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

Ruijgrok, B.J.

Author: Ruijgrok, Bobby
T**itle:** Tapping into semantic recovery : an event-related potential study on the processing of gapping and stripping
Date: 2018-05-31

CHAPTER 5

Replication and norming stimuli

Essential to the process of science is the replication of previous studies in order to validate existing findings before building on them. As has become clear in Chapter 3.5, electrophysiological research on ellipsis, let alone Gapping in Dutch, is scarce. Section 5.1 reports the findings of a replication of Kaan et al. (2013). I thank Wouter Broos for his assistance with organising the stimuli and recording of the EEG data. In sections 5.2 and 5.3 I report norming studies that I carried out to pretest newly designed stimuli to be used in subsequent experiments.

5.1 Validating ERP results: a replication study

5.1.1 Method

Test materials

Using a Latin Square design, 117 quadruplets as described in chapter 3.5 were divided over four lists and complemented with 96 fillers.[1] The experimental paradigm is illustrated again in (1). The four stimulus conditions and additionally the two collapsed conditions (No-Gapping vs. Gapping) are colour-coded, corresponding to colours used in graphs later on.

(1) a. Anouk zond de kaart aan haar vader, en Julia **de** **bloemen** aan
Anouk sent the card to her father, and Julia the flowers to
haar moeder.
her mother.
'Anouk sent the card to her father,
and Julia the flowers to her mother.' (***Plausible Gapping***)

b. Anouk schreef de kaart aan haar vader, en Julia **de** **bloemen**
Anouk wrote the card to her father, and Julia the flowers
aan haar moeder
to her mother.
'Anouk wrote the card to her father,
and Julia the flowers to her mother.' (***Implausible Gapping***)

c. Anouk zond de kaart aan haar vader, terwijl Julia **de** **bloemen**
Anouk sent the card to her father, while Julia the flowers
aan haar moeder stuurde.
to her mother shipped.
'Anouk sent the card to her father,
while Julia shipped the flowers to her mother.' (***Control for condition a***)

d. Anouk schreef de kaart aan haar vader, terwijl Julia **de** **bloemen**
Anouk wrote the card to her father, while Julia the flowers
aan haar moeder stuurde.
to her mother shipped.
'Anouk wrote the card to her father,
while Julia shipped the flowers to her mother.' (***Control for condition b***)

[Kaan et al. (2013)]

[1] The odd number has to do with the fact that, in the original experiment, from the original set of 120 items three had been omitted due to an experimenter error.

To recapitulate, in Kaan et al. (2013), at the determiner, a LAN effect was expected for Gapping conditions ***a-b*** versus No-Gapping conditions ***c-d***, but this was only apparent in a group of participants who scored relatively poorly at the end of sentence task. At the noun following the determiner, an N400 for (***b*** versus ***a***), perhaps followed by a P600, was predicted. Only a P600 effect turned out to be significant; an N400 was arguably detected, but only as a numerical trend. Finally, the authors hypothesised that if syntactic integration is more effortful in Gapping versus No-Gapping constructions, a P600 effect for Gapping versus No-Gapping constructions at the noun should be found. A notable result of the original study is that this effect was apparent for plausible conditions ***a*** versus ***c***.

Participants

Twenty-two native speakers of Dutch with normal or corrected-to-normal vision participated. All participants reported not to have any neurological problems or disease. Due to bad signal (3 participants) and left-handedness (1 participant), four participants were discarded and the analysis below is based on 18 right-handed participants (16 women, 2 men, $M_{\text{Age}} = 23.17$, range 20-27). Participants gave informed consent before the study and were paid €15. The experiment complied with the Ethics Committee regulations of the Faculty of Social Sciences of Leiden University, which approved its implementation.

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. Sentences were presented one word at a time in white letters in Verdana font (18pt) on a black screen using the presentation software E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA). Each sentence was preceded by a fixation cross ("+") which appeared at the centre of the screen and remained there for 1,000 ms. Then, each word was presented for 300 ms, followed by a 200 ms blank screen. A word before a comma was presented with that comma appended; similarly, the last word of each sentence was marked with a full stop. 1,500 ms after offset of the sentence-final word a prompt, *OK of SLECHT* ("OK or BAD"), appeared. The left response button was linked to *OK* and the right one to *SLECHT*. As a means of counterbalancing, half of the participants received the prompt and button choices the other way around. After a response click, a blank screen appeared for 1,000 ms. After every 12 sentences, participants were offered a break. Before starting the experimental phase, six warm-up practice trials were presented to the participants. These sentences bore no resemblance to any of the experimental or filler items.

In addition to the experiment reported above, a working memory test based on a task described in chapter 4.2.3 was carried out. Participants were instructed to count aloud from 1 to 10 at the rate of approximately one digit per second (5 trials). In the second session, they were asked to randomly count aloud numbers from 1 to 10 while monitoring that every number was only mentioned once in each trial: they were not allowed to repeat the same number more than once until all 10 numbers were reported. During the third session participants listened to a random sequence of nine digits between 1 and 10, after which they were asked to say which digit between 1 and 10 had been omitted (5 trials). The last session was as the third session but instead the numbers were presented visually one by one. The working memory test was carried out after the EEG recordings.

The experiment took about 2 hours per participant in total, including set-up.

Apparatus and electrophysiological recording

The electroencephalogram (EEG) was obtained using BioSemi ActiveTwo system (BioSemi B.V., Amsterdam) following the international 10/20 system (originally proposed by Jasper in 1958 and, after modifications, standardised as of 1991 by the American Electroencephalographic Society) Ag/AgCl electrodes (Fp1/2, FC5/6, AF3/4, Fz, CP5/6, CP1/2, Cz, F7/8, F3/4, T7/8, C3/4, Pz, FC1/2, P3/4, O1/2, Oz, P7/8, PO3/4). Four flat electrodes were used to monitor the eye movements (i.e. to obtain an electro-oculogram or EOG): two above and underneath the left eye to measure blinks; two at the external canthi of both eyes to measure saccades. A flat electrode was placed on each mastoid to be used for off-line re-referencing. The EEG signal was recorded using the BioSemi ActiView software at a sampling rate of 512 Hz. Electrode impedance was monitored during installation and running to ensure a low level of noise.

