Computing and recomputing discourse models: An ERP study

Baggio, G.; van Lambalgen, M.; Hagoort, P.

DOI
10.1016/j.jml.2008.02.005

Publication date
2008

Published in
Journal of Memory and Language

Citation for published version (APA):
COMPUTING AND RECOMPUTING DISCOURSE MODELS: AN ERP STUDY

GIOSUÈ BAGGIO, MICHIËL VAN LAMBALGEN, AND PETER HAGOORT

Author affiliations:

GIOSUÈ BAGGIO
F.C. Donders Centre for Cognitive Neuroimaging
Radboud University Nijmegen

MICHIËL VAN LAMBALGEN
Institute for Logic, Language and Computation
University of Amsterdam

PETER HAGOORT
F.C. Donders Centre for Cognitive Neuroimaging
Radboud University Nijmegen and
Max Planck Institute for Psycholinguistics, Nijmegen

Address correspondence to:

GIOSUÈ BAGGIO
F.C. Donders Centre for Cognitive Neuroimaging
Radboud University Nijmegen
P.O. Box 9101
NL-6500 HB Nijmegen
The Netherlands
Phone: +31.024.3610980
Fax: +31.024.3610989
E-mail: g.baggio@fcdonders.ru.nl
ABSTRACT. While syntactic reanalysis has been extensively investigated in psycholinguistics, comparatively little is known about reanalysis in the semantic domain. We used event-related brain potentials (ERPs) to keep track of semantic processes involved in understanding short narratives such as ‘The girl was writing a letter when her friend spilled coffee on the paper’. We hypothesize that these sentences are interpreted in two steps: (1) when the progressive clause is processed, a discourse model is computed in which the goal state (a complete letter) is predicted to occur; (2) when the subordinate clause is processed, the initial representation is recomputed to the effect that, in the final discourse structure, the goal state is not satisfied. Critical sentences evoked larger sustained anterior negativities (SAN) compared to controls, starting around 400 ms following the onset of the sentence-final word, and lasting for about 400 ms. The amplitude of the SAN was correlated with the frequency with which participants, in an off-line probe-selection task, responded that the goal state was not attained. Our results raise the possibility that the brain supports some form of non-monotonic recomputation to integrate information which invalidates previously held assumptions.

KEYWORDS. Semantics; Discourse; Non-monotonicity; ERPs; Sustained anterior negativity.

1. INTRODUCTION

In the past three decades experimental research using event-related potentials (ERPs) has provided numerous insights into word, sentence and discourse comprehension. However, as has been noted, “a cognitive neuroscience approach to language has not as yet merged with linguistic and psycholinguistic research programmes” (Brown & Hagoort 1999). One linguistic research program that may contribute to understanding the basis of meaning in the human brain is semantic theory. Logicians and formal semanticists since the ‘dynamic turn’ (Peregrin 2003) have shifted their attention from describing semantic competence to modeling cognitive update and information exchange. A case in point is a recent proposal by van Lambalgen & Hamm (2004) which regards comprehension as an incremental, yet non-monotonic process whereby temporary structures are set up in working memory and may be later revised on the basis of further discourse information. Although evidence for semantic reanalysis exists (Carreiras et al. 1996; Sturt 2007), the issue has arguably received less attention than it deserves. The purpose of the present study is to contribute filling this gap. We used ERPs to test a processing hypothesis proposed by Baggio & van Lambalgen (2007) as an application of the formal, non-monotonic semantics of tense and aspect of van Lambalgen & Hamm (2004).
1.1. ERP research on semantic processing. Event-related potentials have proved useful to address a number of issues concerning the relative complexity and time course of semantic processes. Kutas & Hillyard (1980) conducted the first ERP experiment in which linguistic factors were successfully manipulated, in this case the semantic plausibility of a word given the preceding sentence context:

(1) a. The officer shot the man with a gun.
   b. The officer shot the man with a moon.

Compared to ‘gun’, the anomalous noun ‘moon’ resulted in a larger negative shift starting around 250 ms after word onset, peaking at 400 ms, and lasting for approximately another 150 ms. This ERP component, called N400 after its polarity and peak latency, is known not to be affected by other unexpected events, such as variations in the physical properties of the stimuli. Larger N400s are also triggered by semantically plausible words which are nevertheless judged as less preferred in a given sentence context (Kutas & Hillyard 1984; Hagoort & Brown 1994), for example ‘pocket’ in (2b):

(2) a. Jenny put the sweet in her mouth after the lesson.
   b. Jenny put the sweet in her pocket after the lesson.

The amplitude of the N400 is also modulated by lexical items which provide information conflicting with the discourse context (van Berkum, Hagoort, & Brown 1999; van Berkum, Zwitserlood, Hagoort, & Brown 2003) or world knowledge (Hagoort et al. 2004). In sum, although every content word evokes an N400, the amplitude of the negative shift appears to be affected by the degree of semantic fit of a lexical item with the preceding context and the knowledge base relevant for its integration.

Semantics-related negativities different from the N400 have also been found. Van Berkum and colleagues recorded ERPs while participants read (van Berkum, Brown, & Hagoort 1999) and listened to (van Berkum, Brown, Hagoort, & Zwitserlood 2003) discourses in which a particular NP in a target sentence could denote either a single referent introduced in the preceding discourse or two equally suitable referents. For instance, (3c), containing the NP ‘the girl’, could follow either the single-referent context (3a) or the double-referent context (3b):
47 (3) a. David had told the boy and the girl to clean up their room before lunch time. But the boy had stayed in bed all morning, and the girl had been on the phone all the time.

b. David had told the two girls to clean up their room before lunch time. But one of the girls had stayed in bed all morning, and the other had been on the phone all the time.

c. David told the girl that ...

Referentially ambiguous NPs, such as ‘the girl’ in (3c) following (3b), elicited a sustained anterior negativity (SAN), emerging 300-400 ms after noun onset and lasting for several hundreds of milliseconds. The SAN differed from typical instances of the N400 in duration (‘sustained’) and scalp distribution (‘anterior’). Motivated by earlier research (Mecklinger et al. 1995; Müller et al. 1997), the time-course and topographical profile of the observed ERPs are taken to suggest that “at least some of the processing consequences of referential ambiguity may involve an increased demand on memory resources” (van Berkum, Brown, Hagoort, & Zwitserlood 2003).

Another study (Münte et al. 1998) reported sustained anterior negativities. ERPs were recorded while subjects read narratives differing in the initial temporal connective:

(4) a. After the scientist submitted the paper, the journal changed its policy.

b. Before the scientist submitted the paper, the journal changed its policy.

