Tapping into semantic recovery : an event-related potential study on the processing of gapping and stripping

Ruijgrok, B.J.

CHAPTER 6

ERP experiments I & II: Structural complexity

This chapter reports two ERP experiments that focus on the effects of the amount of elided structure. I am thankful to Naomi Nota, Olga Kepinska and Ferdi van de Kamp for assistance during some parts of the data collection.

6.1 Modulation of structure in the right conjunct

6.1.1 Introduction

The aim of the experiment reported in this section was to find out the effect of the amount of structure that is elided in gapping-like constructions. In a word-by-word reading task, EEG was recorded while at the same time participants' comprehension was measured. Condition by condition a phrase is cut off. At a fixed measure point across conditions, ERPs were recorded to analyse the effect of increasing amounts of structure. A memory task was included to control for individual variation in relation to memory retrieval (c.f. Kaan et al., 2013). The stimuli used in this experiment were rated by other participants for their acceptability in a computer-administered judgement task (see Section 5.2.2 for the rationale behind the experimental sentences).

At the critical measure point, a copy account predicts a processing cost relatively early, that is, just after encountering a gap. Possibly this would be reflected by a LAN or ELAN component. The subsequent integration of retrieved structure should be relatively easy. A cue-based account, however, predicts the reverse: less retrieval cost and possible ERP effects relatively late in the time course; for example, a P600 reflecting an integration cost.

6.1.2 Methods

Test materials

As explained in chapter 5.2.2, thirty-six quadruplets as in (1) were chosen. For a complete list of test sentences and average acceptability scores see Appendix A. A Latin square design was applied to counterbalance the stimuli so that each participant only saw one sentence of a quadruplet. Along with the 36 test sentences, 72 filler sentences, half of which containing proper names, were added. The test sentences are shown in (1) (repeated from chapter 5.2.2). The critical word *waren* is in bold: ERPs were measured in relation to the presentation of this word. The colour of *waren* corresponds with the type of condition.

(1) a. Omdat Hilde in de voortuin het gazon onderhield, en Ralph
Because Hilde in the front.garden the lawn maintained and Ralph
in de achtertuin de paden harkte, **waren** de buurtgenoten
in the back.garden the paths raked were the neighbours
vrolijk.
happy
'Because Hilde maintained the lawn in the front garden
and Ralph raked the paths in the back garden,
the neighbours were happy.' (***No Gapping***)

b. Omdat Hilde in de voortuin het gazon onderhield, en Ralph
Because Hilde in the front.garden the lawn maintained and Ralph
in de achtertuin de paden, **waren** de buurtgenoten vrolijk.
in the back.garden the paths were the neighbours happy
'Because Hilde maintained the lawn in the front garden
and Ralph the paths in the back garden,
the neighbours were happy.' (***Verb Gapping***)

c. Omdat Hilde in de voortuin het gazon onderhield, en Ralph
Because Hilde in the front.garden the lawn maintained and Ralph
in de achtertuin, **waren** de buurtgenoten vrolijk.
in the back.garden were the neighbours happy
'Because Hilde maintained the lawn in the front garden
and Ralph in the back garden,
the neighbours were happy.' (***Verb-Object Gapping***)

d. Omdat Hilde in de voortuin het gazon onderhield, en Ralph
Because Hilde in the front.garden the lawn maintained and Ralph
ook, **waren** de buurtgenoten vrolijk.
too were the neighbours happy
'Because Hilde maintained the lawn in the front garden
and Ralph too,
the neighbours were happy.' (***Stripping***)

Note that I use a comma to separate clauses as is common in Dutch, especially between two finite verbs. In general, a comma is beneficial for the reader as is suggested by Yang (2010) (see also https://onzetaal.nl/taaladvies/komma-voor-en). I was not interested in any ambiguity or mismatch effects at the point that the main verb of the sentence is processed in Gapping sentences, therefore a second comma is helpful because it would prohibit such possible effects. Also, the finite verb of the main clause consistently disagrees with the subject of the second conjunct in terms of number. The subject of the main clause *de buurtgenoten* is plural, in contrast to the singular subject in the preceding clause *Ralph*. Following Kaan et al. (2013) an additional comma was placed before *en* to ensure that the noun of the right conjunct would not be erroneously interpreted as the second object of the preceding verb.

Participants

Twenty-nine native speakers of Dutch with normal or corrected-to-normal vision took part in this study and were paid €15. Six of them were disregarded from the analysis because they were left-handed (1 participant), because of technical failure (1 participant) or because of too many artefacts according to the (stringent) criteria described below (4 participants). Of the remaining 23 participants, 13 were female and the mean age was 23.26 (age range 19-37). The experiment followed the Ethics Committee regulations of the Humanities Faculty of Leiden University, which approved its implementation. Participants gave informed consent before the study.

Procedure

Participants were comfortably seated in a dimly lit sound-proof room at a distance of approximately 80 cm of a 17 inch CRT monitor. One-hundred-and-eight test sentences were presented in a random order using the presentation software E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA). The sentences were presented word by word in Verdana font (18pt). Each word was presented for 300 ms with a 300 ms fixation cross interval. Presentation of a trial started with a fixation cross for 1,000 ms. Every sentence was followed by a content question to encourage comprehension. A blank screen with a duration of 1,500 ms intervened between the last word of a trial and the presentation of the comprehension question. For half the participants the left response button referred to "YES", for the other half the left button referred to "NO". Participants were given a break after 12 sentences and could proceed at their own pace. The comprehension questions referred to different parts of the sentences equally. Before the actual test, the participants were able to get used to the task with four practice sentences.

The experiment was concluded with a working memory test based on a task described in chapters 4.2.3 and 5.1.1.

The experiment took about 1.5 hours per participant, including EEG set-up.

Apparatus and electrophysiological recording

A description of the recording set-up can be found in chapter 5.1.1.

Data analysis

The EEG data were preprocessed using Brain Vision Analyzer Version 2.0. (Brain Products, Munich, Germany). EOG artefacts were corrected using the Gratton et al. (1983) algorithm. Remaining artifacts were rejected on the basis of the following criteria: trials below -200 μV, above 200 μV, or including a voltage step of 20 μV or more within 200 ms. A low cutoff filter of 0.05 Hz, 24bB/oct and a high cutoff filter of 30 Hz, 24 dB/oct were applied. Epochs of 1,000 ms were computed with a 200 ms pre-stimulus baseline. ERP grand averages were time-locked to the onset of the target word *waren*. As a result, 6.64 % of the trials were excluded from the analysis: of the nine trials presented per condition an average of 8.4 (SE = 0.15) per participant were used.

Both the behavioural data and the electrophysiological data were analysed by means of a repeated measures procedure in **R** using the `ez` package (Lawrence, 2011). To correct for possible sphericity violations in either analysis the Greenhouse-Geisser correction applies. Using the package `lmPerm` (Wheeler, 2010), time windows of interest were empirically determined by permutation tests that were run per sample at each electrode with the independent factor CONDITION from 0 to 800 ms after target onset.[1] This procedure is akin to running multiple t-tests (or ANOVAs) for consecutive, averaged 20 ms time windows (see for example Timmer & Schiller, 2012), however, the advantage of running a permutation test per sample is that the onset and offset of time windows can be determined with even more precision. Also, like non-parametric tests, permutation tests are robust for relatively small samples (Legendre & Legendre, 1998:20). For this thesis, I determined that consecutive significant samples with a duration of at least 50 ms occurring in at least one electrode were taken into consideration for follow-up analyses. These analyses were conducted separately for midline sites (Fz, Cz, Pz) and for the lateral electrode regions: left/right frontal (AF3/4, F7/8, F3/4), left/right central (FC1/2, C3/4, CP1/2), left/right parietal (P7/8, P3/4, PO3/4). In the current experiment a repeated measures ANOVA was planned using within-subjects factors CONDITION (four levels: No Gapping, Verb Gapping, Verb-Object Gapping, Stripping), ANTERIORITY (3 levels: frontal, central, posterior), and, for analyses involving lateral sites, HEMISPHERE (2 levels: left, right).