Data analysis

Using Brain Vision Analyzer Version 2.0 (Brain Products, Munich, Germany), the EEG data were preprocessed before analysis to reduce noise and artifacts as much as possible. EOG artifacts were corrected using the Gratton, Coles, and Donchin (1983) algorithm. Remaining artifacts were rejected and checked visually on the basis of the following criteria: the maximum allowed voltage step was 20 μV/ms, the maximal allowed difference of values was 100 μV in an interval of 200 ms and the lowest allowed activity was 0.5μV. Just as in the original study, a low cutoff filter of 0.16 Hz, 24 bB/oct and a high cutoff filter of 30 Hz, 24 dB/oct were applied. Epochs of 1,300 ms were computed with a 100 ms pre-stimulus baseline. ERP grand averages were time-locked to (*i*) the critical determiner following the position of the elided verb (average percentage rejected: 24.41% of the trials for Gapping and 25.74% No-Gapping conditions) and (*ii*) the noun following the determiner (average percentage rejected: 24.22% for Gapping and 24.60% for No-Gapping).

Again in accordance with the original study, the effect of Gapping versus No-Gapping at the determiner was analysed using the mean amplitude in the

100-200 ms (ELAN) and 400-600 ms (LAN) time windows. An additional time window of 200-400 ms was taken into account. At the following noun, the mean amplitude in the 300-500 ms (N400), 500-700 ms, 700-900 ms (P600), and 900-1,200 ms time windows (late positivity) were analysed.

Analyses were conducted separately for midline sites (Fz, Cz, Pz) and for the lateral electrode regions: left/right frontal (Fp1/2, AF3/4, F7/8, F3/4), left/right central (FC5/6, T7/8, C3/4, CP5/6), left/right parietal (P7/8, P3/4,PO3/4, O1/2). For each time window, a repeated measures analysis was carried out with as within-subjects factors GAPPING, ANTERIORITY (3 levels), and, for analyses involving lateral sites, HEMISPHERE (2 levels). Additionally, for the epochs of the noun position, PLAUSIBILITY of the verb in the first clause and object in the second. Mean voltage-amplitude was considered as the dependent variable in the analysis, and p-values where corrected for sphericity where required.

Throughout this thesis, both the behavioural data and the electrophysiological data were analysed using **R** version 3.3.3 (R Development Core Team, 2008). As can be seen above, I use small capitals to indicate factors (variables). Scripts and data can be found at http://bobbyruijgrok.com/data.

5.1.2 Behavioural results

Average accuracy rates of the acceptabilty judgements were high and no participants were rejected on the basis of accuracy ($M = 87.45\%$, $SE = 0.96\%$). The accuracy scores were similar across conditions ($M_{\text{Plausible Gapping}} = 88.82\%$, $M_{\text{Implausible Gapping}} = 87.70\%$, $M_{\text{Plausible control for a}} = 86.70\%$, $M_{\text{Control for b}} = 86.21\%$). The difference in mean values was not significant as shown by a repeated-measures ANOVA by participants with CONDITION as independent factor and ACCURACY OF SENTENCE COMPREHENSION as dependent variable [$F(3, 51) = 0.32, p = .808$, $\eta^2_G = .010$].[2]

Although the working memory task consisted of four sessions, I will only report the findings of the last three: the first session was meant as a control condition as to adjust the participant's speed of production to one digit per second approximately. Errors were defined as follows: a repetition or an omission of a number in a trial of the self-oredered condition, or an incorrect response in a trial in the auditory and visual conditions. The accuracy ratio of the three test sessions was 67.04% ($SE = 2.87\%$). Per condition the scores were: $M_{\text{Random Counting}} = 66.67\%$, $M_{\text{Auditory Presentation}} = 58.89\%$, $M_{\text{Visual Presentation}} = 75.56\%$. Although numerically the difference between the auditory and visual conditions seemed large, a repeated measures ANOVA by subjects with CONDITION as independent factor and ACCURACY OF NUMBER RECALL as dependent variable yielded only marginal significance [F(2,

[2] Throughout this thesis, in reporting repeated measures ANOVAs I report the generalized eta squared as is proposed by Bakeman (2005) as useful statistic: .02 = small, .13 = medium and .26 = large.

34) = 2.72, p = .080, η^2_G = .084].

The scores of the comprehension task of the ERP experiment were compared with the scores of the working memory task. A slight correlation was found between the variables but this was not statistically significant ACCURACY OF SENTENCE COMPREHENSION and ACCURACY OF NUMBER RECALL [r = .389, p = .110].

5.1.3 Electrophysiological results

ERPs at the determiner

Figure 5.1 shows the ERPs for the Gapping and No-Gapping conditions (i.e. collapsed over plausibility conditions: ***a-b*** and ***c-d***) at the moment the critical determiner was displayed. Relative to No-Gapping conditions a negativity can be observed in the Gapping conditions starting just after 200 ms at all electrodes.

On midline electrodes, the factor GAPPING reached marginal significance in the time window 200-400 ms post-onset [$F(1, 17)$ = 3.44, p = .081, η^2_G = .022]. No other effects could be established.

On lateral electrodes, the factor GAPPING reached significance in the 200-400 ms time window [$F(1, 17)$ = 5.33, p = .034, η^2_G = .018] as well as the 400-600 ms time window [$F(1, 17)$ = 6.01, p = .025, η^2_G = .023]. In the 100-200 ms time window the factor HEMISPHERE yielded significant effects, the left-lateralised electrodes having more negative averaged amplitudes [$F(1, 17)$ = 11.22, p = .004, η^2_G = .042]. Significant effects of HEMISPHERE coincided with significant interaction effects of ANTERIORITY by HEMISPHERE in the 200-400 ms [$F(2, 34)$ = 5.07, p = .012, η^2_G = .007] and 400-600 ms time window [$F(2, 34)$ = 4.23, p = .023, η^2_G = .005]. The interaction effects are visualised in Figure 5.2. As can be seen, left central electrodes show relatively negative mean amplitudes.