‘Before’ sentences elicited a larger sustained negativity, maximal over left anterior sites. At the left frontal electrode, ERP responses to ‘before’ and ‘after’ diverged approximately 300 ms after sentence onset. The effect lasted throughout the sentence and was larger during the second clause. The difference of anterior negativity between ‘before’ and ‘after’ items was positively correlated with participants’ working memory span. Münte et al. argue that the slow negative shift evoked by ‘before’ sentences reflects working memory operations involved in computing a model for (4b) in which the events are represented in their actual order of occurrence. That is, in contrast with (4a), (4b) requires additional memory resources as the two events are mentioned in reverse temporal order.

The connection between sustained anterior negativities and working memory is explicit in van Berkum, Brown, & Hagoort (1999), van Berkum, Brown, Hagoort, & Zwitserlood
COMPUTING AND RECOMPUTING DISCOURSE MODELS: AN ERP STUDY

(2003) and Münte et al. (1998). However, there is no full agreement on a functional account, based on linguistically-informed notions, of these findings. For instance, while van Berkum et al. suggest that the sustained anterior negativity reflects ‘referential processing’, Münte et al. seem to implicate that ‘additional discourse-level computations’ of the temporal and causal profiles of the events described by ‘before’ and ‘after’ sentences are responsible for the observed slow negative shifts. Matters appear to be further complicated by the finding that sustained anterior negativities are elicited by constructions in which complexity at the syntax-semantics interface is affected, as in long-distance wh-dependencies (King & Kutas 1995; Müller et al. 1997; Fiebach et al. 2002; Felser et al. 2003; Phillips et al. 2005).

Despite the differences between the conditions in which sustained anterior negativities have been observed, the proposed processes can be brought under a single umbrella term, which we shall refer to as ‘computing a discourse model’. Formal semantics, at least since Discourse Representation Theory (DRT) (Kamp 1981), has assumed that interpreting definite and indefinite NPs, resolving anaphoric pronouns, determining the order of events, establishing wh-dependencies and other cross-clause and cross-sentence processes concur in the construction of a discourse model, that is, a cognitive representation making a given narrative true. More recent proposals, which build upon DRT and add some sophistication to it, view discourse comprehension as a process in which lexical meanings, references and world knowledge interact to produce consistent discourse representations (van Lambalgen & Hamm 2004; Hamm et al. 2006). Pragmatic constraints and causal/world knowledge are brought to the fore by these accounts. Furthermore, discourse models as envisaged by the theory (called ‘minimal models’, see Section 1.3) can be efficiently computed by artificial neural networks, which account for some capabilities and limitations of working memory (Stenning & van Lambalgen 2005). Therefore, in this framework it becomes possible to raise and address a number of issues concerning the complexity of computing discourse models in working memory. To see in some detail how this could be done, we must first introduce the linguistic phenomenon with which we shall be concerned: the ‘imperfective paradox’.

1.2. The imperfective paradox. Verb phrases (VPs) can be semantically classified as states (‘know’, ‘love’ etc.), activities (‘run’, ‘write’ etc.), accomplishments (‘write a letter’, ‘bake a cake’ etc.), achievements (‘finish’, ‘reach’ etc.) and points (‘flash’, ‘hop’ etc.) (Steedman...
1997). Accomplishments involve the activity from which they are derived. For instance, ‘write a letter’ is constituted by the activity ‘write’ and the direct object ‘a letter’, which need not refer to an existing entity, but carries information about the goal toward which the writing activity is directed. Here we use the term ‘activity’ to denote both the aspectual class of VPs such as ‘write’ in the above classification and the atelic process involved in all accomplishments. We use ‘progressive’ and ‘imperfective’ interchangeably to allow the reader to see the connection between the semantics of the progressive and the imperfective paradox, although this is not entirely correct (Comrie 1976).

Let us consider accomplishments first:

(5) The girl was writing a letter when her friend spilled coffee on the tablecloth.

From (5) the reader would typically conclude that, barring unforeseen circumstances, the girl will attain the desired goal and would thus assent to the statement ‘The girl has written a letter’ (see 2.2.2 and 3.1 for evidence supporting this claim). Such an inference is based on the assumption that spilling coffee on the tablecloth is usually neutral with respect to the writing activity. That is, it is not a typical immediate cause leading to the termination of the activity. It is possible to imagine situations in which writing was temporarily interrupted or even terminated by the accident. However, as the data reported in Sections 2.2.2 and 3.1 will demonstrate, failing to explicitly mention an obstacle in the discourse is sufficient to lead the reader to assume that there was no such obstacle to attaining the intended goal.

We hypothesize that the inference to a goal state is defeasible or non-monotonic, that is, it can be suppressed if the discourse describes an event which terminates the relevant activity:

(6) The girl was writing a letter when her friend spilled coffee on the paper.

Assuming that writing was intended to occur on the same paper sheets on which coffee was spilled, the accident is sufficient to terminate the activity and it is therefore a disabling condition for obtaining a complete letter. Accordingly, on the basis of (6) the reader would assent to ‘The girl has written no letter’.

Suppression can obtain only with accomplishments, not with activities (Rothstein 2004). In accomplishments, the object NP ‘a letter’ expresses the existence of a natural culmination point or ‘canonical goal’ toward which the writing activity is directed, namely a complete letter. Activities, for instance ‘writing letters’, do not involve any such canonical goal. The
use of the bare plural ‘letters’ indicates that the number of letters is (for the speaker and
the hearer) unspecified and that, therefore, the activity has no natural culmination point.
Accordingly, a narrative containing the activity VP ‘writing letters’ will be interpreted as
entailing that ‘The girl has written one or more letters’ regardless of the consequences of
the second event on the writing activity:

(7) The girl was writing letters when her friend spilled coffee on the tablecloth.
(8) The girl was writing letters when her friend spilled coffee on the paper.

There appears to be something paradoxical about (6) in its relation to (5), which is not
found in the pair (7)-(8). Whereas it belongs to the meaning of the accomplishment ‘writing
a letter’ that the writing activity is directed toward the goal state of a complete letter, the
actual occurrence of that consequent state can be denied without contradiction. How can
an essential component of the meaning be denied without destroying meaning itself? This
is the so-called ‘imperfective paradox’.

1.3. Minimal models, inference in the limit, and recomputation. Language processing
amounts to incrementally computing a discourse representation given lexical, syntactic and
contextual constraints (Hagoort 2006). To render computation tractable, discourse models
must be ‘minimal’, that is, in a precise mathematical sense (van Lambalgen & Hamm 2004),
the simplest possible structures making the narrative true. Minimal models behave like
‘closed worlds’, in which only those propositions which are asserted in discourse, or which
can be inferred from it or from background knowledge, are represented as true in the model.
For the remaining cases, a distinction must be drawn. Propositions which are mentioned
in discourse, but are not asserted and do not follow from what is said or from background
knowledge (e.g. the antecedent of a conditional), are represented as false in the minimal
model. In logical terms, these propositions still belong to the finite language upon which
the construction of the minimal model is based. But as long as nothing forces their truth,
they will be taken as false. Propositions which are not part of the finite language – because
they do not occur in the discourse context or in background knowledge – are not included
in the minimal model, that is, they are not represented as either true or false.