6.1.3 Behavioural results

The accuracy on the comprehension questions was on average 92.75% (SE = 1.04%). Per condition the accuracy scores were as follows: $M_{\text{No Gapping}}$ = 88.41%, $M_{\text{Verb Gapping}}$ = 94.69%, $M_{\text{Verb-Object Gapping}}$ = 91.79%, and $M_{\text{Stripping}}$ = 96.14%. The accuracy scores differed significantly between conditions as a repeated measures on the ratios showed [$F(3, 66) = 3.50$, $p = .039$, $\eta^2_G = .089$]. A post hoc comparison with Bonferroni correction was carried out to check which conditions differed specifically. Only between conditions No Gapping (***a***) and Stripping (***d***) a notable difference was found, but this did not reach significance [$p = .088$].

Due to a scripting error, the first session of the working memory task (random counting) consisted of four trials instead of five. The accuracy ratio of

[1] I am very thankful to Cesko Voeten who provided a script to make this work, since it made the procedure of combined data analysis in Brain Vision Analyzer and **R** workable.

the three test sessions was on average 65.83% ($SE = 2.65\%$). Per condition the scores were: $M_{\text{Random Counting}} = 66.30\%$, $M_{\text{Auditory Presentation}} = 60.00\%$, $M_{\text{Visual Presentation}} = 71.30\%$. A repeated measures ANOVA by subjects with CONDITION as independent factor and ACCURACY OF NUMBER RECALL as dependent variable showed that the ratios did not differ between conditions [$F(2, 44) = 1.55$, $p = .223$, $\eta^2_G = .045$].

The scores of the comprehension task of the ERP experiment were compared with the scores of the working memory task. A small, non-significant correlation was found between the variables ACCURACY OF SENTENCE COMPREHENSION and ACCURACY OF NUMBER RECALL [$r = .249$, $p = .251$].

6.1.4 Electrophysiological results

Onset of *waren*

In Figure 6.1, the mean amplitudes of all conditions are depicted as measured from the onset of *waren*. Particularly conditions Verb-Object Gapping (***c***) and Stripping (***d***), show a large and sustained negativity as compared to the No Gapping condition (***a***). However, deviations appear to already begin before onset, yielding an erratic baseline. Analysis of effects after such a distorted baseline are unreliable, since they may be attenuated by effects earlier on.[2]

6.1.5 Towards an alternative measure point

The accuracy of the comprehension questions showed that the participants understood the stimuli almost perfectly in all conditions; nonetheless the fewest mistakes were made in condition ***d***, which may not be surprising considering that this condition contained the least amount of "new information" in the right conjunct. Sentence comprehension barely correlated with the working memory scores. It is notable that, while the acceptability study reported in 5.2 showed that condition ***d*** was rated relatively the lowest (together with condition ***c***) in terms of acceptability, it was the best understood according to the comprehension scores of the current experiment. Possibly, the issue of the least amount of new information again plays a role. The comprehension task requires that all of the information of the foregoing sentence remains accessible in memory, which is easier when there is less (new) information. In addition, it could be that the sub-type of Gapping, namely Stripping (as in condition ***d***), is more frequently used than Gapping (as in conditions ***b***-***c***) in Dutch.

A limitation of this study concerns my decision to measure ERPs at the main verb following the subordinate clause, *waren*. This point was chosen because an earlier measure point, immediately before *waren*, would face the problem of different and hence incomparable phrases across four conditions.

[2] I am thankful to the people that attended my poster at the CNS meeting in Boston in 2014 and who politely pointed out my classic rookie mistake – overlooking the erratic baseline.

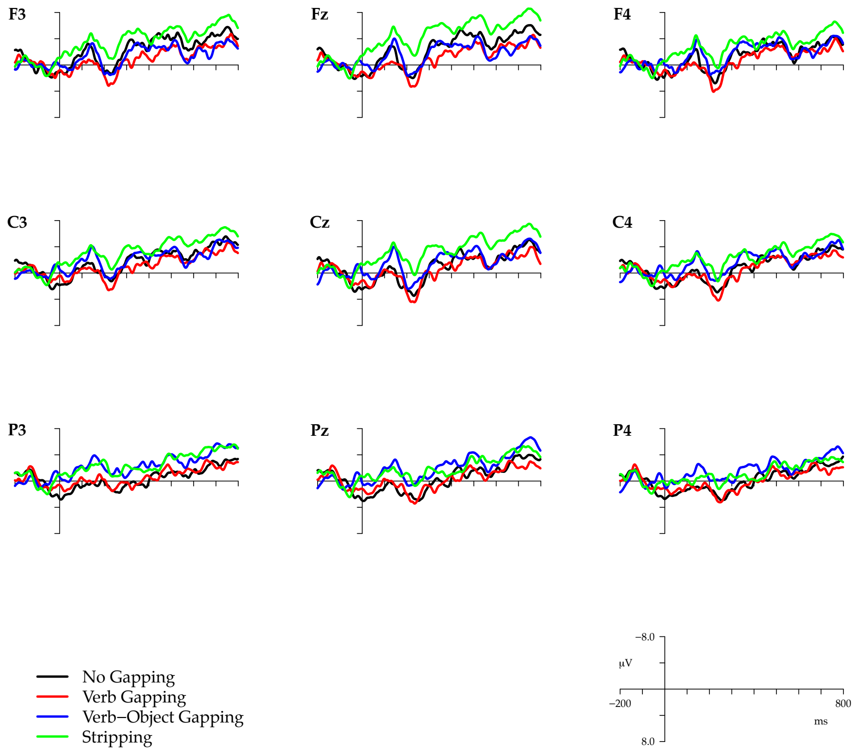

Figure 6.1: Grand averages of all conditions at the onset of *waren* at electrode sites F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. Corresponding example sentences can be found on page 109.

However, my planned measure point appeared to be too late since a proper baseline could not be determined, as explained above. It may be the case that once the elision (or *ook*, for that matter) was being processed, the search for the antecedent started immediately. Also, the comma that was presented with the word before the main verb is a sign for the reader that the resolution process may start, as it marks a clause boundary. This process could have caused a so-called "spill-over effect", as has been noted by Steinhauer and Drury (2012): an effect elicited before the onset of *waren* prevents us from drawing conclusions about any possible component elicited at the onset of *waren*. Steinhauer and Drury (2012) used the spill-over effect to explain unjustified ELAN effects and it may also be applicable for the current study. It seems, therefore, that we tapped into the recovery process at least 300 ms late. Such a delay could even mean missing the recovery process in a condition such as ***b***, in which the ellipsis was small and could potentially have been resolved before *waren*.

A solution to overcome spill-over effects is to use a different baseline, for example, the proper name in the second conjunct. However, this would lead to analysing large time windows, especially in conditions ***a***-***b***. With the current data, this would in turn lead to excluding a considerable amount of trials because of an increased number of artefacts. Considering the already small number of trials in the current design, calculating an alternative baseline was not an option. Additionally, the comma marks a clause boundary as well a prosodic boundary. It is conceivable that the comma induced the ellipsis resolution process. Therefore, it would be even more appropriate for us to find out whether differences between conditions are apparent on words that appeared just before *waren* and that can be compared with appropriate counterparts. In the next section, I will explore a post hoc analysis of such words.