To investigate whether the overall effect was attenuated by individual variation, the mean differences in amplitude in all three time windows between the Gapping and No-Gapping conditions, collapsed over the left anterior electrodes (Fp1, AF3, F7, F3), were analysed with respect to (*i*) sentence judgement accuracy of the experimental items and (*ii*) the accuracy of the working memory task. No significant correlations could be established.

ERPs at the noun

Effects of semantic integration between the noun and the elided verb were first analysed in relation to the factor PLAUSIBILITY. ERPs at the critical noun are displayed in Figure 5.3 for the Plausible Gapping and Implausible Gapping conditions (***a*** and ***c***).

While a negative deflection can be observed at around 400 ms, no significant effect of PLAUSIBILITY could be established in the 300-500 ms time win-

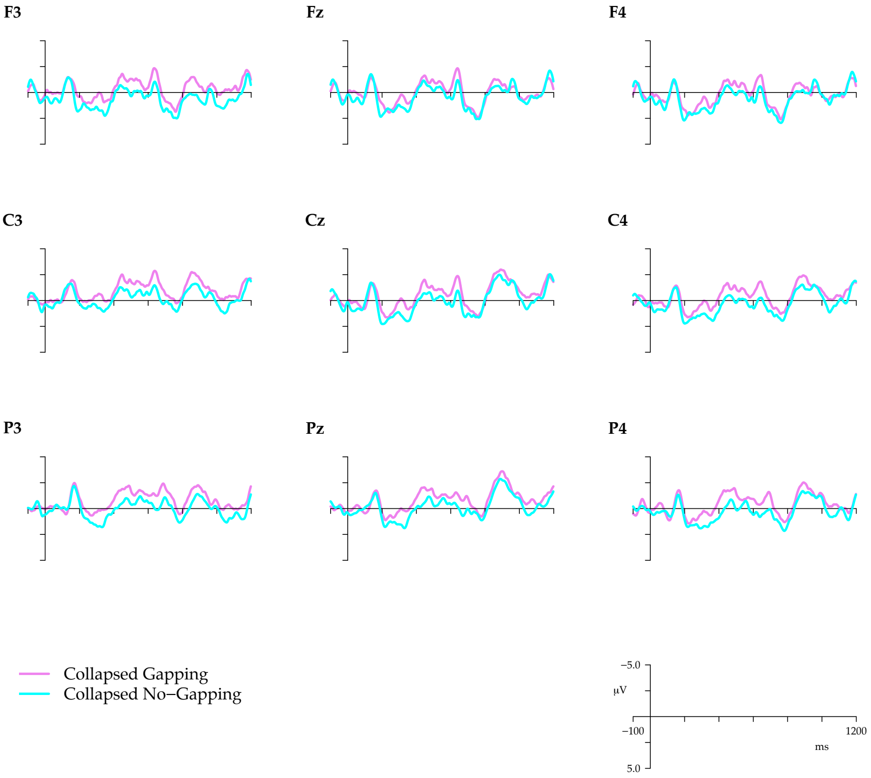

Figure 5.1: Grand averages of collapsed Gapping conditions (***a*** and ***b***) compared to No-Gapping conditions (***c*** and ***d***) at onset (y-axis) of the determiner (*de*) at electrode sites F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. Corresponding example sentences can be found on page 86.

dow. However, on midline electrodes, an effect of ANTERIORITY was apparent [$F(2, 34) = 8.02, p = .007, \eta^2_G = .063$]. In the same time window on lateral sites an effect of HEMISPHERE could be observed [$F(1, 17) = 10.57, p = .005, \eta^2_G = .029$]. These effects were due to relatively negative amplitudes at right-lateralised centro-parietal sites.

In Figure 5.3 a late positivity for Implausible Gapping can be observed most prominently at electrode Pz. While no significant effects for the factor PLAUSIBILITY were found in later time windows (after 500 ms), on midline sites an effect of ANTERIORITY was established in the 500-700 ms time window [$F(2, 34) = 5.12, p = .022, \eta^2_G = .023$] and 700-900 ms window [$F(2, 34) = 8.68, p = .007, \eta^2_G = .004$]. Again, these effects were due to relative negative amplitudes at centro-parietal sites. In the 700-900 ms window on lateral

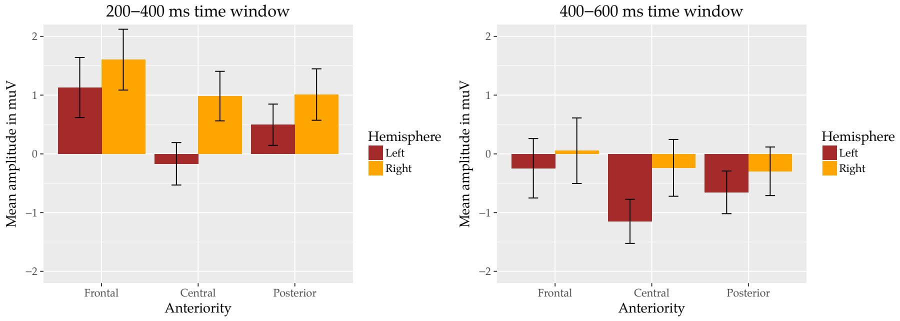

Figure 5.2: Error bar graphs of interaction effects of Anteriority by Hemisphere at the determiner (*de*) on lateral electrodes in 200-400 ms and 400-600 ms time windows.

sites, an additional interaction effect of Plausibility by Hemisphere was found [$F(1,17) = 4.68, p = .045, \eta^2_G = .002$]. Figure 5.4 shows that implausible items caused counter effects on the mean amplitude in relation to left and right electrodes, the left hemisphere being implicated in relatively large negativity.

