigate which prosodic cues speakers use (i.e., duration of segments and pauses, prefinal lengthening, height of pitch accents) to disambiguate a German case ambiguity like (2).

2.1. Production

2.1.1. Material

The experimental sentences contain a matrix and subordinate clause. The matrix clause contains a sentence adverb followed by three NPs and a verb. In the matrix clause a local ambiguity arises with respect to the NP1-NP2 complex. In (2a) NP2 is a possessive modifier of NP1. In (2b) NP2 represents a dative object. In both readings, the linear succession of words up to and including NP3 is identical as is the phonological structure of tones. All NPs are associated with a rising pitch accent (L*H), and the right edge of the matrix clause is associated with a high boundary tone (H%) indicating a sentence continuation [8]. A prosodic difference between (2a) and (2b) may arise by means of pitch register differences as has been shown for German [9]. Sentences are disambiguated on encountering the verb.

(2a) L*H L*H L*H H%
Neulich hat [der Mann der Nachbarin] # ein Haus gesehen,
Recently the man$_{\text{-NOM}}$ the neighbour$_{\text{-GEN}}$ a house$_{\text{-ACC}}$ see
"Recently the man of the neighbour saw a house, that…"

(2b) L*H L*H L*H H%
Neulich hat [der Mann] # [der Nachbarin] ein Haus geschenkt,
Recently the man$_{\text{-NOM}}$ the neighbour$_{\text{-DAT}}$ a house$_{\text{-ACC}}$ give
"Recently the man gave the neighbour a house, that…"

The stimuli are controlled for number of syllables, stress pattern and gender. All NP1s are masculine disyllabic trochees, all NP2s are feminine trisyllabic trochees and all NP3s are of neuter gender and monosyllabic. The experimental sentences are highly sonorant to allow for a maximally accurate pitch analysis. Each stimulus sentence was assigned to both the genitive and the dative condition. The resulting 12 sentences were interspersed with numerous fillers and fed into linger software [5]. The experimental sentences were pseudo-randomized for each subject such that sentences of the same condition did not appear adjacently and corresponding sentences of the two conditions had a maximal distance.

2.1.2. Subjects

18 speakers participated in the experiment. All were female undergraduate students at the University in Potsdam and were residents of Potsdam/Berlin surrounding areas. All were native speakers of German and reported no speech or hearing impairment. They either received course credit or were paid for participation.

2.1.3. Method

For each sentence, a context question in broad focus, spoken by a male native voice, had been previously recorded. The contexts were presented together with a target sentence both visually on screen and auditory over headphones. The items were presented on a 15'' computer screen. Participants were asked to read and listen to the context and then speak out the answer displayed on the screen as a response to the question. Subjects were familiarized with the task through written and verbal instructions, followed by three practice trials. In case of hesitations or false starts, participants were asked to repeat the sentence. Recordings took place in a sound-proof chamber equipped with an AT4033a audio-technica studio microphone, using a C-Media Wave soundcard at a sampling rate of 44.1 kHz with 16 bit resolution. Presentation flow was controlled by the experimenter, and participants were allowed to take a break whenever they wanted.

All 216 (18x12) target sentences were hand-annotated by a trained student and subjected to phonetic analysis using Praat software [1]. Duration of the adverbial phrase, of each NP plus the pause between NP1 and NP2 were measured. Pitch analysis was conducted using a Hanning window of 0.4 seconds length with a default 10 ms analysis frame. The pitch contour was smoothed using the Praat [1] smoothing algorithm (frequency band 10 Hz) to diminish microprosodic perturbations. Stylized pitch tracks were calculated (Figure 1). For this purpose, each constituent in (2) was divided into five equal-sized intervals, and the mean F0 (in Hz) per interval was aggregated over all speakers and sentences for each interval. The resultant values were interpolated for each condition.

2.1.4. Results

Figure 1 shows the mean pitch track, averaged over all speakers, of the ambiguous sequence, i.e., the adverbial phrase, the first NP, the second NP and the third NP for both, the genitive and the dative condition.