Post hoc analysis of comma effects

It is possible to compare nouns with a comma in conditions ***b*** and ***c*** – amounting to Gapping constructions – with their counterparts which are not accompanied by a comma in the No Gapping condition ***a***. Two different nouns can be taken into consideration for comparison. The first is the object noun. Below I repeat the example of the stimulus set. In (2) it is shown that between conditions ***a*** and ***b*** the object nouns (in bold) can be compared. I hypothesise that an ERP effect can be measured from the onset of ***paden*** in condition ***b*** as compared to condition ***a***, showing the effect of the resolution process of the elided verb of the second conjunct.

(2) a. Omdat Hilde in de voortuin het gazon onderhield, en Ralph
Because Hilde in the front.garden the lawn maintained and Ralph
in de achtertuin de **paden** harkte, waren de buurtgenoten
in the back.garden the paths raked, were the neighbours
vrolijk.
happy
'Because Hilde maintained the lawn in the front garden
and Ralph raked the paths in the back garden,
the neighbours were happy.' (***No Gapping***)

b. Omdat Hilde in de voortuin het gazon onderhield, en Ralph
Because Hilde in the front.garden the lawn maintained and Ralph
in de achtertuin de **paden**, waren de buurtgenoten vrolijk.
in the back.garden the paths, were the neighbours happy
'Because Hilde maintained the lawn in the front garden
and Ralph the paths in the back garden,
the neighbours were happy.' (***Verb Gapping***)

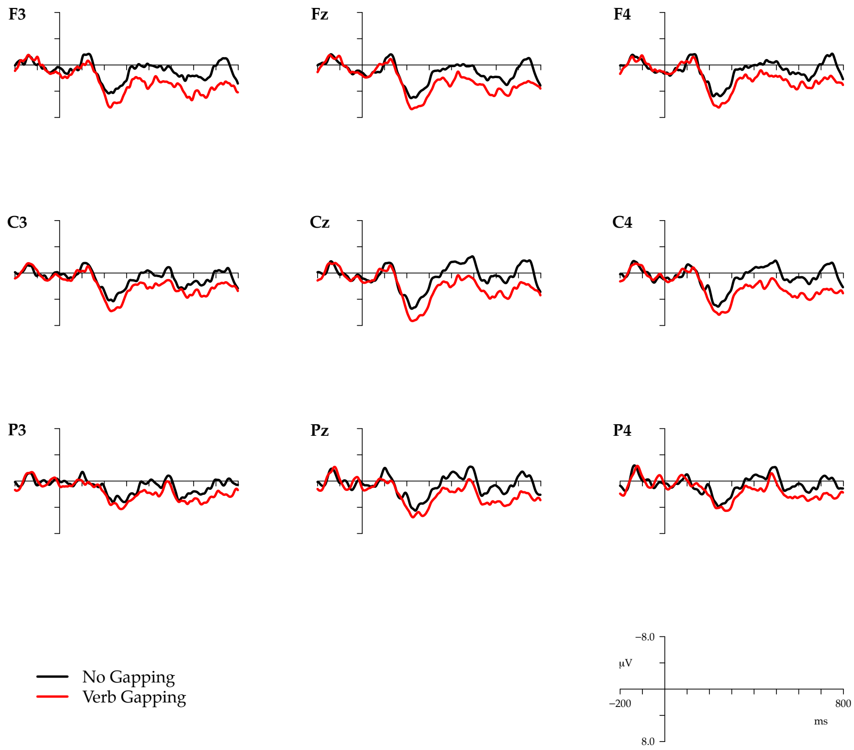

Figure 6.2: Grand averages of conditions No Gapping (***a***) and Verb Gapping (***b***) at onset (y-axis) object noun (*paden*) at electrode sites F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. Corresponding example sentences can be found on page 114.

Plots of the grand averages were evaluated together with a permutation test per sample per electrode. Time windows of interest were empirically determined between 220-520 ms and 570-780 ms. As a check, a repeated measures ANOVA was run for the intermediate time window of 520-570 ms. No effects could be found for the factor CONDITION on midline sites [$F(1, 22) = 1.83, p = .189, \eta^2_G = .027$] and on lateral sites [$F(1, 22) = 1.18, p = .288, \eta^2_G = .017$]).

Object noun: effects in the 220-520 ms time window
On midline electrodes, the factor CONDITION reached significance [$F(1, 22) = 5.37, p = .030, \eta^2_G = .070$]. Additionally, a marginal effect of ANTERIORITY was found [$F(2, 44) = 3.69, p = .054, \eta^2_G = .013$]. A pairwise comparison showed a significant difference between frontal and central

electrodes [$p = .007$, $M_{\text{Frontal}} = 1.83$ ($SE = 0.56$), $M_{\text{Central}} = 0.85$ ($SE = 0.58$)]. No interaction between CONDITION and ANTERIORITY was found. On lateral electrodes, only an effect of CONDITION could be established [$F(1, 22) = 4.54$, $p = .044$, $\eta^2_G = .054$].

Object noun: effects in the 570-780 ms time window

The effect of CONDITION was similar between midline electrodes [$F(1, 22) = 4.87$, $p = .038$, $\eta^2_G = .077$] and lateral sites [$F(1, 22) = 4.88$, $p = .038$, $\eta^2_G = .066$]. No other effects or interactions were found.

Adjunct noun: effects in the 300-430 ms time window

The second noun that can be considered is the noun in the adjunct phrase *in de achtertuin*. (3) shows the relevant conditions. As compared to condition ***a***, it is predicted that in condition ***c*** an effect can be measured at ***achtertuin***, representing the recovery of the VP *het gazon onderhield*.

(3) a. Omdat Hilde in de voortuin het gazon onderhield, en Ralph
Because Hilde in the front.garden the lawn maintained and Ralph
in de **achtertuin** de paden harkte, waren de buurtgenoten
in the back.garden the paths raked, were the neighbours
vrolijk.
happy
'Because Hilde maintained the lawn in the front garden
and Ralph raked the paths in the back garden,
the neighbours were happy.' (***No Gapping***)

c. Omdat Hilde in de voortuin het gazon onderhield, en Ralph
Because Hilde in the front.garden the lawn maintained and Ralph
in de **achtertuin**, waren de buurtgenoten vrolijk.
in the back.garden, were the neighbours happy
'Because Hilde maintained the lawn in the front garden
and Ralph in the back garden,
the neighbours were happy.' (***Verb-Object Gapping***)

Plots of the grand averages were evaluated together with a permutation test per sample and per electrode. A time window of interest was determined between 300-430 ms.

On midline electrodes, the factor CONDITION reached significance [$F(1, 22) = 4.83$, $p = .039$, $\eta^2_G = .048$] as well as on lateral sites [$F(1, 22) = 6.69$, $p = .017$, $\eta^2_G = .059$]. No other effects or interactions could be found.