To further analyse integration effects of the elided verb at the position of the noun, the factor Gapping was taken into account. In Figure 5.5, the difference between Plausible Gapping and Plausible No-Gapping conditions (***a*** and ***c***) are displayed. Relative to No-Gapping a large positive deflection can be observed for the Gapping condition.

On midline electrodes, an effect of Gapping was found in the 700-900 ms window [$F(1,17) = 6.56, p = .020, \eta^2_G = .037$] and in the 900-1,200 ms window [$F(1,17) = 6.40, p = .022, \eta^2_G = .034$].

No effect of Gapping could be established on lateral sites.

5.1.4 Discussion

In contrast to the original study, a negativity could be demonstrated at the determiner as the ERPs show an (E)LAN-like effect. This was hypothesised as a possible outcome. The interaction of Anteriority by Hemisphere in later time windows can be explained by the relative negative amplitudes at central sites orientated at the left. In that sense, the negative component has a relatively central distribution in this study. Considering that the factor Gapping was most prominent in the 200-400 ms and 400-600 ms time windows, the component looks like a LAN rather than an ELAN. Crucially, the effect of Gapping was not attenuated by individual variation, yet might indeed be considered as indexing prediction processes (as was suggested in the original study). Although Gapping and No-Gapping conditions were balanced across experimental items, Gapping sentences in this study were in fact in the minority if one takes all stimuli, including fillers, into account. Of the 96 filler items,

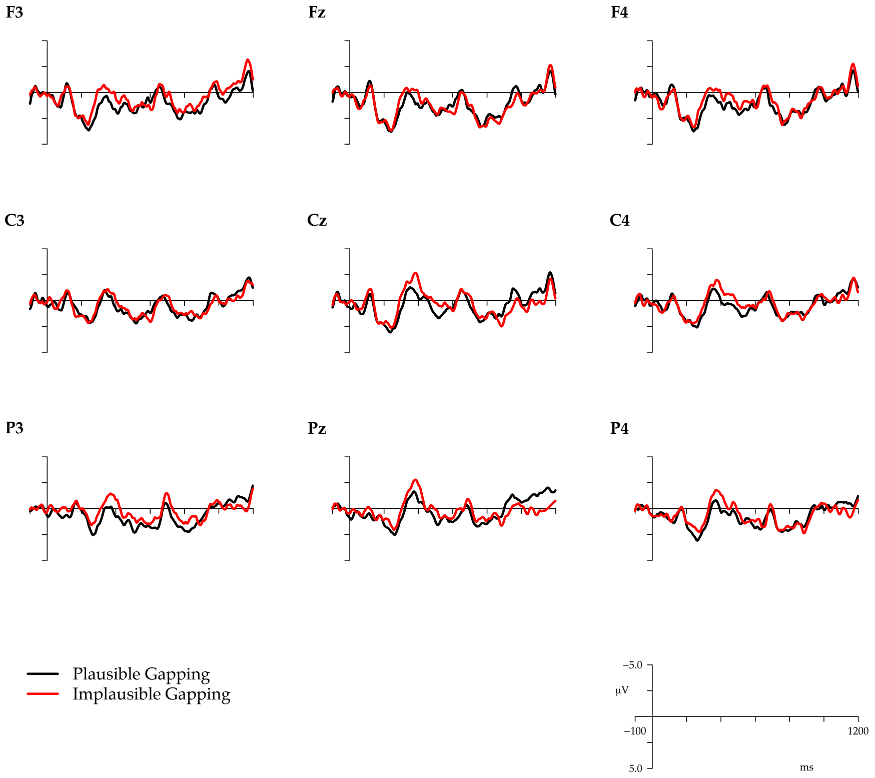

Figure 5.3: Grand averages of Plausible Gapping condition (***a***) and Implausible Gapping condition (***b***) at onset (y-axis) of the noun (*bloemen*) at electrode sites F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. Corresponding example sentences can be found on page 86.

only 16 contained Gapping constructions, notably containing a coordination with the connective "maar".

In line with the original study, the factor PLAUSIBILITY did not yield an N400 effect at the position of the noun. Although it was numerically apparent, it was not statistically significant. Possibly, a time window of 200 ms is too large, meaning that an N400 component in this design might be expressed at a shorter latency.

The P600 effect for the factor PLAUSIBILITY in the original study could not be corroborated in this replication. A late positive deflection was visible but it was not statistically significant. The interaction effect of PLAUSIBILITY with HEMISPHERE shows that implausible items yielded opposing mean amplitudes – negative in the left hemisphere and positive in the right hemisphere.

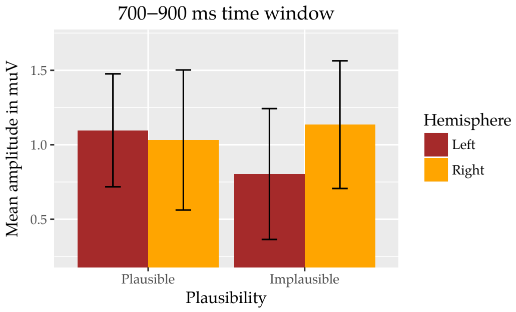

Figure 5.4: Error bar graph of interaction effect of PLAUSIBILITY by HEMISPHERE on the noun (*bloemen*) on lateral electrodes in the 700-900 time window.

This interaction may explain why no straightforward P600 could be established.

Of the most interest in relation to this thesis is the effect of the factor GAPPING, which could be corroborated in the 700-900 ms and 900-1,200 ms time windows. At the position of the noun, a process of integration may be assumed and it seems likely that this is expressed by late positive P600-like deflections. In addition, a close look at Figure 5.5 points the attention to earlier time points. It seems that a positivity is already apparent at an early stage at around 350 ms. Again, it could be that analyses using shorter time windows may have revealed significant effects here.