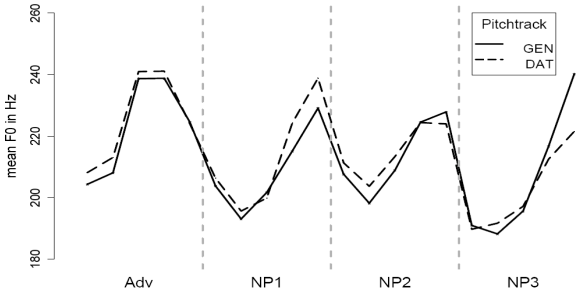

Figure 1: *Time-normalized pitch tracks, based on five measuring points per constituents, showing the mean across all speakers (genitive sentences (GEN, solid line); dative sentences (DAT, dashed line))*

Figure 1 shows that NP1 in the dative condition and NP3 in the genitive condition is realized with a higher f0-excursion compared to the other condition, respectively.

Statistical analysis confirmed these observations. We fit a multilevel model [5] using crossed random factors subject and item, and condition (GEN, DAT) as fixed factors. The analysis relied on f0 maximum as dependent variable. The statistical comparison revealed a significant effect of f0 maximum within the NP1 interval (GEN: 237.39 Hz; DAT: 250 Hz, t=3.09) and a significant effect of f0 maximum within the NP3 interval (GEN: 255.12 Hz; DAT: 237.67, t=6.92). The analysis of segmental duration (Figure 2) showed a significant effect of duration within the first (GEN: 397ms; DAT: 627ms, t=5.97), the second (GEN: 658ms; DAT: 554ms, t=-4.21) and the third NP interval (GEN: 465ms; DAT: 413ms, t=-3.81). Additionally, the statistical comparison of pause duration between NP1 and NP2 in both condition revealed a significant effect (GEN: 28ms; DAT: 104ms, t=2.94).

Speakers significantly altered their production of the utterance in ways that were consistent with the intended structure (i.e., providing a higher pitch excursion on and a longer duration of NP1 plus a subsequent prosodic break in the dative condition (2b) as opposed to the genitive condition (2a)).

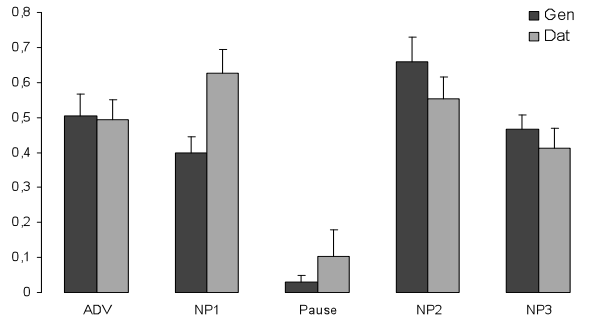

Figure 2: *Mean duration times in seconds per constituents and pause across all speakers.*

2.2. Perception

To test whether listeners use these different prosodic cues observed in the production experiment for disambiguation, a series of perception experiments were conducted.

2.2.1. Material

20 sentences of each case condition were pronounced by a trained speaker incorporating the intonational patterns of the production study, (i.e. lengthening of NP1 in the dative condition followed by a longer pause and a higher rise in pitch on NP1 compared to the genitive condition). Prosodic structures were confirmed by phonetic measurements of f0 and duration for all words in the temporarily ambiguous region. Sentences that did not match the mean intonational contour of the production study were replaced by new recordings. Subsequently, the disambiguating part of all target sentences (i.e., the verb and the following relative clause) was cut off. The resulting sentence fragments were taken to test the extent to which listeners use the information in auditory parsed sentence fragments to predict upcoming entities.

2.2.2. Method

Each sentence fragment was presented to 20 listeners in a quiet room via loudspeakers. In a forced-choice task subjects listened to each sentence fragment up to and including NP3, and then were asked to complete the sequence by choosing the better fitting of two offered verb continuations (genitive verb/dative verb) in a questionnaire. The target sentences were intermixed with numerous filler sentences that varied in prosodic and syntactic structure, in order to discourage response strategies.