One important upshot of the theory is that the occurrence of a goal state can be inferred
from a minimal model of a discourse containing an accomplishment in the past progressive.
As soon as the sentence ‘The girl was writing a letter’ is processed, the system constructs
a minimal model in which the goal state (a complete letter) is attained at some time later
than the interval referred to by the progressive. Two remarks concerning this crucial point
are in order. First, interpretation is based on the ‘closed world assumption’: if no disabling
condition is described in discourse (so far), it will be (temporarily) assumed that there is no
obstacle interfering with the writing activity. Second, the conclusion that eventually a letter
is accomplished is an instance of predictive inference or, more precisely, inference in the limit:
given that writing is asserted to hold some time in the past, that it can be assumed there are
no obstacles for the writing activity, that some form of inertia holds (writing continues if it
is not hindered by external forces), and that a letter is a finite object, it can be expected that
the process will converge – ‘in the limit’ – to a complete letter. This holds for both neutral (5)
and disabled (6) accomplishments. Now, when the initial model is extended with a ‘when’
clause describing an event which terminates the writing activity (i.e. a disabling condition),
the goal state inference will be suppressed. The subordinate clause ‘when her friend spilled
coffee on the paper’ will lead to the retrieval of causal knowledge from semantic memory
to the effect that the coffee accident terminated the writing activity. Spilling occurred during
the writing process, from which follows that the accident took place before a complete letter
was obtained. The writing event can be imagined as an open interval, where the goal state
(a complete letter) is no longer part of the structure. We shall use the term ‘recomputation’
to refer to the suppression of the goal state inference when the subordinate clause in (6) is
processed. Because (5) describes a neutral scenario, the goal state derived while processing
the progressivized VP is maintained in the final model. In conclusion, whereas (5) involves
an extension of the initial discourse model, (6) might induce a recomputation. Since (7) and
(8) do not involve a canonical goal, they will require an extension only.

1.4. Predictions for ERPs. The only difference between neutral and disabled activities (e.g.
‘writing letters’) is the noun in the subordinate clause, ‘tablecloth’ or ‘paper’. In both cases
the initial model is simply extended, thus we expect to observe only local ERP differences
related to the integration of the differing nouns. As ‘tablecloth’ is less semantically expected
in the context of the other lexical items occurring in the sentence compared to ‘paper’, we
expect a larger N400 for the former compared to the latter word.
Processing a 'when' clause following an accomplishment (e.g. 'writing a letter') involves integrating the differing nouns and, in the disabling case, recomputing the initial discourse representation. Also in this case, the neutral noun ‘tablecloth’ is predicted to evoke a larger N400 compared to the disabling ‘paper’, reflecting a lower degree of semantic relatedness with the preceding context. In our ERP study Dutch materials were used, where the verb in subordinate clauses occupies the sentence-final position (see 2.1). The temporal and causal information provided by verbs in ‘when’ clauses is necessary to initiate the recomputation process (Baggio & van Lambalgen 2007). Thus, the ERP effects of what we have analyzed as recomputation are expected to surface at the sentence-final verb ‘spilled’ (‘morste’ in our Dutch stimuli, see 2.1 and Table 1).

One additional prediction is that the amplitude of the ERP effect evoked by disabled accomplishments is correlated with the the frequency with which readers infer that the goal state was not attained. Recomputation is expected to evoke an ERP shift in each trial in which a negative judgment concerning the attainment of the goal is made. Therefore, the larger the number of such inferences, that is, the larger the number of trials in which recomputation occurred, the larger the amplitude of the ERP component. The method and results of an ERP study in which these predictions were tested are described below.

2. METHOD

2.1. Materials. The set of Dutch materials used in the experiment included 160 test and 160 filler items. Each test item included two context sentences providing a neutral, obstacle-free setting for the events narrated, four target sentences (A)-(D), and two probe pairs (Table 1). Target sentences were constructed by manipulating the aspectual class of the progressive VP (activity or accomplishment) and the effects of the event introduced by the ‘when’ clause (neutral or disabling) on the event described in the main progressive clause. All progressive VPs were instances of the Dutch periphrastic ‘was/waren NP aan het V_{inf}’ construction. This solution is to be preferred to the use of the Dutch simple past which, in some cases, is aspectually ambiguous between perfective and imperfective readings. Accomplishments differed from activities in the object NP only: an indefinite (‘een brief’/’a letter’) was used for accomplishments, a bare plural (‘brieven’/’letters’) for activities. Disabling and neutral subordinate clauses differed only in the prepositional or object NP, for instance ‘papier’ and
Neutral and disabling events were distinguished based on the experimenters’ judgment (but see 2.2.2 for some data supporting these choices). Probe pairs (E) were used with activities and (F) with accomplishments.

Fillers were 160 sentences of varying length, structure and content. Analogously to test items, fillers were preceded by two neutral context sentences and followed by a probe pair. Target sentences described an event consistently, as in (9), or inconsistently, as in (10), with factual knowledge (see Hagoort et al. (2004) for an experiment based on these stimuli):

(9) Dutch trains are white and very crowded.
(10) Dutch trains are yellow and very crowded.

Probes were of the type ‘Trains in the Netherlands are white.’/‘Trains in the Netherlands are yellow.’ These fillers were chosen to add variety to the materials while preserving the task used for test items.

Four test versions were constructed, comprising randomized lists of test and filler items. The task was identical for critical and filler sentences. Participants had to select the correct probe based on the information provided by the context and target items. Mean length, raw and lemma frequency of the differing nouns in the NP of subordinate clauses were matched using the CELEX Dutch corpus (Baayen et al. 1996). Mean length was 7.9 letters (SD=2.46) for neutral and 7.75 (SD=2.79) for disabled cases, and was kept below 12 letters in any case. Raw frequency per million words was 1113 (SD=2462) for neutral and 1096 (SD=2792) for disabled cases. Lemma frequency per million words was 1730 (SD=3559) for neutral and 1666 (SD=3585) for disabled cases. The length of sentence-final verbs was identical across conditions and was kept below 12 letters in any case. Cloze probabilities of sentence-final verbs were normed in a dedicated pre-test discussed below.

2.2. Pre-tests.

2.2.1. Cloze probability test. In order to determine the cloze probabilities of sentence-final verbs, context sentences followed by a target sentence with the final word blanked were presented to a group of thirty-two native speakers of Dutch (mean age 22.5, 27 female). Participants were requested to fill in the blank with the first word that came to their mind.