A few caveats are in order though. Firstly, negative deflections observed at the determiner may have had the effect of amplifying any positive effect in the epochs of the noun. Pre-stimulus activity may be problematic for the evaluation of critical time points (Luck, 2014:256). In that sense, a positivity could be seen as artefactual effect. Future designs should overcome this problem. Secondly, the analysis of this replication is based on 18 participants instead of 30 in the original study, which yields less statistical power. Nevertheless, the effect sizes for the effect of GAPPING on midline sites in the 700-900 ms and 900-1,200 ms time windows are relatively large.

5.1.5 Conclusion

In addition to an evaluation of previous studies a proper study should commence with an attempt to replicate previous published findings. Unfortunately, this prerequisite is generally seen as an unrewarding task and therefore often left out. Although results of a replication study may deviate from the original, they may still give insight as to how to proceed. The current replication gave rise to a result that was hypothesised, but which was not apparent in the original study. A LAN-like component was found that can be regarded

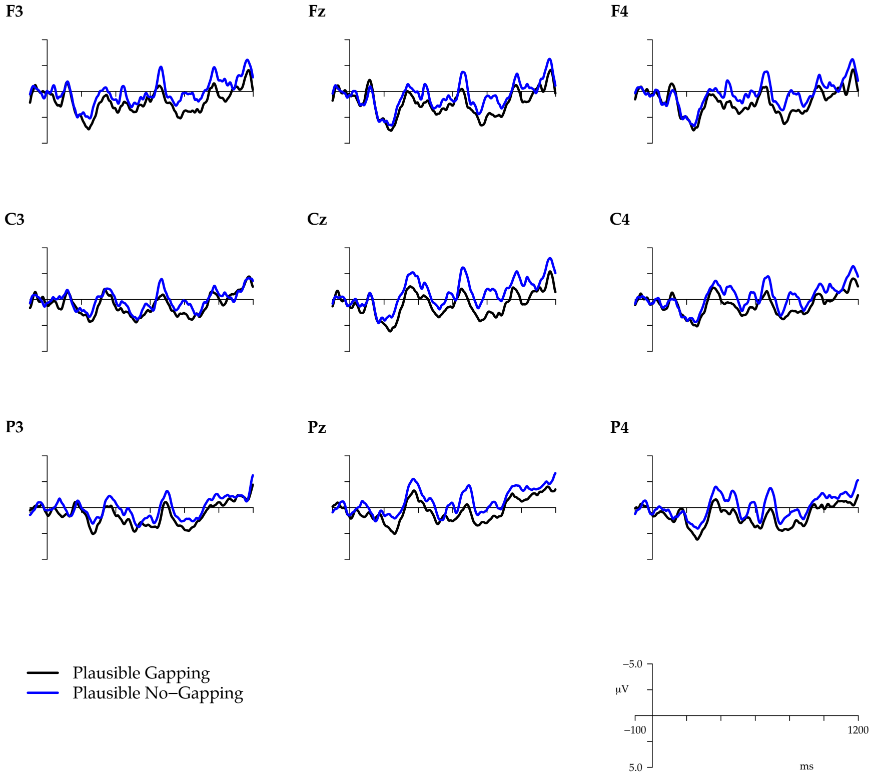

Figure 5.5: Grand averages of Plausible Gapping condition (***a***) and Plausible No-Gapping condition (***c***) at onset (y-axis) of the noun (*bloemen*) at electrode sites F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. Corresponding example sentences can be found on page 86.

as an index of prediction. This effect seems marginally sensitive to individual variation, but may in fact be due to the relative frequency of Gapping items in the stimulus list. Furthermore, the effects of implausible items appeared to be less strong than in the original study. Again, no N400 was found and additionally a P600 was only numerically visible. However, the replication does corroborate processes of integration of a plausible elided verb at the critical noun, interpreted in the original study as being on a par with the integration of object *wh*-phrases. Gapping of plausible phrases, then, seems to be most appropriate to investigate further and this will be pursued in the continuation of the current research.

5.2 Norming stimuli I: acceptability of structural elision

5.2.1 Purpose

Throughout the experiments reported in this thesis I make use of sentences as stimuli. Preferably the stimuli should be designed such that they can be used in different experiments. This method allows us to compare results from different experiments. Furthermore, I wish to use grammatical and interpretable stimuli to investigate Gapping and Stripping. During the ERP experiments, participants will answer a comprehension question after every stimulus. On the one hand, I can make sure that participants actually *read* the sentences, on the other, comprehension scores can be analysed with respect to the complexity of the ellipsis.

Test sentences should be minimal pairs. Given that during the ERP experiments sentences will be presented by means of a word by word reading task, a fixed measure point – one word – is required to compare effects of ellipsis between conditions.This section is a report of a pilot study of stimuli in which structural complexity of the ellipsis was manipulated: phrases are cut off step by step (condition by condition) reducing the amount of overt structure step by step. The goal of this pretest is to ascertain the acceptability of the stimulus sentences, in order to be able to reject uninterpretable stimuli and gain awareness of acceptability differences across the stimuli set. The ERP experiments described in Chapter 7 were designed on the basis of the tested items.

5.2.2 Method

Building on previously used materials

Since only one peer-reviewed ERP study of Gapping processing in Dutch (the replicated study reported in section 5.1 above) had been published at the time I started this research project, it seemed most practical to develop stimuli on the basis of the test sentences from that study. As a first step, I designed 44 quadruplets as exemplified in (2).

(2) a. Omdat Hilde in de voortuin het gazon onderhield en Ralph in
Because Hilde in the front.garden the lawn maintained and Ralph in
de achtertuin de paden harkte, waren de buurtgenoten vrolijk.
the back.garden the paths raked, were the neigbours happy
'Because Hilde maintained the lawn in the front garden and Ralph raked the paths in the back garden, the neighbours were happy.'