2.2.3. Results

The results of the sentence completion experiment (Figure 3) display a very clear pattern in that in 67% the correct gentitive verb was selected when a gentive sequence was presented. When a dative sequence was presented, listeners selected in 93% the prosodically fitting dative verb. For the statistical, frequency-based analysis, we fit a multilevel model [5] using crossed random factors subject and item, and sentence sequence (GEN, DAT) as fixed factors. The analysis relied on the verb as dependent variable. The statistical comparison revealed a significant effect of the selected verb in that independent of the presented auditory sequence, listeners select the prosodically correct sentence continuation significantly ($z=5.16$, $p>0.01$) more often, confirming the actual use of prosodic cues to identify the intended syntactic structure.

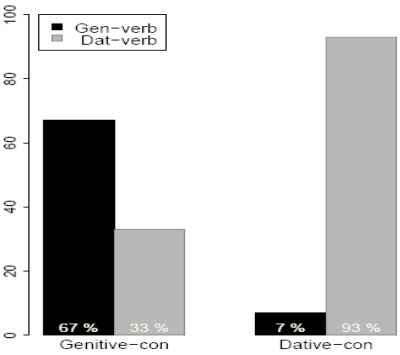

Figure 3: *Percentage of the selected sentence continuation*

2.2.4. Perception with duration only

Given that listeners reliably differentiate between the case ambiguous sentences on the basis of natural stimuli (Figure 3), we wanted to investigate in a following perception experiment whether listeners use durational- or rather f0-cues for disambiguation. To that end, a perception experiment with manipulated speech was conducted. The f0-contour was flattened at a 120 Hz-level by using a Praat script [1] while the durational properties of the sentences have not been changed. The f0-manipulated sequences were then auditory presented via loudspeakers to 20 participants. The target sequences were again intermixed with filler sentences from four unrelated experiments to avoid response strategies. Participants listened to each manipulated sequence and then were asked to complete the sentences in the forced-choice manner described above.

2.2.5. Results

Solely on the basis of durational cues, listeners selected the prosodically fitting sentence continuation in 65% when they listened to genitive sequence (Figure 4). When confronted with a dative sequence, subjects selected in 87% the correct verb. The statistical, frequency-based analysis confirmed these observations and revealed a significant effect of the selected verb ($z=7.50$, $p<0.01$). Comparing these results with the outcome of the unmanipulated stimuli, a similar pattern emerges in that independent of the presented auditory sequence, listeners select significantly more often the fitting verb. This result suggests that listeners draw upon durational cues when resolving syntactic ambiguity.

2.2.6. Perception with f0 only

To test if isolated pitch information is as sufficient as isolated durational cues for disambiguation, an experiment using manipulated prosodic information has been carried out. That endeavour was approached by manipulating the durational cues in the following way: the duration of each constituent had been adjusted according to the mean duration for every constituent in order to have each manipulated constituents carrying the mean of the length of the original sounds (see [3]

for a description of this method). To avoid response strategies favouring the dative interpretation, pause duration between NP1 and NP2 of the dative condition was reduced to equal pause in the genitive condition. The durational manipulated sequences were interspersed with filler sentences from four unrelated experiments and were auditory presented via loudspeakers to 20 participants. Participants listened to each manipulated sequence and were, likewise to the preceding perception experiments, asked to complete the fragments.

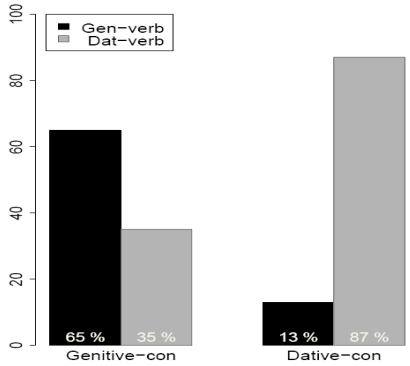

Figure 4: *Percentage of the selected sentence continuation, duration only*

2.2.7. Results

The results of the durational manipulation are depicted in Figure 5. The outcome shows that independent of the presented auditory sequence, listeners choose the dative continuation more often (83 to 85%). The interaction of the auditory sequence and the selected verb is non- significant ($z=1.61$, $p=0.039$). Thus, solely pitch information does not be sufficient to disambiguate the sequences. The manipulated durational properties of the words and pauses disallow listeners to disambiguate the sentences, in both conditions.