Four versions (40 items per condition), randomized and balanced across conditions, were
constructed. Mean cloze probabilities were not different between the conditions (all comparisons using T-tests, $P > 0.05$) in each test version as well as in the entire set.

2.2.2. Entailment questionnaire. A paper-and-pencil judgment task was also administered. Thirty six Dutch native speakers (mean age 22.5, 28 female) were presented with the context followed by a target sentence and a probe pair. The task was to select the appropriate probe. Negative probes were more frequently chosen for disabled accomplishments than for the other conditions. Neutral activities (SI) (see Table 1) showed the lowest mean of negative responses ($M=2.72$, $SD=3.22$), followed by disabled activities (B) ($M=8.06$, $SD=7.05$), neutral accomplishments (C) ($M=10.03$, $SD=9.23$) and disabled accomplishments (D) ($M=25.14$, $SD=8.02$) (see Baggio & van Lambalgen (2007) for details).

2.3. Participants. Thirty one students participated in the ERP experiment. Of these, 7 were left out of the final analysis due to a high number (>20%) of trials contaminated by artifacts. This left us with twenty four participants (mean age 22.5, 17 female), with no history of neurological, psychiatric or cognitive disorders. Subjects were selected from the database of the F.C. Donders Centre for Cognitive Neuroimaging at the Radboud University Nijmegen. Participants received € 8 per hour or course credits. None of the subjects who took part to the pre-tests participated in the ERP experiment.

2.4. Procedure. After applying the electrodes (see 2.5), participants were conducted into the experimental room and were asked to sit in front of a video monitor. The stimuli were presented as follows: the two context sentences were displayed together on a single screen (white on black background) for a variable duration (6, 7 or 8 s), depending on the length of the sentences themselves; next the target sentence, one of (A)-(D), was presented on the screen word-by-word (600 ms SOA, 300 ms word duration; white on black background); the target sentence was preceded and followed by a fixation cross, presented for 1500 ms; finally, the probe pair, one of (E)-(F), was shown on the screen (red on black background) and remained visible until the participant gave a button-press; the probes were followed by a fixation cross which lasted for 1500 ms. The same presentation parameters were used for fillers. Participants were instructed to read each sentence carefully, to blink only when the fixation cross was shown and to select the correct probe by pressing one of two buttons.
(left or right on the button box) as quickly and accurately as possible. The position on the screen (top or bottom) of the positive and negative probe corresponded to the left and right button respectively, and was counterbalanced across test versions. In this way, participants could not prepare their motor response before the probe pair was presented on the screen. The experiment took about 2 hours and was divided into 24 blocks of 10 trials each.

2.5. **Recording.** EEG and EOG signals were recorded using Ag/AgCl electrodes. The EOG was measured from 4 electrodes: one at the outer canthus of each eye, one below and one above the left eye (FE). The EEG was measured from 28 electrodes, arranged according to American Electrophysiological Society conventions: FE, FF, F7, F3, Fz, F4, F8, FC5, FC1, FCz, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CE, CF, CP6, P7, P3, Pz, P4, P8, O1, O2. Two additional electrodes were placed on the left and right mastoids, the former serving as the reference during the measurement. All EEG and EOG electrodes were re-referenced off-line to a linked mastoid. EEG electrodes were attached to an elastic cap, whereas EOG and reference electrodes were applied using two-sided adhesive decals external to the cap. Electrode impedance was kept below 5 k\(\Omega\) throughout the experiment. The EEG/EOG was amplified by a multichannel BrainAmp DC system, with a 500 Hz sampling rate, a low pass filter set at 125 Hz and a 10 s time constant.

2.6. **Data Analysis.** Data analysis was conducted using FieldTrip,\(^1\) a MATLAB package for processing EEG signals. The following transforms were applied to each subject’s dataset. Segments corresponding to the noun and the sentence-final verb were extracted from the EEG with an interval of 200 ms preceding and 800 ms following stimulus onset. Baseline correction used the 200 ms interval preceding the onset of nouns, and the 100 ms interval following the onset of sentence-final verbs. The latter choice was effected so as to prevent ERP differences in the 400-600 ms interval following the onset of the nouns from biasing the baseline correction for the ERPs evoked by sentence-final verbs in the same time interval. The use of such a baseline seems acceptable on grounds that the expected recomputation effect at the verb would not affect such largely exogenous components as the N1. Artifact rejection was based on two FieldTrip functions: the first detects and rejects all trials that contain activity exceeding a threshold of ±100 \(\mu V\); the second identifies and discards trials.

\(^{1}\)For more information, see [http://www2.ru.nl/fcdonders/fieldtrip/](http://www2.ru.nl/fcdonders/fieldtrip/)
contaminated with eye movements or blinks by means of thresholding the $z$-transformed
value of the raw data in the EOG channels, preprocessed using a band-pass filter of 1-15 Hz.
A 30 Hz low-pass digital filter was applied to the segmented, artifact-free data. ERPs were
obtained for each subject by averaging over trials in each experimental condition. A 5 Hz
low-pass filter was used to produce the waveforms shown in Figures 2-5. Topographical
plots and statistical analyses were however based on the 30 Hz low-pass filtered data.

For the analysis of behavioral responses we employed two repeated-measures ANOVA
models with Subject as the random effect, Aspectual Class (Activity/Accomplishment) and
Subordinate Clause Type (Neutral/Disabling) as fixed effects, and the mean value of either
negative judgments (negative probes selected in the response task) or decision times in each
condition as the dependent variables.

Statistical analyses of ERP data used a non-parametric randomization procedure (Maris
2004; Maris & Oostenveld 2007) which took as input mean amplitude ($\mu V$) values in each
condition in time bins of 100 ms, starting from the onset of the relevant word and ending
800 ms after, and produced as output a cluster of electrodes (min. 1, max. 28) in which the
difference between the conditions was significant in each time bin, the sum of $T$-statistics
in that cluster and Monte Carlo estimates of $P$-values.

For the correlation analysis (see Section 1.4), we calculated the difference between the
ERPs evoked by sentence-final verbs in subordinate clauses – disabled (D) minus neutral
(C) – following accomplishments at anterior sites (FE, FF, F7, F3, Fz, F4, F8 averaged) in
the 500-700 ms interval after the onset of the sentence-final verb (see 3.2 for motivation).
Pearson’s product-moment correlation was computed to determine whether the amplitude
difference in ERPs varied with the number of negative responses, quantified again as the
difference of negative judgments between disabled (D) and neutral (C) accomplishments.
The correlation analysis was done on a per-subject basis (i.e. each pair of data points in the
correlation corresponded to a single subject’s data).

