Philipps Universität Marburg

Quantity and Quality: Not a Zero-Sum Game.

A computational and neurocognitive examination of human language processing

Inaugural-Dissertation

zur Erlangung der Doktorwürde

des Fachbereichs
Germanistik und Kunstwissenschaften
der Philipps-Universität Marburg

vorgelegt von

Phillip Marshel Alday
aus Raleigh, North Carolina, USA

Marburg an der Lahn, 2015

Vom Fachbereich Germanistik und Kunstwissenschaften der Philipps-Universität Marburg

als Dissertation angenommen am: 12. Februar 2015

Tag der Disputation: 18. März 2015

Gutachter:

Prof. Dr. Ina Bornkessel-Schlesewsky
Prof. Dr. Dr. Joakim Nivre
Prof. Dr. Richard Lewis

for Dr. Elizabeth Hall and Dr. Denise Della Rossa,
who knew that I could never leave language completely behind

Acknowledgements

nunc est bibendum

Horace

A dissertation is somehow like a child – you spend years worrying about little details, struggling through the day to day, and then, suddenly, it's time to send the not-so-little one out into the world. And much like a child, it takes a village to 'raise' one.

I would like to thank my advisor and *Doktormutter*, **Ina Bornkessel-Schlesewsky** for her guidance and support, as well as my additional readers, **Joakim Nivre** and **Rick Lewis**, who kindly consented to reviewing such an interdisciplinary mess. Joakim has also supported, advised and taught me during my code sprints in Uppsala, and for that I'm particularly thankful.

The **Marburg University Research Academy (MARA)** through the Graduiertenzentrum Geistes- und Sozialwissenschaften kindly supported me twice in recent years, financing a two-week stay in Uppsala and a conference presentation in San Diego. **The *datorlingvistik* group at Uppsala Universitet**, and especially **Marco Kuhlmann**, were kind enough to host me for those two weeks and take time from their busy schedules to discuss my ideas on combining aspects of dependency parsing with psycholinguistic theory, and were exceedingly helpful and patient with my at-times very basic questions.

Large parts of the work presented here are based on work performed in the neurolinguistic laboratories of Philipps-Universität Marburg and Johannes Gutenberg-Universität Mainz and would not have been possible without the support of the lab personnel and student assistants. In particular, I would like to thank **Anika Jödicke**, **Brita Rietdorf** and **Greta Frank** for their help and support as well as **Fritzi Milde** for her heroic effort in annotating the natural story for several different psycholinguistic theories.

I would also like to thank my fellow *Marburger*, both present and past, who joined me in the day-to-day joys and struggles of academic life, **Karen Henrich née Bohn**, **Miriam Burk**, **Alexander Dröge**, **Sabine Frenzel**, **Katerina Kandylaki**, **Johannes Knaus**, **Franziska Kretzschmar**, **Laura Maffongelli**, **Fritzi Milde**, **Muralikrishnan**, **Elisabeth Rabs**, **Jona Sassenhagen**, **Sarah Tune** and **Fiona Weiß**.

To **Franziska Kretzschmar**, I am especially thankful for her helpful advice on manners professional and academic as well as the insightful discussions on the nature of ERP components. Along with **Harm Brouwer**, Franziska contributed invaluable feedback on my fledgling ERP

theory, and I would like to thank both of them for their help. **Jona Sassenhagen**, the sometimes victim and the sometimes cause of my neuroses, was a constant source of *Auf- und Anregung* and profoundly shaped my intellectual and personal development. **Nathan Alday** was always a good sounding board on technology and a source of good programming discussion.

Alexander Dröge, **Greta Frank**, **Barbara Güldenring**, **Johannes Knaus** and **Franziska Kretzschmar** read drafts of this disseration at various stages of its evolution. Together they helped me improve the clarity of my presentation as well as eliminate many, many typos and more than a few ungrammatical sentences.

I would not have survived the stresses of graduate school without the support of some great friends. **Oliver Schallert** (*?oui!*) and **Alexander Dröge** (*'s'up y'all*), and their respective significant others, **Lea Schäfer** and **Jaruwan Dröge née Junnum**, made sure that I never took myself too seriously and were always there in spirit(s) to support me. **Karen Henrich**, **Megan Fariello**, **Laura Maffongelli** (*ciao Mama!*), **Sonja Müller** and **Hannah Bayer** likewise watched out for me, from both near and far, and **Barbara Güldenring** was a great swim buddy. **Johannes Knaus**, a kindred soul in offically being an unofficial IT department, was always there with a software tip and open ear, and **Bettina Knaus** supported me and guided me through the trials of bureaucratic hell. **Greta Frank** has tirelessly supported me and my endeavors with love and kindness. **Adam Osborn** continues to be the best friend I could hope for.