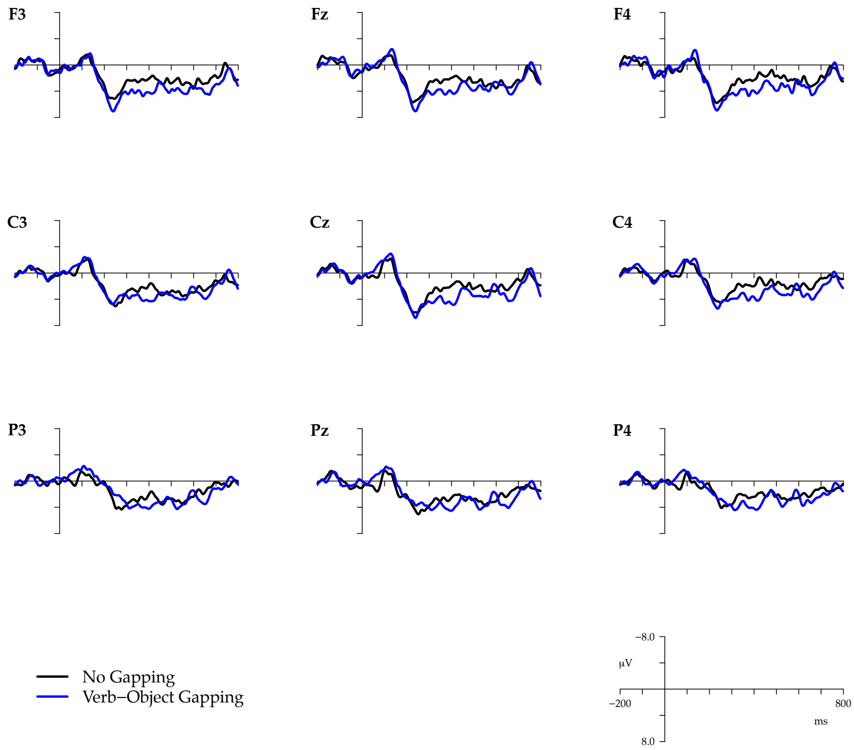

Figure 6.3: Grand averages of conditions No Gapping (***a***) and Verb-Object Gapping (***c***) at onset (y-axis) adjunct noun (*achtertuin*) at electrode sites F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. Corresponding example sentences can be found on page 116.

6.1.6 Discussion and conclusion

In this experiment, I attempted to tap into the resolution process of Gapping-like constructions while manipulating the amount of structure to be recovered. As it turned out, my planned measure point was not an appropriate critical time point. In a post hoc analysis of time points earlier in the experimental sentences, it appears that indeed effects of CONDITION could be established. Effects are apparent before the intended measure point *waren*.

Rather counter-intuitively, the recovery of the verb *onderhield* in condition ***b*** yielded a larger effect (lasting for 510 ms in total) compared to the recovery of the VP *het gazon onderhield* in condition ***c*** (lasting for 130 ms in total). Although it is possible that the sustained positivity in condition ***b*** is caused by individual variation, I will argue that it is composed of two consecutive positive deflections.

Regardless of whichever account one wants to relate the results, it was expected that the elision of a larger structure would result in a larger effect overall. With the current results we could provisionally conclude that the recovery of a VP is more effortful than to connect a fully elided VP with a remnant adjunct phrase. Thus, in the post hoc analysis it appears that it depends on the type of elided structure (quality) rather than the amount of structure (quantity). To recover a VP a predicate argument relation between elided verb and remnant object needs to be established. It could be that this relation is readily available in case a full VP is retrieved. The integration of a full VP with a remnant adjunct phrase is predicted to be relatively easy by structural accounts (including Copy α) which assume that an adjunct attaches directly to a VP node. However, it is assumed to encounter relatively more problems in searching for the exact structure in memory. A cue-based approach, that may translate different information types as cues, predicts the reverse: relative ease of retrieval, but higher integration cost as different information types need to be matched. Therefore, the two proposed mechanisms may both account for the relative ease of processing an elided VP and adjunct phrase. It is still unclear how these mechanisms may help to explain why the recovery of a VP is relatively more costly.

In both manipulated conditions, the effect of CONDITION amounted to positive deflections starting around 300 ms. Further, it seems that two consecutively positive components can be related to a relatively effortful resolution process. This contrasts with the negativity as found on the determiner (that was followed by a P600 on the subsequent noun) in the replicated study in Chapter 5.1. The first positive component found in the current experiment could be in part connected to a closure positive shift (CPS). It has been argued that the appearance of a comma triggers subvocal prosodic sentence phrasing – marking a prosodic boundary (Drury, Baum, Valeriote, & Steinhauer, 2016; Steinhauer & Friederici, 2001). Yet, the latency of the CPS may differ as a function of task or language used in the experiment (see for example Peter, McArthur, & Crain, 2014). In that sense, a negativity as found in the replication study could be outweighed by a subvocal prosodic effect, yielding a different polarity. Note that the negativity was found on a function word – a determiner – which is usually not assigned meaningful prosodic content, contrasting with remnants that survive by virtue of focus assignment. However, if it were only a reflection of subvocal prosody, we should have found a similar latency and duration of the component in all conditions. As this is not the case, I suggest that the early positivity is sustained by both 'acoustic' and linguistic cues (see Peter et al., 2014 for a discussion that both acoustic and linguistic cues are involved in the CPS in 'normal' sentences). The early positivity marks the start of the resolution process to retrieve missing information, to integrate it with the remnant structure and to arrive at an interpretation. I further suggest that the relative difficulty of integration of retrieved material is then reflected by the secondary positivity which I relate to a P600.

Besides the CPS, the early positivity could be a family member of a more domain-general P300 which has been related to surprisal effects (in so-called "oddball tasks") and context updating (see for example Donchin, 1981; Donchin & Coles, 1988). In particular, in an overview and discussion of the P300, (Polich, 2007:2130) states that the amplitude and latency of a P300 varies as a function of task demands to the extent that "for tasks that require greater amounts of attentional resources, [the] P300 amplitude is smaller and peak latency is longer". While a context updating account may be in line with a linguistic resolution process, it should be noted that a P300 may be simply reflecting attentional, evaluative, or memory mechanisms (see for a review Nieuwenhuis, Aston-Jones, & Cohen, 2005). Still, in the experimental linguistic literature, Gouvea et al. (2010) reported that an anterior positivity between 300-500 ms was elicited by a condition during which a *wh*-dependency was resolved, marking the onset of a retrieval process. These authors also suggest that "manipulations that impact the number and type of syntactic relations that are attempted should change the amplitude and/or duration of the P600" (2010:174), which may be in line with the secondary positivity as established on the object noun.

However, a caveat is in order. At the critical nouns, the clauses are "wrapped up" up to that point. This means that retrieved information is integrated with the remnants. In condition ***b*** it can be seen that there is more "new" information than in condition ***c***. This extra information may have caused a spill-over effect, increasing the demands on the integration process. While this cannot be explained by a copy account, it can be accounted for by means of cues, as new information – say, every contrasting remnant – may interfere with old information. It has been shown that a cue-based account is apt to provide an explanation for interference effects (see for example Martin & McElree, 2008). As a consequence of the difference in information load, it is difficult to compare the effect on the object noun with the effect on the adjunct noun. In that sense, the effect does depend on a spill-over effect rather than on the nature of the elided structure. In the following experiment, I control for a possible spill-over effect.

6.2 Modulation of structure in the left conjunct

6.2.1 Introduction

Still being concerned with the effect of elision of different structure size and in order to control for spill-over effect in the second conjunct, I designed a follow-up ERP experiment. In this experiment, I utilise Stripping since it will help to control for the amount of new information in the second conjunct. In Stripping, the structure to be elided can be manipulated in the left conjunct while the size of the remaining structure is constant in the right conjunct. Let us have a look at the stimuli to further explain the logic and hypotheses for the current design.