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 neigbours happy
'Because Hilde maintained the lawn in the front garden and Ralph the paths in the back garden, the neighbours were happy.'

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 neigbours happy
'Because Hilde maintained the lawn in the front garden and Ralph in the back garden, the neighbours were happy.'

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 neigbours happy
'Because Hilde maintained the lawn in the front garden and Ralph too, the neighbours were happy.'

Condition *a* represents the control sentence: a fully-fledged structure with all phrases in place. In condition *b*, the verb is elided in the right conjunct, in condition *c* the verb with the object are elided, and in condition *d* every phrase in the right conjunct except for the subject is stripped and replaced by 'too'. While the original sentences are made of conjunctions, in the new stimuli a conjunction is captured within a subordinate adjunct. The motivation to do so, was to be able to cut off phrases step by step while having a stable measuring point: the main verb *waren*. At this point, the ellipsis should be resolved. Furthermore, the completion of the subordinate clause does not hinge on the main clause as would be the case with a subject clause (e.g. *That John bought a book surprised his mother.*). In such sentences the main verb needs the subject – the whole subordinate clause – in order to integrate the arguments. As a consequence, this process may overshadow the ellipsis resolution mechanism. As can be seen in (2), the stimuli are closely related to the crucial stimuli as used by Kaan et al. (2013), repeated here in (3).

(3) a. Hilde onderhield het gazon in de voortuin en Ralph de
Hilde maintained the lawn in the front.garden and Ralph the
paden in de achtertuin.
paths in the back.garden
'Hilde maintained the lawn in the front garden and Ralph the paths in the back garden.'

b. Hilde onderhield het gazon in de voortuin terwijl Ralph de
Hilde maintained the lawn in the front.garden while Ralph the
paden in de achtertuin harkte.
paths in the back.garden raked
'Hilde maintained the lawn in the front garden while Ralph raked the paths in the back garden.'

As discussed in Chapter 3.5 Kaan et al., compared (3a) with (3b), which differ in structure. By contrast, my aim is to compare measurement point(s) between sentences with the same structure. Kaan et al. utilised the noun phrase *de paden* as measuring point. Note that in (3a), this phrase is in a main clause, while in (3b) it is in a subordinate clause. As explained earlier, they reasoned that the determiner is expected in (3b) and not expected in (3a). In that sense, their results are contingent on expectancy effects which are partly induced by the clause type, i.e. the conjunction.

Only 44 stimuli could be used of the available 117 from Kaan et al since some of their original stimuli contained noun phrase modifiers. A disadvantage of such sentences for the purpose of cutting off phrases step by step is, that such modifiers cannot appear on their own and hence cannot be used in the proposed setting. For example, in (4) *de staking* cannot be separated from *van de monteurs*. This problem does not arise with adjuncts as is shown in (5).

(4) a. Renate organiseerde de staking van de monteurs.
Renate organised the strike of the mechanics
'Renate organised the strike of the mechanics.'

b. *Van de monteurs organiseerde Renate de staking.
of the mechanics organised Renate the strike
int: 'Of the mechanics Renate organised the strike.'

(5) a. Renate organiseerde de staking in de ochtend.
Renate organised the strike in the morning
'Renate organised the strike of the mechanics.'

b. In de ochtend organiseerde Renate de staking.
in the morning organised Renate the strike
'In the morning Renate organised the strike.'

Other sentences discarded from Kaan et al's original set contained either potential ambiguities or adjuncts that differed in semantic function between conjuncts.

In the original sentences, the objects, such as *het gazon* in (3a), are all definite expressions. Since we expected that non-generic objects would be more difficult to interpret in the proposed conditions *c* and *d*, we changed them to indefinite objects – where possible. As we can see in (6), an object that refers to exactly one of a set may cause an odd reading when it is elided in the right conjunct.

(6) a. Nina arranged the grill and Ruben hooked up the tap.
b. ?Nina arranged the grill and Ruben too.
c. Nina arranged a grill and Ruben too.

In (6b), it is hard to believe that one and the same grill is arranged twice, while in (6c) it is plausible that two people arranged two grills separately. The difference here is easily explained in terms of the definiteness of the NPs. Definite NPs in (6a) and (6b) refer to unique (some scholars use the term "familiar") entities in the context. Note, that the difficulty caused by uniqueness does not (immediately) arise with so-called weak definites such as *het gazon* in (3a) above.

Again on the basis of material used in Kaan et al. (2013), fillers were designed. (7a) is an example of a plausible filler and (7b) is an example of an implausible filler.

(7) a. Terwijl Gerda op de bank televisie keek, zat Sanne aan tafel te
While Gerda on the couch television watched sat Sanne at table to
puzzelen.
puzzle
'While Gerda watched TV on the couch, Sanne solved a crossword at the table.'

b. Nadat Esmee de post bij de villa bezorgde, keek de hond luid
After Esmee the mail at the villa delivered looked the dog loudly
naar haar.
at her
int: 'After Esmee delivered the mail at the villa, the dog looked at her loudly.'

While all test sentences started with the conjunction *omdat* 'because', fillers started with *omdat* 'because', *aangezien* 'since', *doordat* 'as a result of', *nadat* 'after', *voordat* 'before', or *terwijl* 'while'. Fillers differed in word length between 9 and 21 words. Thirty-six plausible fillers and 32 implausible fillers were constructed. A full list of the stimuli of this pretest can be found in Appendix B.

As discussed in section 3.3, complexity in ellipsis is subject to inconclusive evidence, Copy α and the cue-based mechanism predicting comparable results. However, in this design, it is not the form of the antecedent which changes but the complexity of to be recovered material, which is possible when using Gapping-like constructions. This allows us to compare different sizes of structure elisions within one sentence. In line with the suggestion of Poirier, Wolfinger, Spellman, and Shapiro (2010), we hypothesise that if more structure is elided, this might affect processing load.