Auch meiner deutschen Familie möchte ich danken. **Heidi und Pit Harrer** haben mich wie ihren eigenen Sohn aufgenommen, mich immer willkommen geheißen und immer geschaut, dass ein Heimatsferner trotzdem ein Zuhause hat. **Regina Schrack-Frank und Georg Frank** haben ebenfalls auf mich aufgepasst, mich gut empfangen, und sich um mich gekümmert.

Finally, I would like to thank my parents and family for putting up with my international (mis)adventures. I hope that this finally clears up what I've been doing all these years.

Phillip M. Alday

Marburg, January 2015

Contents

Acknowledgements iv

1. Introduction 2
- 1.1. The extended Argument Dependency Model (eADM) . . . 3
  - 1.1.1. Cue-based Processing . . . 3
  - 1.1.2. Beyond Behavior: Neurocognition . . . 4
- 1.2. The Goal of Modeling . . . 8
  - 1.2.1. Mechanisms and Properties . . . 8
  - 1.2.2. Levels of Abstraction . . . 9
- 1.3. The New Statistics, Parameter Estimation and Mixed-Effects Models . . . 10
- 1.4. Towards a More Precise Formulation of the Actor Heuristic . . . 11

2. Looking for Arguments That Get Stuff Done: Greediness and the Actor Strategy 12
- 2.1. Dependency Parsing . . . 15
- 2.2. Dependency Parsing as a Cognitive Representation . . . 16
- 2.3. Transition-based Parsing . . . 17
- 2.4. A First Attempt . . . 18
  - 2.4.1. Properties of Appropriate Feature Models . . . 19
  - 2.4.2. Non-Traditional Dependency Structures . . . 20
  - 2.4.3. Preliminary Results . . . 20
- 2.5. Review and Outlook . . . 22

3. Distinctness as a Numerical Quantity 24
- 3.1. Brief Summary of Methods and Results . . . 24
  - 3.1.1. Mathematical Formalisms . . . 24
  - 3.1.2. Results . . . 26
  - 3.1.3. Future Directions . . . 27
- 3.2. Relevance . . . 28
- 3.3. Publication . . . 28

4. Decisions, Decisions: Quantifying Cue Contributions 66
- 4.1. Brief Summary of Methods and Results . . . 66
- 4.2. Relevance . . . 67
- 4.3. Publication . . . 67

5. Natural Stories: New Perspectives on ERPs and the Role of Frequency 87
5.1. Brief Summary of Methods and Results . . . 87
5.1.1. Methods . . . 87
5.1.2. Results . . . 88
5.2. Relevance . . . 89
5.3. Publication . . . 90

6. The New Old Thing: Memory, Models, and Prediction 128
6.1. A Neurocomputational Proposal . . . 128
6.1.1. Supporting Assumptions and Hypotheses . . . 128
6.1.2. Central Proposal . . . 130
6.1.3. Implications . . . 133
6.2. Divergence from Existing Theories . . . 134
6.3. Relationship to Previous Computational Work on the Actor Heuristic . . . 135
6.4. Review and Outlook . . . 136

7. Conclusion 137

Bibliography 139

A. Non-Replicable LOO Results 146
A.1. Ambiguous . . . 147
A.1.1. Prominence Labels . . . 147
A.1.2. Grammatical-Relation Labels . . . 148
A.2. Unambiguous . . . 149
A.2.1. Prominence Labels . . . 149
A.2.2. Grammatical-Relation Labels . . . 150

Eidesstattliche Versicherung 151

Summary in English 152

Zusammenfassung in deutscher Sprache 153

Curriculum Vitae (English) 154

Curriculum Vitae (Deutsch) 156

We cannot describe how the mind is made without having good ways to describe complicated processes. Before computers, no languages were good for that. Piaget tried algebra and Freud tried diagrams; other psychologists used Markov Chains and matrices, but none came to much. Behaviorists, quite properly, had ceased to speak at all. Linguists flocked to formal syntax, and made progress for a time but reached a limit: transformational grammar shows the contents of the registers (so to speak), but has no way to describe what controls them. This makes it hard to say how surface speech relates to underlying designation and intent–a baby-and-bath-water situation. I prefer ideas from AI research because there we tend to seek procedural description first, which seems more appropriate for mental matters.