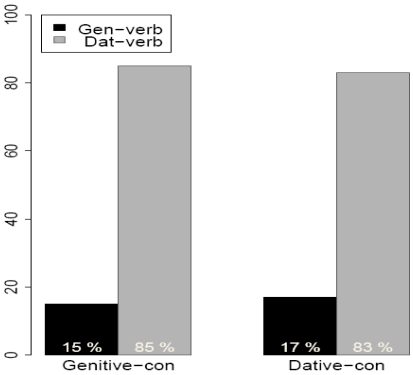

Figure 5: *Percentage of the selected sentence continuation, f0 only*

3. Conclusions

The purpose of the production study was (i) to investigate whether speakers disambiguate German case ambiguous sentences by prosodic means and (ii) investigating which prosodic strategies are employed by speakers in doing so. To this end a total of 126 case ambiguous sentences, produced by 12 speakers, were recorded and phonetically analyzed. The results from these analyses revealed that all speakers were able to reliably disambiguate the sentences by prosodic means. Final lengthening, insertion of pauses and F0 rise was used to make the intended meaning explicit. In a sequence of perception experiments listeners´ reaction to naturally spoken sentences was compared to manipulated sentences. The perception of separated f0-manipulated and durational-manipulated stimuli show that duration, but not F0, is a sufficient cue to syntactic structure in sentence processing.

The phonetic cues exploited by speakers in production to disambiguate a sentence stand in contrast to the prosodic cues used by listeners to disambiguate sentences. Not all measurable differences are used in parsing prosody. With respect to the parsing process, our present findings can be taken as evidence for an incremental parsing strategy according to which ambiguous nominal elements are preferably interpreted as single participants of a ditransitive event in the absence of durational information (Figure 5). Once durational information is available for the processor, the parser correctly interprets ambiguous nominal elements and assigns the intended syntactic structure (Figure 4).

In sum, our findings show that prosodic information, more precisely, durational information, guide ambiguity processing in spoken language in German, as Lehiste and colleagues [3] have been found for English.

4. Acknowledgements

This research was supported by a DFG grant to the project "Prosody in Parsing" at the University of Potsdam. Thanks for helpful comments and suggestions are due to Caroline Féry, Lyn Frazier, Bob Ladd, Shravan Vasishth and Gerrit Kentner. We are also grateful to Bernadett Smolibocki who provided considerable technical assistance and Gerrit Kentner who spoke the stimuli for the comprehension experiments.

5. References

[1] Boersma, P. & Weenink, D., "*Praat-doing phonetics by computer*", Online resource: http://www.praat.org, 1997-2009.

[2] Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, 7, 107-122.

[3] Lehiste, I., Olive, J. P. & Streeter, L. A. (1976). The role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America*, 60, 1199-1202.

[4] Linger: "*Linger-a flexible platform for language processing experiments*", Online resource: http://tedlab.mit.edu/~dr/Linger/

[5] Bates, D. & Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes.* R package version 0.9975-11.

[6] Frazier, L. & Fodor, J. (1978). The sausage machine: A new two-stage parsing model. *Cognition,* 6, 291-325.

[7] Price, P., Ostendorf, M., Shattuck-Hufnagel, S. & Fong, C. (1991). The use of prosody in disambiguation. *Journal of the Acoustical Society of America*, 90, 2956-2970.

[8] Féry, C. (1993). *German Intonational Patterns*. Tübingen: Niemeyer.

[9] Féry, C. & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *J.Phonetics*, 36, 680-703

[10] Radford, A., Anderson, S. R. & Bresnan, J. (1988). *Transformational Grammar: A First Course*. Cambridge: CUP.

[11] Clifton, C., Speer, S. & Abney, S. (1991). Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *J.Memory and Language*, 30, 251-271.

[12] Schütze, C. T. & Gibson, E. (1999). Argumenthood and English prepositional phrase attachment. *J.Memory and Language*, 40, 409-431.