Participants

Twenty participants participated and received €3 for their cooperation. Two participants did not obey the instructions: one took too much time to complete the experiment, the other appeared to have misunderstood the task. Two additional participants were invited as substitutes. The results below are based on twenty participants (four male; $M_{\text{Age}} = 24.45$, range 18-49).

Procedure

The items were divided over four lists using a Latin Square design. Each list contained only one member of each quadruplet and each participant rated only one list. The stimuli, which were interspersed with the 68 fillers described above (32 uninterpretable and 36 interpretable), were presented in an individually randomised order using the software PsychoPy (Peirce, 2007, 2009). Uninterpretable sentences had a well-formed structure but contained mismatching lexical items. Participants were asked to rate the sentences on a seven-point scale (see section 3.1 for a discussion on acceptability tests). They were encouraged to take into account the structure as well as the interpretability of the presented sentences. Also, they were asked to react as quickly as possible to obtain intuitive responses. Before the actual test, which contained 112 sentences, participants completed a practice session of 21 sentences. The experimental session took 25 minutes at the most.

5.2.3 Results

The mean ratings were calculated per quadruplet and per sentence. Quadruplets of test sentences of which one item had an average score below 4 were disregarded. Since the stimuli would be counterbalanced in the subsequent ERP experiment so that each participant only saw one sentence of a quadruplet, the number of quadruplets should be dividable by 4. Of the 38 remaining quadruplets an additional 2 quadruplets were removed on the basis of lowest scores per quadruplet and per sentence. After applying these criteria, thirty-six quadruplets remained for the following analysis. One implausible filler sentence was rated 5.20 on average. This filler was excluded along with the eight discarded quadruplets.

Condition	Mean	N	Standard Error
a	5.63	36	.11
b	5.49	36	.09
c	5.15	36	.11
d	5.17	36	.10
Total	5.36	144	.05

Table 5.1: Means of rating of test sentences per condition after correction.

In Table 5.1 the average ratings of the remaining test sentences are listed. The mean rating of plausible and implausible filler sentences was $M = 6.55$ $[SE = 0.07]$ and $M = 2.19$ $[SE = 0.06]$, respectively. The means of the test sentences were evaluated using a one-way ANOVA. Between four test conditions a main effect of CONDITION was found, $[F(3, 140) = 5.21, p = .002, \eta^2_p = .100]$. A Bonferroni post hoc analysis of planned contrasts revealed that condition *a* differed marginally from condition *b* $[p = .069]$, but it differed significantly from condition *c* $[p = .011]$ and condition *d* $[p = .022]$. No other significant differences were apparent.

5.2.4 Discussion

The stimuli in this acceptability test consisted of plausible fillers, implausible fillers, and test sentences – the items of main interest. Relative to the control condition, the test conditions displayed a decline in ratings as more and more structure was elided. As expected, condition *a*, the control condition without ellipsis, was rated the highest while sentences with more elided structure were judged lower. Especially the inclusion of an object in the ellipsis (conditions *c-d*) had an effect on the mean ratings. Note though, that the steps between conditions *b, c, d* were not significant. Notably, numerically, the difference between the Gapping condition *c* and the subtype of Gapping (Stripping) condition *d* in which more structure was elided was almost equal.

The decreasing ratings relative to the control condition could be related to the "amount of repair" of structure as discussed in Chapter (1). In that sense, more elided structure may amount to a relative processing cost, while Stripping constructions (condition *d*) might be easier to repair than Gapping constructions. It will be interesting to see to what extent a processing cost affects comprehension of elliptical sentences and how this is reflected in terms of ERPs. In the ERP experiments in which a comprehension task will be included I will try to establish this.

One may ask why the test sentences were generally judged less acceptable than the plausible fillers. A tentative explanation could be that the test sentences consisted of three clauses instead of two as is the case in the fillers. Possibly, participants found sentences with more clausal content more difficult. During the debriefing of the experiment some of the participants indeed pointed to the issue of "too much information" in one sentence. Additionally, corpus research could be helpful to investigate to what extent the form of the test items differs from that of the filler items in terms of frequency. Stimuli with relatively more elided structure were rated relatively low. As mentioned above, this could be down to processing cost, but it could also be that such sentence forms are not frequently used. Note that low frequency items usually correlate with processing difficulty (see for example Levy, 2008).

5.2.5 Conclusion

The goal of this pretest was to check which of the quadruplets, that were designed on the basis of the first ERP experiment on Dutch Gapping, could be used in the planned ERP experiments reported in this thesis. By conducting a computer administered experiment in which the test sentences were presented together with plausible and implausible fillers, thirty-six of 44 quadruplets appeared to have adequate acceptability ratings. This means that these stimuli are considered as acceptable by native speakers of the language in terms of structure and interpretation. Compared to the control condition, a tendency of acceptability to decline as more structure is elided was observed. This could indicate that, when relatively more structure has to be recovered, processing load increases. Using the pretested stimuli in ERP experiments, I will try to shed light on the nature of processing mechanisms. Additionally, I will be able to compare acceptability judgement data from this pilot to comprehension data that will be collected and analysed in Chapter 6.

5.3 Norming stimuli II: acceptability of quantifiers

5.3.1 Purpose

In this norming study, proposed test sentences with semantic difficulty were tested for acceptability by native speakers. Items were included to compare the quantifiers *elke* "every" and *alle* "all" with the determiner *de* "the" in Gapping conditions and Stripping conditions. The latter modulation is tested in the ERP experiment reported in Chapter 7. In other items the additive marker *ook* "too" contrasts with the polarity marker *niet* "not". These items are included for follow-up experiments (not reported in this thesis).

5.3.2 Method

Participants

Forty native speakers of Dutch (10 male; $M_{\text{Age}} = 22.24$, range 19-31) participated and received €5 compensation.

Stimuli

On the basis of the original data set used by Kaan et al. (2013), ninety-five quintuplets as in (8) below were designed.