6.2.2 Methods

Test materials

On the basis on the Stripping condition of the previous experiment a stimulus set of 36 quadruplets was compiled as in (4):

(4) a. Omdat Koen een kast verving, en Judith **ook**, waren de bewoners perplex.
Since Koen a cabinet replaced and Judith too were the inhabitants perplexed
'Since Koen replaced a cabinet
and Judith too,
the inhabitants were perplexed.' (***VP Stripping***)

b. Omdat Koen een enorme kast verving, en Judith **ook**, waren de bewoners perplex.
Since Koen a huge cabinet replaced and Judith too were the inhabitants perplexed
'Since Koen replaced a huge cabinet,
and Judith too,
the inhabitants were perplexed.' (***VP-Adjective Stripping***)

c. Omdat Koen in de woonkamer een kast verving, en Judith **ook**, waren de bewoners perplex.
Since Koen in the living.room a cabinet replaced and Judith too were the inhabitants perplexed
'Since Koen replaced a cabinet in the living room,
and Judith too,
the inhabitants were perplexed.' (***VP-Adjunct Stripping***)

d. Omdat Koen in de woonkamer een enorme kast verving, en Judith **ook**, waren de bewoners perplex.
Since Koen in the living.room a huge cabinet replaced and Judith too were the inhabitants perplexed
'Since Koen replaced a huge cabinet in the living room,
and Judith too,
the inhabitants were perplexed.' (***VP-Adjective-Adjunct Stripping***)

As can be observed, a baseline was constructed in condition ***a*** in which the

VP *een kast verving* is stripped in the second conjunct and is resolved at the word *ook*, which will be the critical measure point. In the first non-baseline condition (***b***) the adjective *enorme* is added to the object noun of the VP, to expand the VP. Condition ***c*** in this set equals the Stripping condition of the experiment reported in section 6.1 and contains the VP of condition ***a*** with the addition of the adjunct *in de woonkamer*. In the last manipulation in the current set, the baseline VP has the addition of both the adjective and adjunct. Clearly, the second conjunct has the same length in all conditions and consequently, contains the same amount of new information, as there is only one contrasting element (*Judith*). This was done to overcome the suggested spill-over effects during wrap-up and integration at the end of the clause. In the current design, an increase of information load is constrained to the elided structures.

Since the measure point will be at a word that will be presented with a comma, I predict – on the basis of the findings in the previous experiment – a positive component which reflects the effect of prosody and the start of the recovery process. The amplitude and duration of the positivity is further expected to be modulated by the amount of elided structure – the least in condition ***b*** and the most in condition ***d***. I further hope to corroborate previous effects in terms of two consecutive positive components. The first component, around 300 ms after onset of *ook*, would be understood as (partly) related to processes of retrieval. A secondary positive component occurring around 600 ms after onset will be considered as a measure of integration cost. I hypothesise that, if retrieval mainly depends on searching for and finding syntactic structure, a syntax-related ERP should be found. This is predicted by Copy α. Alternatively, retrieval cost that is reflected by modulations of the first positive component may be linked to several information types – not exclusively syntax-related. Possible modulations of integration cost as reflected by the second component may be connected to the cue-based retrieval account, since a copy account predicts a relatively easy integration.

The stimuli were interspersed with 72 fillers – the same as in the previous experiment – and were again counterbalanced using a Latin Square design.

Participants

Thirty-three native Dutch participants with normal or corrected-to-normal vision took part in this study and were paid €15. Two of them were disregarded from the analysis because of technical failure and three participants were taken out because of too many artefacts. Four participants were excluded because they were left-handed. Of the remaining 24 participants that were taken into consideration for further analysis, 8 were male and the mean age was 22.30 (range 19-33). The experiment followed the Ethics Committee regulations of the Humanities Faculty of Leiden University, which approved its implementation. Participants gave informed consent before the study started.

Procedure

Participants were comfortably seated in a dimly lit sound-proof room at a distance of approximately 90 cm from a 19 inch LCD monitor. One-hundred-and-eight test sentences were randomly presented word by word in Verdana font (36pt) for 300 ms per word with a 300 ms blank screen interval using the presentation software E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA). Presentation of a trial started with a fixation cross for 1,000 ms. Every sentence was followed by a content question to ensure participants' attention. A blank screen with a duration of 1,500 ms intervened between the last word of a trial and the presentation of the comprehension question. For half the participants the left response button referred to "YES", for the other half the left button referred to "NO". Participants were given a break after every 12 sentences and could proceed at their own pace. The comprehension questions referred to different parts of the sentences equally. Before the test stimuli, the participants were able to get used to the task with four practice sentences.

The experiment was concluded with a working memory test as used in the previous experiment in this chapter.

In total, the experiment took about 1.5 hours per participant, including set-up.

Apparatus and electrophysiological recording

A description of the recording set-up can be found in chapter 5.1.1.

Data analysis

The EEG data were preprocessed using the same criteria as the previous analysis. Epochs of 1,000 ms were computed with a 200 ms pre-stimulus baseline and ERP grand averages were time-locked to the onset of the target word *ook*. 10.53% of the trials were excluded from the analysis: of the presented nine trials per condition an average of 8.05 ($SE = 0.13$) were used.

The repeated measures ANOVA on the ERPs had as independent factor CONDITION consisting of four levels (A, B, C, D) of which the first is understood as baseline. As in the other experiments, three levels of the factor ANTERIORITY (frontal, central, posterior) and two levels of the factor HEMISHERE (left, right) were taken into consideration.

6.2.3 Behavioural results

On average, the accuracy on the comprehension questions was 92.25% ($SE = 1.57\%$). Per condition the accuracy scores were: $M_{\text{VP Stripping}} = 92.60\%$, $M_{\text{VP-Adjective Stripping}} = 93.52\%$, $M_{\text{VP-Adjunct Stripping}} = 90.74\%$, and $M_{\text{VP-Adjective-Adjunct Stripping}} = 92.13\%$. The accuracy scores did not differ significantly between conditions as shown by a repeated measures ANOVA on the scores [$F(3, 69) = 0.42, p = .660, \eta^2_G = .004$].

The accuracy of the three test sessions of the working memory task was on average 68.61% ($SE = 2.45\%$). Per condition the scores were: $M_{\text{Random Counting}} = 62.50\%$, $M_{\text{Auditory Presentation}} = 59.17\%$, $M_{\text{Visual Presentation}} = 84.17\%$. A repeated measures ANOVA by subjects with CONDITION as independent factor and ACCURACY OF NUMBER RECALL as dependent variable showed that the scores differed between conditions [$F(2, 46) = 9.44, p < .001, \eta^2_G = .202$]. A post-hoc comparison with Bonferroni correction showed that the visual condition differed significantly from the random condition [$p = .014$] and auditory condition [$p < .001$].

Again, the scores of the comprehension task of the ERP experiment were compared with the scores of the working memory task. A small, non-significant correlation was found between the variables ACCURACY OF SENTENCE COMPREHENSION and ACCURACY OF NUMBER RECALL [$r = .213$, $p = .318$].

6.2.4 Electrophysiological results

In Figures 6.4 (on page 125), 6.5 (on page 126), 6.6 (on page 127), the grand averages of conditions ***b***, ***c***, ***d***, respectively are depicted relative to the baseline (condition ***a***). After evaluation of the graphs and a permutation test on all electrodes per sample with independent factor CONDITION, two time windows of interest were determined: 300-480 ms and 530-790 ms. In order to check the effects of the factor CONDITION in the intermediate time window 480-530 ms, a repeated measures ANOVA was run. No effects could be found on midline sites [$F(1, 23) = 0.87, p = .462, \eta^2_G = .017$] and on lateral sites [$F(1, 23) = 2.03, p = .120, \eta^2_G = .032$]).

Effects in the 300-480 ms time window

On midline sites, the factor CONDITION reached marginal significance [$F(3, 69) = 2.41, p = .074, \eta^2_G = .040$], while also a marginally significant interaction of CONDITION by ANTERIORITY was found [$F(6, 138) = 2.42$, $p = .059, \eta^2_G = .016$]. For convenience, the means of voltages per condition grouped by anteriority are shown in Figure 6.7 (on page 128).