3. Results

3.1. Behavioral data. Neutral activities (S1) showed the lowest mean of negative responses
(M=4.08, SD=2.87), followed by disabled activities (B) (M=5.83, SD=4.51), neutral accom-
plishments (C) (M=9.58, SD=9.96) and disabled accomplishments (D) (M=18.13, SD=11.16).
The distribution of the data in the different conditions appears rather similar, as indicated by box height and whisker length in Figure 1a. However, disabled accomplishments have a more spread-out distribution, suggesting that inference patterns were less uniform across participants. ANOVAs revealed significant main effects of Aspectual Class and Subordinate Clause Type, and a significant interaction between the two factors (Table 2, Figure 1a). The observed pattern of responses supports the linguistic views outlined above and replicates our previous findings (see 2.2.2). There is no difference in decision times (Table 2, Figure 1b): (S1), M=2111 ms, SD=677 ms; (B), M=2100 ms, SD=680 ms; (C), M=2070 ms, SD=646 ms; (D), M=2086 ms, SD=710 ms.

3.2. Event-related brain potentials.

3.2.1. Nouns. Figure 2 displays the ERP topographies and waveforms elicited by activities. An N1 component peaking at approximately 100 ms is followed by a F component with a trough at about 200 ms. The amplitude of the N1 and F is not different between disabled and neutral clauses: no significant clusters were found between 0 and 300 ms (all contrasts, \( P > 0.1 \)). The N1-F complex is followed by an N400. The amplitude of the N400 is larger in neutral (‘tafelkleed’) than in disabling (‘papier’) clauses (Figure 2b): significant clusters with a central distribution were found between 300 and 500 ms (Table 3, Figure 2a). No difference between neutral and disabling clauses was found after 500 ms.

Figure 3 displays the ERP topographies and waveforms elicited by accomplishments. Also in this case, an N1-F complex can be observed. There is no difference between neutral and disabling clauses, as no significant clusters between 0 and 300 ms were found (all contrasts, \( P > 0.1 \)). The amplitude of the N400 is again larger in neutral (‘tafelkleed’) than in disabling (‘papier’) clauses (Figure 3b). The effect lasts longer than the N400 observed in activities: significant clusters with a central distribution were found between 300 and 600 ms (Table 3, Figure 3a). No difference between conditions was found after 500 ms.

There is no overall difference between the two aspectual classes. Cluster-based \( T \)-tests comparing the N400 effects in activities and accomplishments (corresponding to testing the main effect of Aspectual Class in a parametric model) produced no significant clusters between 300 and 600 ms from noun onset (all contrasts, \( P > 0.1 \)). No difference was found in any of the remaining time bins.
3.2.2. Sentence-final verbs. Figure 4 displays the ERP topographies and waveforms elicited by activities. Contrary to what we had observed at the noun, there is no difference between the N400 elicited by neutral and disabling clauses. Moreover, there is no difference between conditions in any of the remaining time bins (Table 3).

Figure 5 displays the ERP topographies and waveforms elicited by accomplishments. No difference between disabling and neutral clauses was observed in either the N1-F complex or the N400: no significant clusters between 0 and 400 ms were found (all contrasts, \( P > 0.1 \)). While disabled and neutral activities do not result in any robust differential effect in later time bins (400-800 ms, Table 3, Figure 4), disabling verbs following accomplishments evoked larger negative shifts compared to neutral verbs (Table 3, Figure 5). The effect emerges at about 400 ms following the onset of sentence-final verb, lasts for approximately 400 ms, and is larger over the more anterior scalp sites, in particular of the left hemisphere. Based on its temporal profile and scalp distribution, we take this effect to be an instance of sustained anterior negativity (SAN). The magnitude of the SAN effect is correlated with the frequency of negative judgments in the response task \((r = -0.415, T(22) = -2.140, P = 0.043; \text{Figure 6})\): the higher the number of negative responses, the larger the amplitude of the SAN.

No difference between the two aspectual classes was found. Cluster-based \( T \)-statistics comparing mean ERP amplitudes in activities and accomplishments, again corresponding to testing the main effect of Aspectual Class in a parametric model, produced no significant clusters between 0 and 800 ms from noun onset (all contrasts, \( P > 0.1 \)).

4. DISCUSSION

The ERP results reported above can be summarized as follows. The N400 elicited by nouns is larger in neutral than in disabling clauses, following both activities and accomplishments. This can be explained by the lower degree of semantic association with the preceding words (‘writing’, ‘letter’ or ‘letters’) of the noun in neutral clauses (‘tablecloth’) compared to the noun in disabling clauses (‘paper’). On the basis of our processing model, we predicted that disabled accomplishments would induce a different ERP response at the sentence-final verb compared to neutral accomplishments. This corresponds to the difference between the recomputation and the extension of the initial discourse model (see Section 1.3). The effect was expected to be (i) absent in activities and (ii) correlated with the frequency with which
participants infer that the goal state was not attained. These predictions were borne out. Disabled activities did not modulate ERPs at the verb. Disabled accomplishments evoked sustained anterior negativities (SANs). Moreover, a correlation of the SAN amplitude with the frequency of negative judgments was observed. Taken together, these results seem to offer some support for the recomputation hypothesis. Below we address a few alternative explanations of the data and some related outstanding issues.

4.1. Alternative explanations and outstanding issues.

4.1.1. Local integration. An alternative account of the data would relate the observed effect to difficulty in integrating the sentence-final verb into the ‘local’, clause-level context, rather than to suppressing a ‘global’, discourse-level inference. If this were correct, a modulation of the N400 should be expected, possibly correlated with differences in cloze probabilities. However, as reported above, cloze probabilities do not differ between conditions (see 2.2.1). Also, there was no difference in the N400s elicited by sentence-final verbs (see 3.2), which were moreover lexically identical across conditions. Following earlier proposals (Oosterhout 1997), we see the sustained anterior negativity as reflecting difficulty in constructing a discourse-level representation of disabled accomplishments. Supported by further experimental evidence, the recomputation hypothesis could provide a more explicit characterization of at least one instance of sentence-final ‘wrap-up effects’, in terms of restructuring the initial model.

4.1.2. Response frequency. Another alternative account would be based on the observation that sentences requiring a negative response (disabled accomplishments) are relatively less frequent than sentences requiring a positive one (activities and neutral accomplishments), the projected ratios being respectively 1/4 and 3/4 (see 3.1 for the actual behavioral data). On this view, a modulation of the P3 component (Donchin 1981; Ruchkin et al. 1990) might be expected, inversely correlated with the frequency of negative judgments given to (D): the less frequent the negative responses, the larger the amplitude of the P3. However, in our experiment no P3 response was observed and, moreover, the correlation was rather the inverse: the more frequent the negative responses, the larger the amplitude of the sustained anterior negativity.
4.1.3. Monotonicity and possible worlds semantics. An important issue is whether the observed sustained anterior negativity can be explained by a monotonic account of the progressive. Further, it may be asked whether the data reported here constitute compelling evidence for non-monotonicity and against monotonicity. One such alternative explanation can be formulated in possible worlds semantics (Kripke 1963; Dowty 1979).