(8) a. Koen verving de kast in de woonkamer, en Judith de lamp in de gang.
Koen replaced the cabinet in the living.room and Judith the lamp in the hall
'Koen replaced the cabinet in the living room, and Judith the lamp in the hall.'

b. Koen verving elke kast in de woonkamer, en Judith de lamp in de gang.
Koen replaced every cabinet in the living.room and Judith the lamp in the hall
'Koen replaced the cabinet in the living room, and Judith the lamp in the hall.'

c. Koen verving de kast in de woonkamer, en Judith niet.
Koen replaced the cabinet in the living.room and Judith not
'Koen replaced the cabinet in the living room, and Judith did not.'

d. Koen verving de kast in de woonkamer, en Judith ook.
Koen replaced the cabinet in the living.room and Judith too
'Koen replaced the cabinet in the living room, and Judith too.'

e. Koen verving elke kast in de woonkamer, en Judith ook.
Koen replaced every cabinet in the living.room and Judith not
'Koen replaced every cabinet in the living room, and Judith too.'

As I have explained in Chapter 2.4.2, quantifying expressions may be a burden on mechanisms of movement and/or copying since additional structural information has to be analysed. Therefore, I created stimuli to test the difference between quantified phrases and phrases containing a definite article. Condition *a* is the same as the plausible Gapping condition that was used in the replication of Kaan et al. 2013 reported earlier. This condition contrasts with condition *b* in which the determiner of the object in the left conjunct is replaced by a quantifier. In condition *c*, the negative polarity marker at which the ellipsis is resolved can be compared to the (positive) additive marker in condition *d*. In turn, condition *d* can be contrasted with condition *e* to estimate the difference between a determiner and a quantifier in Stripping constructions. The latter comparison will be further explored in Chapter 7 which reports an ERP experiment that was designed to focus on the semantic aspect of retrieval and integration processes.

Procedure

The items were counterbalanced over five lists. Each list contained only one member of each quintuplet and each participant rated only one list. The stimuli, which were interspersed with an additional 93 fillers of which 22 uninterpretable, were presented in an individually randomised order using the software PsychoPy (Peirce, 2007, 2009). Uninterpretable sentences had a well-formed structure but contained mismatching lexical items (similar to the uninterpretable items used in the pretest described above). Participants were instructed to take into account the structure as well as the interpretability of the presented sentences, and to rate the sentences on a seven-point scale. To obtain intuitive responses, they were asked to react as quickly as possible. Before the actual test, which contained 188 sentences, participants completed a practice session of 21 sentences. The session lasted 30 minutes on average.

5.3.3 Results

Due to a scripting error, three conditions of one stimulus set were wrongly coded and presented as the same condition. Therefore, the analysis is based on the remaining 94 stimuli sets. The mean ratings were calculated per sentence. In Table 5.2, the means and standard errors are listed for the five test conditions with mean scores higher than 4.

Condition	**Mean**	**N**	**Standard Error**
a	5.85	92	.07
b	5.16	75	.07
c	5.23	80	.06
d	5.40	86	.07
e	4.99	70	.08
Total	5.36	403	.04

Table 5.2: Means of rating of test sentences per condition after correction.

The mean ratings of plausible and implausible filler sentences were $M = 6.55$ [$SE = 0.07$] and $M = 2.80$ [$SE = 0.06$], respectively. Table 5.2 shows that low mean scores coincide with a relatively high exclusion rate. In general, the items containing quantifiers were judged least acceptable. Since conditions *d* and *e* are to be tested in the ERP experiment reported in Chapter 7, these items were analysed in more detail. From the data set, 42 pairs of conditions *d* and *e* were chosen such that they matched in terms of their mean ratings. Items within such a pair maximally differ in 1.25 average acceptability score points. The range of average scores among chosen items was 4.38-6.50; means of condition *d* [$M = 5.46$, $SE = 0.10$] and condition *e* [$M = 5.32$, $SE = 0.08$] did not differ significantly [$t(41) = 1.41, p = .166, d = .218$].

5.3.4 Discussion

Since the sentence structures in this norming study more closely resemble the original stimuli tested by Kaan et al. (2013) than the stimuli in the first norming study, it was easier to construct a larger set of stimuli. As a consequence, a set consisting of conditions *d-e* could be chosen in which the means differ minimally. Note that therefore we have the luxury of controlling the effect of acceptability in subsequent experiments using these stimuli, but this is not possible for the stimuli set derived from the first norming study. At least numerically, the sentences with quantifiers were rated lower than the other conditions, followed by condition *c*, which contained negation at the ellipsis site. In this sense, semantic difficulty seems to correlate with lower ratings, that is, acceptability may decrease as a function of semantic complexity.

As was the case in the first norming study, the elliptical sentences were rated lower than the plausible fillers. It was proposed in the first norming study that this may be down to the inclusion of three clauses in the sentence structure. Since the elliptical sentences in the current set do not have this property, it may in fact be the case that ellipsis is less acceptable than fully-fledged sentences in general. It should be noted though that "acceptability" is not only a measure of grammaticality but it is also dependent on the relative difficulty of interpretation and therefore likely related to a relative processing cost that may resemble the resolution process. In the subsequent ERP experiments, this will be investigated in more detail.

5.3.5 Conclusion

A norming study was carried out to ascertain the acceptability of stimulus sentences containing Gapping and Stripping constructions which differed in terms of semantic complexity. From the pool of tested sentences a set has been selected for use in the ERP experiment described in Chapter 7, where semantic complexity is investigated. In contrast to the result of the first norming study, a set could be compiled in which the means of acceptability differed only minimally. Consequently, the factor ACCEPTABILITY need not be considered as factor in the ERP experiment on semantic complexity.

Additional stimulus sets that have been tested in this section may be used in future experiments – for example, as a follow-up of the current thesis (e.g. a comparison between the additive markers *ook* and *niet* to investigate negated elisions). In the remaining chapters, however, we will be concerned with the four ERP experiments that have been conducted.