On lateral sites, the factor CONDITION reached significance [$F(3, 69) = 4.82, p = .004, \eta^2_G = .050$]. A pairwise comparison with Bonferroni correction showed that both conditions ***a*** and ***b*** differed significantly from condition ***c*** and condition ***d***. Table 6.1 (on page 124) summarises these comparisons together with the values for the means, standard errors and p-values. While no interactions could be established, a marginal effect of HEMISPHERE was noticed [$F(3, 69) = 3.89, p = .061, \eta^2_G = .009$], as the left-localised electrodes were slightly more positive [$M_{\text{Left}} = 1.39$ ($SE = 0.20$), $M_{\text{Right}} = 0.81$ ($SE = 0.21$)].

Effects in the 530-790 ms time window

While the effect of CONDITION was marginally significant on midline electrodes [$F(3, 69) = 2.73$, $p = .051$, $\eta^2_G = .053$], a significant effect of ANTERIORITY was found [$F(2, 46) = 3.97$, $p = .036$, $\eta^2_G = .015$]. A pairwise comparison with Bonferroni correction showed that the average amplitude of electrode Pz was significantly more positive than electrode Fz [$p = .05$, $M_{Fz} = 0.39$ ($SE = 0.49$), $M_{Pz} = 1.66$ ($SE = 0.41$)].

On lateral sites the effect of CONDITION was significant [$F(3, 69) = 4.61$, $p = .005$, $\eta^2_G = .067$]. A pairwise comparison with Bonferroni correction showed that condition ***a*** differed significantly from condition ***c*** and condition ***d***. Condition ***b*** differed from condition ***c*** with marginal significance and differed significantly from condition ***d***. Finally, condition ***c*** differed significantly from condition ***d***. Table 6.2 (on page 124) summarises the means and standard errors of the four conditions and the *p*-values of the multiple comparisons that were used to determine differences between all conditions at lateral sites.

			Comparison (*p*-value)			
Condition	Mean (μV)	*SE*	***a***	***b***	***c***	***d***
a	0.41	0.32	-	1	.006	$<$ .001
b	0.32	0.30		-	.004	$<$ .001
c	1.44	0.27			-	.164
d	2.39	0.28				-

Table 6.1: Means and standard errors of the amplitudes and *p*-values of the Bonferroni-corrected multiple comparisons of the four test conditions in the 300-480 ms time window at lateral sites.

			Comparison (*p*-value)			
Condition	Mean (μV)	*SE*	***a***	***b***	***c***	***d***
a	-0.59	0.32	-	1	.027	$<$ .001
b	-0.58	0.34		-	.062	$<$ .001
c	0.47	0.21			-	.014
d	1.75	0.32				-

Table 6.2: Means and standard errors of the amplitudes and *p*-values of the Bonferroni-corrected multiple comparisons of the four test conditions in the 530-790 ms time window at lateral sites.

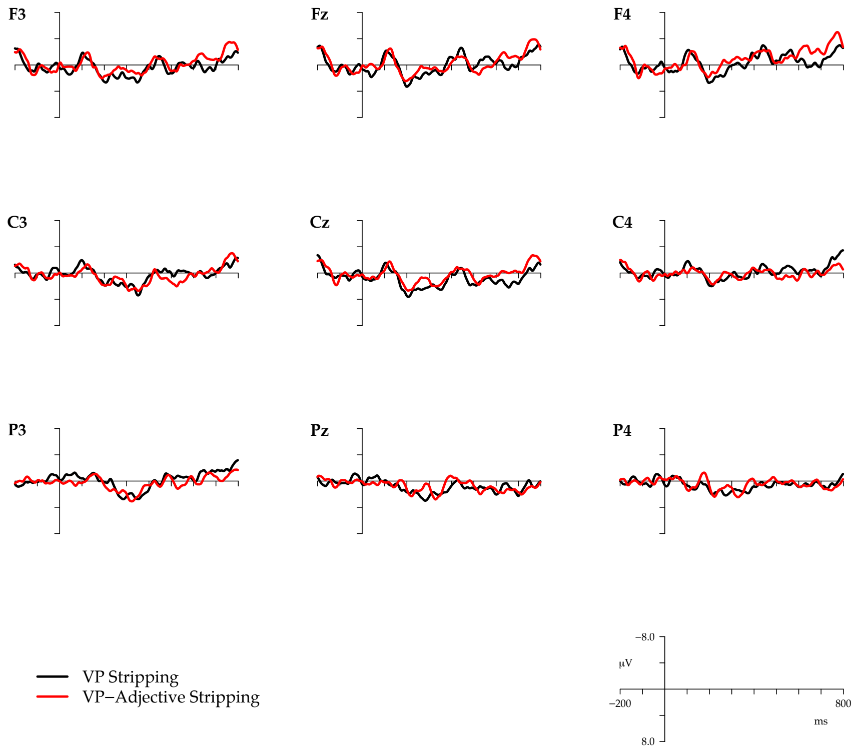

Figure 6.4: Grand averages of VP Stripping condition (***a***) and VP-Adjective Stripping condition (***b***) at onset *ook* at electrode sites F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. Corresponding example sentences can be found on page 120.

6.2.5 Discussion

In the current experiment, no difference in comprehension accuracy between conditions could be established, while at the same time the factor CONDITION related to significant effects in the electrophysiological data. In that sense, this is an example of diverging results of different measurement types as discussed in Chapter 4.1.3. Based only on offline behavioural data, I would have concluded that interpretation of Stripping does not depend on the amount of structure to be retrieved and integrated, which is in line with a cost-free resolution mechanism, as is predicted by both a Copy α and cue-based accounts. However, the online data indicate that the parser does require additional ef-

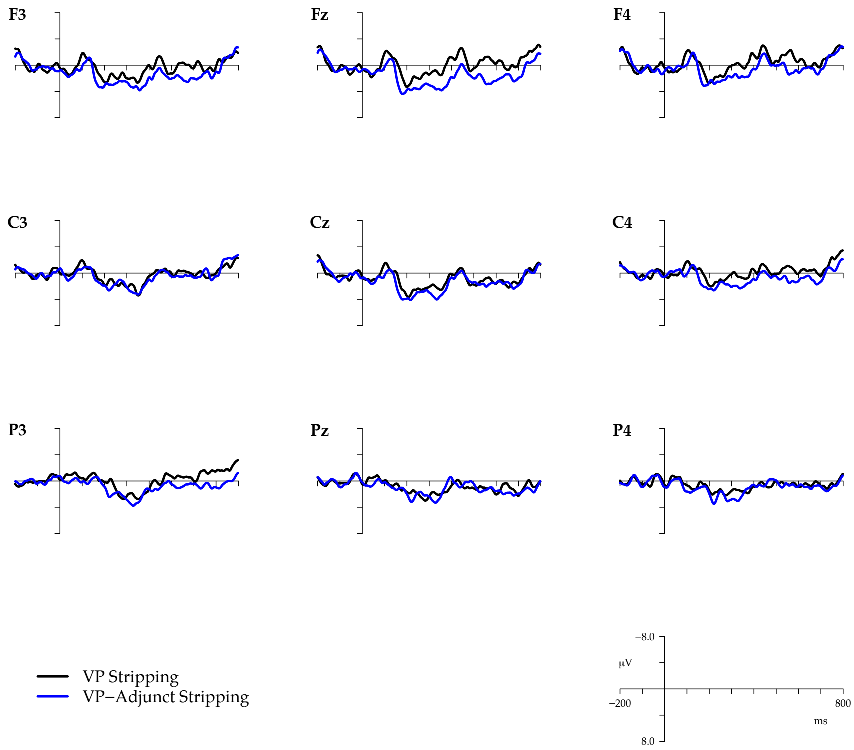

Figure 6.5: Grand averages of VP Stripping condition (***a***) and VP-Adjunct Stripping (***c***) at onset *ook* at electrode sites F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. Corresponding example sentences can be found on page 120.

fort as a function of the size of the structure to be retrieved and integrated. As was suggested in the post-hoc analysis of the experiment described in section 6.1.5, the pattern of two consecutive positive deflections may be interpreted as reflecting a two-stage resolution mechanism.