In possible worlds semantics, the progressive denotes a stage of a process which, if it does not continue in the actual world, has chances of continuing in some other possible world (de Swart 1998). The latter may be called ‘inertia worlds’, courses of events in which the process is not disturbed by external forces and is therefore brought to a successful end. In his analysis of the progressive, Dowty (1979) claimed that the following are equivalent:

1. ‘The girl is writing a letter’ is true in the actual world;
2. ‘The girl will have written a letter’ is true in all so-called ‘inertia worlds’, worlds which are identical with the present world until ‘now’, but then continue in a way most compatible with the history of the world until ‘now’.

These insights can be rendered into processing terms. Processing neutral accomplishments involves moving from the actual world, a snapshot of which is provided by the progressive clause, to some inertia world, in which the goal state is eventually attained (the behavioral data reported in 3.1 show that such an inference is drawn). By contrast, processing disabled accomplishments amounts to proceeding from the actual world to a relatively unexpected ‘non-inertia world’, in which the process is disrupted by some event, such as spilling coffee on the paper. Accessing a world in which the goal state is not attained may be surprising. The sustained anterior negativity may then be construed as an index of surprise or some other equivalent notion. This account is monotonic, as for both neutral and disabled accomplishments interpretation involves shifting from the actual world to another accessible world. A simple extension of the initial model is performed in both cases.

This analysis is seemingly in conflict with the non-monotonic one. Still, there is no real opposition between the two as regards these data. In both accounts, an initial commitment to the occurrence of the goal state is made. In the non-monotonic approach this takes the

---

2This particular combination of possible worlds semantics and surprise was proposed by an anonymous reviewer of this paper, hence our choice of discussing it here. The same applies (for different reviewers) to the issues discussed in section 4.1.4 and 4.1.5.
form of a defeasible inference leading to a minimal model, whereas in the monotonic theory
it is rather a prior, positive expectation concerning the attainment of the consequent state,
and thus a lower probability assigned to its failure. This commitment is necessary, because
accessing a possible world in which the goal state is not satisfied can be surprising only
if there is such a prior expectation. Moreover, in both accounts a subsequent revision of
the initial commitment is made. In the non-monotonic analysis this is a recomputation of
the minimal model, whereas in the monotonic one it is a recomputation of the initial (low)
probability associated with the possible world in which the goal state is not attained.

So there is recomputation and non-monotonicity built into this ‘monotonic’ theory too.
The possible worlds account is monotonic as far as models are concerned (models are always
extended, and never recomputed), but expectations change non-monotonically (probabilities
are recomputed). While prima facie opposed, the two accounts are in this respect similar.
Our reasons for preferring a non-monotonic account, one in which models are recomputed,
derive from a priori considerations. First, there are forms of non-monotonic inference for-
mally strongly related the non-monotonic reasoning in the progressive which cannot be
captured by Bayesian updates (Stenning & van Lambalgen 2008). Second, minimal models
and non-monotonic inference can be implemented in neural networks (see 4.2 below). Last,
our account of the progressive is embedded into a larger non-monotonic framework (van
Lambalgen & Hamm 2004; Stenning & van Lambalgen 2008), covering other phenomena
in reasoning and language processing in children, adults, as well as patients with ADHD
(Attention Deficit Hyperactivity Disorder) (van Lambalgen et al. 2008) and ASD (Autistic

We must note however that our study was designed to test a particular non-monotonic
theory of the progressive, and not to discriminate between monotonic and non-monotonic
accounts of the same phenomenon. The latter task would require, for one, a well-specified
entirely monotonic theory – that is, one which does not involve recomputation of models,
probability values or other processing parameters – and, moreover, a set of predictions in
which the two proposals would actually differ. This is admittedly hard, apart from being
beyond the scope of the research reported here. Hence the need to emphasize the direction
along the theory-observation path which is relevant here: although it can be argued that our
non-monotonic theory leads to predictions that are consistent with the observed sustained
anterior negativity, it is clearly not the case that the data support only this particular theory.

4.1.4. 

**Interruption and termination of activities.** It may be argued that, compared to disabled
accomplishments, disabled activities are inherently simpler because they involve at most
an *interruption* of the activity, for example writing letters in (B), which may be continued on
some other paper sheets. Accomplishments might leave a more definite ERP trace because
they lead to the *termination* of the activity, for example writing a letter in (D), which cannot
be continued being there only a single sheet. On this view, the sustained anterior negativity
would not be related to model recomputation (as opposed to monotonic extension), but to
the termination (as opposed to the interruption) of the activity. Such an explanation follows
from the seemingly plausible notion that computing a model in which the effects of a given
event are more ‘catastrophic’ should also be more difficult. Here semantic theory comes to
our rescue and suggests that such notion is in fact misguided.

One issue that plays a role here is a type/token distinction concerning the noun ‘letter’.
In the token interpretation of ‘letter’ as referring to some particular scribbles on a particular
piece of paper, there is indeed a difference between interruption and termination. However,
on a type interpretation of ‘letter’ as referring to particular content which can be inscribed
on any piece of paper, the activity and the accomplishment case seem comparable, in that in
both cases the girl has to reach for a new piece of paper. On the type reading, one wouldn’t
even expect a difference in behavioral responses. Nonetheless, since a behavioral difference
was observed, it seems the token reading is what subjects adopt. On this assumption, it can
be shown that, contrary to the alternative proposal, there is *more* computation going on in
the interruption case compared to the termination case – if goal states are not taken into
account; if they are, the pattern is reversed as implied by the recomputation hypothesis. It
seems harder to compute a model in which an activity is first interrupted, then re-initiated,
compared to computing a model in which the activity is just terminated (van Lambalgen &
Hamm 2004). The alternative account would predict a larger sustained anterior negativity
for activities compared to accomplishments, which does not fit the experimental results.
Also in this case, however, we are ready to acknowledge that a different model, in which
terminations are shown to be *more* costly than interruptions, and in which goal states are not
4.1.5. *Goal states and underspecification.* The processing model adopted here implies that, as soon as an accomplishment in the past progressive is encountered, the system constructs a semantic representation in which the goal state is satisfied. Processing the clause ‘The girl was writing a letter’ amounts to computing a minimal model in which the writing activity leads to a complete letter, which is therefore part of the resulting discourse structure. As we have hypothesized, such computation is defeasible, that is, the model can be recomputed if further discourse information implies that the goal state is not satisfiable, as in (6). One may ask whether the claim that the goal is part of a minimal model of the progressive clause is at all tenable. A seemingly more plausible account would assume that an underspecified model, in which it is left undetermined whether the goal state is attained or not, is computed while the progressive clause is processed, and a decision is made only at the subordinate clause.