In both time windows, the factor CONDITION yielded significant effects, most prominently at lateral sites. Although the interaction of CONDITION with ANTERIORITY was only slight, it appeared that in the first time window frontally distributed electrodes showed relatively more positive amplitudes in conditions in which the adjunct needs to be retrieved (conditions ***c***-***d***). Because of this frontal distribution, it is tempting to connect the positivity to a P300, in particular the so-called P3a, that is normally related to general frontal lobe attention mechanisms possibly subserving attentive language comprehension, since the manipulation in question is linguistic. Although only the behavi-

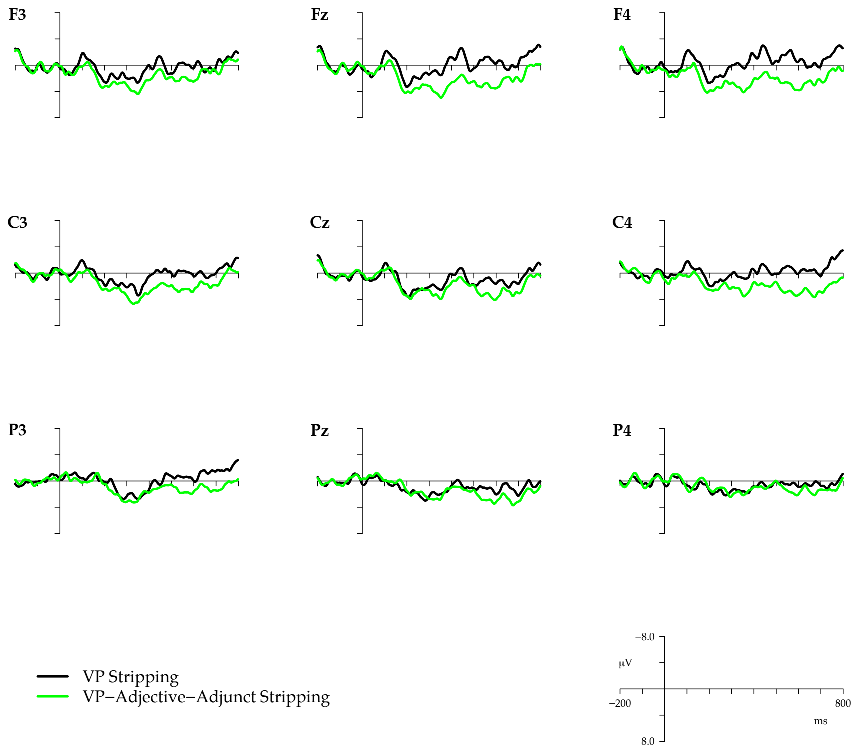

Figure 6.6: Grand averages of VP Stripping condition (***a***) and VP-Adjective-Adjunct Stripping (***d***) at onset *ook* at electrode sites F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. Corresponding example sentences can be found on page 120.

oural data of the the working memory was analysed, it does not seem to be the case that general working memory plays a large role. On the one hand, the correlation between comprehension accuracy and scores on the working memory task was not significant. On the other hand, P300 effects that are related to working memory tend to correlate with activity in parietal brain areas (Polich, 2007).

Following Friederici (2002)'s model of auditory sentence comprehension, the frontal lobe can be implicated in working memory processes, but these are not reflected by positive deflections. An ELAN effect is connected with memory of syntactic structure, while semantic features are usually reflected by negative deflections between 300-500 ms. All of these negativities are assumed to originate from frontal areas. In general, the working memory data in this experiment follow the trend as seen in earlier tests: the visual condition is

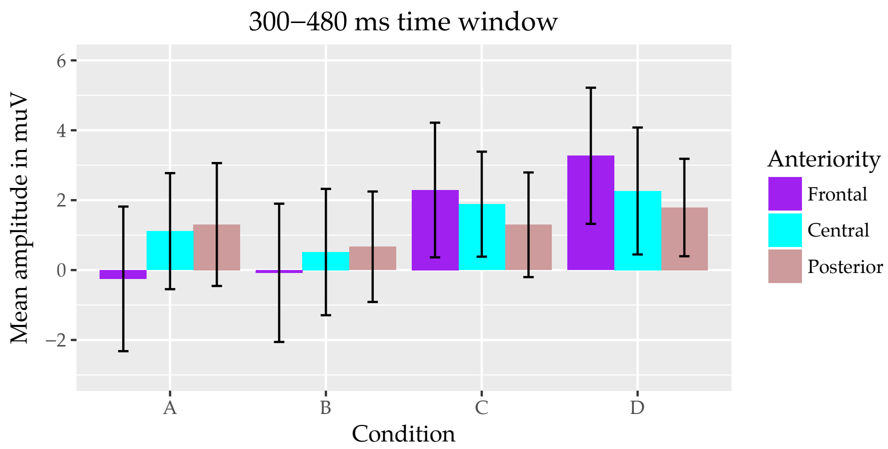

Figure 6.7: Error bar chart of the the means of voltages per condition at onset *ook* in time window 300-480 ms, grouped by anteriority on midline electrodes.

performed best, followed by random counting; the lowest scores are seen in the auditory task; the scores on the memory task barely correlate with the sentence comprehension ratios. All in all, it seems that working memory as assessed does not play a large role.

Since the critical word appeared with a comma, some part of the first positive deflection should be expected to be related to prosodic phrasing which can be measured as a CPS component (Closure Positive Shift). This seems on a par with Drury et al. (2016)'s study that reports comma effects starting at 300 ms – particularly at frontal electrodes – after onset of the critical word in sentences such as in (5).

(5) a. John said Mary was the nicest [...].
b. John, said Mary, was the nicest [...].

Typical CPS effects were shown at proper names "John," and "Mary," (presented with comma) when compared to their comma-less counterparts. However, as noted in section 6.1.6, a CPS may be sustained by acoustic as well as linguistic cues, which might explain why it can be modulated by adding structure that needs to be retrieved – as was done in the current experiment. Thus, the retrieval of the antecedent partly depends on prosodic, and possibly, syntactic and semantic information types.

It remains difficult to understand the exact nature of linguistic information types that are retrieved. That is, the relative import of these types is still unclear. Before I started the ERP experiments and in line with Friederici (2002) mentioned above, I expected to find effects related to the retrieval of a fully-fledged syntactic structure early on, possibly in the form of an ELAN component. However, again, I cannot be sure whether such an effect was apparent since it might have been cancelled out by a CPS and, possibly, other P300-related processes. Therefore, associating the early positivity directly with a Copy α mechanism would be tenuous.

If an early syntax-related component is not found, this does not necessarily entail that no structural information is retrieved. With respect to linguistic information types during retrieval, the early positive component could be a member of the P600 family. It has been argued that the latency of a P600 may vary as a function of ease of retrieval. For example, (Gouvea et al., 2010:175) claim that P600 may occur quite early if "retrieval processes that are needed to initiate structure building" – such as during the resolution of *wh*-dependencies – is relatively easy. Interestingly, such dependencies elicited a positivity between 300-500 ms that had initially a more anterior scalp distribution before shifting to a more 'standard' posterior distribution reflecting integration processes. Yet, the retrieval of an object *wh*-phrase is qualitatively different from retrieving elided structure. A *wh*-phrase is kept in memory pending integration after which it can be fully interpreted, whereas an elided phrase needs to be searched in memory to be recovered as it has already been fully interpreted. In other words, contrasting with ellipsis resolution, if a listener encounters a *wh*-phrase, the parser 'knows' that it will be needed later on.

Recall that Kaan et al. (2004) also found a fronto-central positivity between 300-500 ms during the resolution of Gapping in English, which they cautiously related to the retrieval of the preceding verb information, without being clear as to the specifics of this information. The retrieval processes converge as to their timing and although syntactic information may be necessary to proceed to the integration phase, it does not necessarily mean that this information amounts to a fully-fledged structural representation at the point of retrieval.

Given that we may regard the early positivity to be connected to a P600, it is notable that a P600 has been related to both syntactic processes and semantic processes, ranging from processes of syntactic revision and repair (for example Friederici et al., 2002), syntactic integration (for example Kaan et al., 2000) to semantically related mechanisms (see for example Burkhardt, 2007). It has been proposed by Kaan and Swaab (2003) that late frontal positivities are triggered by "discourse complexity" whereas posterior positivities relate to the repair of ungrammatical structure. While an interaction between CONDITION and ANTERIORITY was statistically – albeit marginally – demonstrated in the first time window on midline electrodes, in the second time window no such interaction was apparent. Possibly, prompted by a prosodic break, "discourse complexity" was observed by the processor already early on. That is, a more fully interpreted chunk is analysed upon retrieval.

While the early component seems to be an amalgam of several (domain-general and linguistic) processes related to prosodic phrasing, attention and retrieval, the positive component as found in the second time window was predicted to the extent that the additional retrieved information would have an impact on the integration part of the process – if no complete syntactic structure was retrieved. Notably, a gradual difference between conditions in later time windows was most significant (see Tables 6.1 and 6.2 for a comparison of numbers). In line with Kaan et al. (2013)'s result, I regard these effects as related to the P600 family involved in integration processes. Since the retrieved chunks of interpreted structure consist of different information types we may conclude that the integration with the remnant structure is relatively costly.

All in all, it seems that ellipsis resolution follows an inference-by-inference parsing mechanism, during which structure is built step-by-step and processed information is integrated. Notably, as I mentioned in Chapter 3.3, Frazier and Clifton (2001) suggest that Gapping resolution might follow such a mechanism. While this incremental procedure is widely acknowledged for sentence processing (as noted in the first chapter, Section 1.3.2), a main locus of integration processes in a relatively late time window is assumed to be reflected by late positivities. In 'normal' sentences, integration hinges on retrieval of incoming words. Difficulty of retrieval of words and semantic complexity can be demonstrated by a negative deflection around 400 ms after onset (i.e. N400). This mechanism forms the basis of a recent and computationally sustained model put forward by (Brouwer & Hoeks, 2013)'s "Retrieval-Integration" account. During ellipsis processing, the retrieval of lexical items is 'replaced' by the retrieval of missing linguistic structure.

Since a possible antecedent for ellipsis is encoded in memory moments before the elision, it is conceivable that the retrieval process of the antecedent starts slightly faster than lexical retrieval. Possibly, and on a par with lexical retrieval, during retrieval of elided structure a semantic representation is targeted at first instance, but it may be steered by additional information types. As with lexical retrieval, the amplitude of the component related to retrieval is modulated by the relative complexity of "search-and-find". During the integration phase several levels are being linked. Bearing in mind that integration processes were reflected as a broadly distributed late positivity, it may be concluded that several processes work in parallel, but also, that it remains unclear what the relative weight is of each of these processes.

By and large, Copy α seems untenable to account for Gapping and Stripping. However, it is still an open question to what extent structural information is analysed. For example, it could be that during integration (part of) this information is "shared" (Frazier & Clifton, 2005) or "recycled" (Arregui et al., 2006; Frazier, 2013). At the same time, the results are not wholly compatible with the predictions of the cue-based account as proposed by Martin and McElree (2008). However, we could assume that the retrieval stage is affected by the propositional content as generated by additional structure. While a clear and detailed description of cues is still lacking, in principle, different information types may all potentially be described in terms of cues. Since a cue-based account refers to more fully interpreted structure, I suggest that some cue directly points to propositions. As I have argued in Chapter 2.4.2, in the case that an adjunct is added, a secondary *proposition* becomes available, which possibly induces an additional retrieval step. Considering it to be a form of "cue overload", the additional propositional content may give rise to a discourse complexity that can be measured as a frontally distributed positivity. Coherent with an inference-by-inference parsing mechanism, after retrieval of a set of cues, the integration stage applies during which different levels of analysis are unified.

In a nutshell, although processes of retrieval and integration are visible as positive deflections, they remain unnoticed in terms of a behavioural measure of the kind I used in this experiment.

6.3 Conclusion

In two experiments, I investigated the role of the size of structure in Gapping and Stripping constructions. While the design of the first experiment was not optimal, I was able to interpret the collected data in a post-hoc manner, being very cautious as to the conclusions I drew. Since a comma was present at the critical measure point in the alternative analysis, a major role of prosodic phrasing was apparent as reflected by a CPS. Furthermore, the amount of new information in the right conjunct – the clause in which the amount of structure was modulated – differed between conditions. Therefore, it was not possible to make a straightforward comparison between conditions and a follow-up experiment was proposed. Provisionally, I concluded that the recovery of elided structure starts at around 300 ms after onset of the critical word and is reflected by positive deflections; furthermore that a secondary positive component can be demonstrated signalling more complex integration processes.

In the second experiment, I controlled for the amount of new information in the right conjunct and manipulated the amount of structure to be deleted in the left conjunct. Again, the analysis was influenced by a comma that was presented with the critical word and a CPS-like component was found. However, since its amplitude was modulated by the amount of structure to be retrieved I suggested that other processes, linguistic as well as domain-general (attentional, evaluative, or memory mechanisms), are at work too. Additionally, a secondary positive deflection was apparent in conditions with relatively large deletions. In line with the findings of the first experiment as well as the replicated study in Chapter 5.1, I suggested that the secondary (late) positivity relates to a P600 and reflects the relative difficulty of integration of additional propositional content.

The main goal of the experiments was to differentiate between two models of ellipsis resolution that I discussed in Chapter 3.3 and which I suggested predictions of in Chapter (1). I hypothesised that under a Copy α account, manipulation of gapped structure size would be reflected by early ERPs related to retrieval of a fully-fledged syntactic structure. Alternatively, I suggested that a cue-based account predicts relative ease of retrieval but a relatively more costly integration process. With the current data, I was able to show that both retrieval and integration processes may be affected by modulation of structure. However, it does not seem to be the case that retrieval processes are only driven by syntax-related mechanisms. I suggest that at least (subvocalic) prosodic phrasing and processes related to attention are involved. I have further argued that a discourse complexity might be noticed as early as the retrieval starts, that is, retrieval processes target more fully interpreted chunks. In the next chapter I further investigate the import of semantic information types during retrieval and integration.

As it stands, it seems that a cue-based account should be able to account for the findings in this chapter, though a clear description of how exactly a cue should be formulated is still unclear. What is clear is that the current data are supported by theoretical accounts that acknowledge the multidimensional character of ellipsis resolution, which I have discussed at length in Chapter 2.4. Future research is needed to investigate how theoretical constructs may feed into a processing account based on cues.