The main problem with an underspecification-based account is that, while it is true that the information provided by the progressive clause is insufficient for determining whether the goal was attained (which would motivate the construction of an underspecified model at that stage), it is *not* the case that sufficient information is contributed by the subordinate clause. While disabling clauses provide evidence that the activity was terminated, and thus license the inference that the goal was not attained, no evidence concerning the satisfaction of the goal state is derivable from neutral clauses. This is a consequence of the well-known ‘frame problem’ (McCarthy & Hayes 1969), which implies that it is impossible to enumerate all the effects and non-effects of an event. For example, that ‘spilling coffee on the tablecloth does not affect the writing activity’ (if that is the case) is not stored in declarative memory, but must be inferred. This is an instance of ‘closed world reasoning’, which was described above (see 1.3). In a ‘closed world’, it is assumed that no obstacle to attaining the goal state occurred. Therefore, a letter was completed. The behavioral data reported above show that subjects draw this inference or, equivalently, they are more likely to give positive responses to neutral accomplishments. Processing models based on underspecification – or on parallel processing, for that matter – would have to explain why that very same conclusion (‘the girl wrote a letter’) is not drawn when the system is faced with the relevant input (the VP in the
progressive), and is instead delayed until the end of the sentence, where critical information
is nonetheless still missing. The hypothesis that the goal state inference is drawn when the
input is given seems to be more consistent with the available evidence on immediacy and
incrementality in discourse processing (Hagoort & van Berkum 2007).

This line of reasoning speaks also to the issue of the potential influence of the primary
response task on on-line interpretive processes. It can be argued that the system may have
carried out a number of inferences on-line in order to facilitate a response when the probes
were presented, but would have processed the same sentences in an underspecified manner
if no response task was administered. The brain would therefore compute representations
which are merely ‘good enough’ for the task at hand, striking a balance between efficiency
and cost minimization (Ferreira et al. 2002; Ferreira & Patson 2007; Douglas & Martin 2007).

We grant that this is a possibility, which cannot be excluded based on either our data or our
processing model. It can however be suggested that, although comprehension probes do
not occur in actual language use, it is possible to imagine ‘language games’ in which hearers
are required to make interpretive commitments and form a belief concerning the potential
outcomes of a process described using the progressive. Our experiment may be taken as a
laboratory study of such real world situations, but is not intended as a realistic account of
all situations in which progressive constructions are uttered and understood. Further work
is needed to investigate the influence of the response task on on-line ERP measures.

4.2. Recomputation in working memory networks. Minimal models can be regarded as
the stable states of associated neural networks. It has been shown that recurrent networks
can compute or approximate (depending on the expressiveness of the logical formalism)
the semantic operators based on which minimal models are constructed (Hitzler et al. 2004;
Stenning & van Lambalgen 2005; Stenning & van Lambalgen 2008). In this framework,
recomputation can be modeled as the readjustment of connection strengths driven by a
simple form of back-propagation called ‘perceptron learning’ (Rosenblatt 1962). Computing
a minimal model of the progressive clause will correspond to the network settling into one
such ‘attractor’ or stable state. Further computation on the initial model brings the network
from its initial stable state to another stable state, corresponding to the new minimal model.
Importantly, there is a large difference in the overall pattern of network activity in disabled
compared to neutral accomplishments. If the initial model is monotonically extended, as in the neutral case (4), a number of units will be activated which were previously silent, while the activation state of the remaining units, including those representing the goal state (the complete letter), will remain unaltered. But if the initial minimal model is recomputed upon encountering the subordinate clause in (6), units which were silent will be activated and the activation patterns across some units which were previously active will be readjusted. For instance, the units representing the goal state (the complete letter) will no longer be active. In the neural network this is achieved by successive applications of perceptron learning.

Even though in both cases the network processes the subordinate clause by settling into a new attractor state, the transition in the disabling case requires an extensive adjustment of the connection weights of the units representing the goal state. Recomputation thus results in a more costly state transition. It remains an open question whether biologically plausible networks can also approximate the semantic operators which give rise to minimal models. Firing rate models, for instance, have been used to implement operations in connectionist networks (e.g. multilayer perceptrons) of the kind required by the construction of minimal models (de Kamps & van der Velde 2002). Interestingly, recurrent excitation in firing-rate models can account for several aspects of persistent activity in prefrontal cortex neurons during working memory tasks (Durstewitz et al. 2000). Recurrent networks thus suggest a plausible mechanistic link between recomputation and sustained anterior negativities, and in general between working memory processes and sustained anterior negativities (King & Kutas 1995; Müller et al. 1997; Münte et al. 1998; van Berkum et al. 1999; Fiebach et al. 2002; Felser et al. 2003; van Berkum et al. 2003; Phillips et al. 2005).

As we noted in the introduction, a cognitive neuroscience of language needs to bridge the gap between psycholinguistic and formal models of specific aspects of language on the one hand, and the neural architecture underlying neurophysiological measures on the other hand. For a number of reasons (Poeppel & Embick 2005) this is a daunting task, which we do not claim to have adequately solved. However, tentatively the following can be said. There is no indication or proof that the sustained anterior negativity is a language-specific ERP effect. Most likely, it reflects the recruitment of neurophysiological activity that might be generated in prefrontal cortex, and is triggered by different cognitive operations which
build upon working memory capacity. For this purpose, the prefrontal cortex is a plausible
candidate from a neurobiological point of view. In the light of our model, the sustained
anterior negativity is taken to index the recomputation following the blocking of the goal
state in accomplishments, and the recruitment of working memory resources required for
this recomputation. In other cases, the demand might be triggered by different cognitive
operations, as in the work by Münte et al. (1998). In general, what we seem to obtain with
ERPs, is a many-to-one mapping from cognitive models to neuronal implementation. This
however in no way invalidates our interpretation, which is based on combined constraints
from the cognitive and neuronal levels of analysis.

The research presented in this paper extends the range of phenomena to which ERPs
can be applied, by testing a processing hypothesis which derives from a formal semantics
of tense and aspect. This open the way to combining formal and philosophical theories of
meaning with experimental data as made available by cognitive neuroscience techniques.
Our ERP results raise the possibility that the brain supports some form of non-monotonic
recomputation to integrate information which invalidates previously held assumptions. It
is a task for future research to provide more stringent tests of monotonic vs. non-monotonic
models of semantic processing and cognitive update more generally.

Acknowledgments. We wish to thank Oliver Bott, Travis Choma, Bart Geurts, Fritz Hamm,
Karl Magnus Petersson, Jos van Berkum, Roel Willems and three anonymous reviewers for
many useful comments on earlier versions of the paper. We are grateful to the Netherlands
Organization for Scientific Research (NWO) for support under grant 051.04.040.
References


Context sentences

De deur van de woonkamer was gesloten. Binnen speelde de radio klassieke muziek.

*The door of the living-room was closed. Inside played the radio classical music.*

Target sentences

(A) Het meisje was brieven aan het schrijven toen haar vriendin koffie op het tafelkleed morste.

*The girl was letters on the to-write when her friend spilled coffee on the tablecloth.*

(B) Het meisje was brieven aan het schrijven toen haar vriendin koffie op het papier morste.

*The girl was letters on the to-write when her friend spilled coffee on the paper.*

(C) Het meisje was een brief aan het schrijven toen haar vriendin koffie op het tafelkleed morste.

*The girl was a letter on the to-write when her friend spilled coffee on the tablecloth.*

(D) Het meisje was een brief aan het schrijven toen haar vriendin koffie op het papier morste.

*The girl was a letter on the to-write when her friend spilled coffee on the paper.*

Probe sentences

(E) Het meisje heeft een of meer brieven geschreven.

*The girl has one or more letters written.*

(F) Het meisje heeft een brief geschreven.

*The girl has a letter written.*

Table 1. Examples of stimulus sentences. *a*Literal translation. *b*Paraphrase.
<table>
<thead>
<tr>
<th></th>
<th>Categorical responses</th>
<th>Decision times</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspectual Class</td>
<td>$F(1, 23) = 21.65$</td>
<td>$F(1, 23) &lt; 1$</td>
</tr>
<tr>
<td></td>
<td>$P &lt; 0.001$</td>
<td></td>
</tr>
<tr>
<td>Subordinate Clause Type</td>
<td>$F(1, 23) = 23.60$</td>
<td>$F(1, 23) &lt; 1$</td>
</tr>
<tr>
<td></td>
<td>$P &lt; 0.001$</td>
<td></td>
</tr>
<tr>
<td>Aspectual Class × Subordinate</td>
<td>$F(1, 23) = 17.20$</td>
<td>$F(1, 23) &lt; 1$</td>
</tr>
<tr>
<td>Clause Type</td>
<td>$P &lt; 0.001$</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 2.** Summary of ANOVA statistics for behavioral data.
<table>
<thead>
<tr>
<th>Time</th>
<th>Noun</th>
<th>Sentence-final verb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Activities</td>
<td>Accomplishments</td>
</tr>
<tr>
<td>300-400 ms</td>
<td>$T(22) = -16.54$</td>
<td>$T(22) = -60.11$</td>
</tr>
<tr>
<td></td>
<td>$P = 0.026$</td>
<td>$P &lt; 0.001$</td>
</tr>
<tr>
<td>400-500 ms</td>
<td>$T(22) = -57.02$</td>
<td>$T(22) = -78.69$</td>
</tr>
<tr>
<td></td>
<td>$P &lt; 0.001$</td>
<td>$P = 0.002$</td>
</tr>
<tr>
<td>500-600 ms</td>
<td>$T(22) = -18.58$</td>
<td>$T(22) = -71.09$</td>
</tr>
<tr>
<td></td>
<td>$P = 0.022$</td>
<td>$P &lt; 0.001$</td>
</tr>
<tr>
<td>600-700 ms</td>
<td></td>
<td>$T(22) = -39.16$</td>
</tr>
<tr>
<td>700-800 ms</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** Summary of cluster-based statistics for the ERP data. Disabling and neutral clauses are compared at the noun and at the sentence-final verb, for activities and accomplishments, in bins of 100 ms starting from word onset. The first significant effects occurred at 300-400 ms. Empty cells denote the absence of significant clusters.
Figure 1. Behavioral data. (a) Boxplot of the categorical responses. (b) Boxplot of the decision times. Conditions are represented on the abscissa (see Table 1 for the correspondences with the labels). Negative responses and decision times are plotted on the ordinate. The solid line within the boxes indicates the median, box height is equal to the interquartile range, whiskers represent adjacent values, and empty circles denote outliers. The maximum of potential negative responses is 40.
Figure 2. Activities, noun. (a) Grand-average (N=24) topographies displaying the mean amplitude difference between the ERPs evoked by the noun in neutral compared to disabled activities. Circles represent electrodes in a significant cluster. (b) Grand-average (N=24) ERP waveforms from frontal, central and parietal electrode sites time locked to the onset (0 ms) of the noun in neutral and disabled activities. Negative values are plotted upward.
FIGURE 3. Accomplishments, noun. (a) Grand-average (N=24) topographies displaying the mean amplitude difference between the ERPs evoked by the noun in neutral compared to disabled accomplishments. Circles represent electrodes in a significant cluster. (b) Grand-average (N=24) ERP waveforms from frontal, central and parietal electrode sites time locked to the onset (0 ms) of the noun in neutral and disabled accomplishments. Negative values are plotted upward.
Figure 4. Activities, sentence-final verb. (a) Grand-average (N=24) topographies displaying the mean amplitude difference between the ERPs evoked by the sentence-final verb in disabled compared to neutral activities. (b) Grand-average (N=24) ERP waveforms from frontal, central and parietal electrode sites time locked to the onset (0 ms) of the verb in neutral and disabled activities. Negative values are plotted upward.
**FIGURE 5.** Accomplishments, sentence-final verb. (a) Grand-average (N=24) topographies displaying the mean amplitude difference between the ERPs evoked by the sentence-final verb in disabled compared to neutral accomplishments. Circles represent electrodes in a significant cluster.

(b) Grand-average (N=24) ERP waveforms from frontal, central and parietal electrode sites time locked to the onset (0 ms) of the verb in neutral and disabled accomplishments. Negative values are plotted upward.

(C) Het meisje was een brief aan het schrijven toen haar vriendin koffie op het tafelkleed morste. (Accomplishment, neutral)

'The girl was writing a letter when her friend spilled coffee on the tablecloth.'

(D) Het meisje was een brief aan het schrijven toen haar vriendin koffie op het papier morste. (Accomplishment, disabling)

'The girl was writing a letter when her friend spilled coffee on the paper.'
FIGURE 6. Scatter plot displaying the correlation between the amplitude of the sustained anterior negativity elicited by disabled accomplishments and the frequency of negative responses ($r = -0.415, T(22) = -2.140, P = 0.043$). The mean difference of negative responses between disabled and neutral accomplishments is plotted on the abscissa. The mean amplitude difference at fronto-polar and frontal electrodes between disabled and neutral accomplishments in the 500-700 ms interval following the onset of the sentence-final verb is plotted on the ordinate.