Marvin Minsky, *Music, Mind, and Meaning*

1. Introduction

Wir müssen wissen, wir werden wissen.

David Hilbert

In 1980, Kutas and Hillyard discovered a component of the human event-related potential (ERP) that would come to dominate the study of human language processing.[1] More than thirty years later, we still lack a standard theory of the neurocognitive mechanism behind the N400. Similar problems exist for the other major components associated with language, such as the P600.

How can it be that a well-characterized, reliable experimental phenomenon lacks a standard theory? Part of the problem is that our current theories are not precise enough to be truly falsifiable. Presented with a challenge for a favorite theory, we always have the option of weaving complicated narratives. Some of this story telling is perhaps necessary in a young field dealing with complex phenomena like cognitive neuroscience — we simply do not know enough to make the educated guesses necessary for more precise theories. However, some of this is related to the types of statistical and analytical methods we apply to our data.

From reaction times to eye tracking to EEG and fMRI, all of the major experimental methods in the study of human sentence processing deliver quantitative data with incredible precision in at least one dimension. Yet, the majority of traditional statistical tools, experimental methods and theory look for qualitative, even categorical, contrasts. When all of our precision is reduced to "greater than", "less than", or "equal to", it is not surprising that we are unable to adequately discriminate models and their predictions. We need models that make precise predictions if we want a precise description of language processing. And only quantitative models are precise.

This dissertation presents a first attempt, through a number of complementary approaches, towards quantifying and refining a model of human language processing. We begin with a short introduction to the model in question, the extended Argument Dependency Model (eADM, Bornkessel-Schlesewsky and Schlesewsky 2009, 2013), and continue with some philosophical, yet important considerations concerning the goal of modeling and role of statistics. The chapter concludes with a brief outline of the remaining chapters.

[1] The single-minded focus on this one component has led Steve Small to suggest that many papers should be published in the (non-existent) *Journal of the N400*.

1.1. The extended Argument Dependency Model (eADM)

Traditional notions of language, stemming from a philological perspective more than 2000 years old,[2] typically involve concepts such as "subject" or "object", i.e. grammatical relations based upon morphosyntactic marking. However, even such a simple notion as "subject" quickly reveals itself as anglocentric and breaks down when we look at the variation present in the world's more than 6000 languages. Traditional subject markers — nominative case, agreement with the verb, even canonical word order — seem to be the artifact of English informed by classical languages. And if we could resolve such issues for languages with nominative-accusative alignment, a quick look at ergative languages[3] will reveal how complex the situation is from a cross-linguistic perspective.

Moreover, when we consider that the supposed goal of sentence processing is comprehension, i.e. the extraction of meaning, then the classical subject is rather unhelpful, with trivial examples like the passive voice and expletive subjects[4] demonstrating that subjecthood is at best a single cue for the semantic relations actually conveyed by a given sentence. In a series of studies in the 1980s demonstrating that syntax was not sufficient for predicting sentence interpretation cross-linguistically, MacWhinney and Bates put forth a model based on competition between processing cues (Bates, McNew, et al. 1982; MacWhinney, Bates, and Kliegl 1984; MacWhinney and Bates 1989; Bates and MacWhinney 1989). This model can be viewed as the predecessor to the eADM, and as such, warrants a closer look.

1.1.1. Cue-based Processing

Cues in the Competition Model include traditional morphosyntactic features such as case, agreement and word order as well as more semantic and even pragmatic-phonological features such as animacy and stress. The interaction of these cues varies across languages and the language-specific weighting (*cue strength*) depends on *cue validity* (how helpful the cue actually is in determining an interpretation) (MacWhinney, Bates, and Kliegl 1984). Cue validity in turn depends on the *cue availability* (called "cue applicability" in MacWhinney, Bates, and Kliegl 1984) and *cue reliability* (how informative a cue is). For example, morphological case in German is often ambiguous and thus not always available; however, unambiguous case marking provides, when present, for a single possible interpretation. Case marking thus has strong cue reliability yet weaker cue availability, which still yields a large cue validity and thus cue strength, seen in the nigh deterministic nature of unambiguous case marking in German.

[2] Pāṇini, often cited as the first grammarian, wrote his grammar of Sanskrit in the 4th century BC (Meyer et al. 1909).

[3] Ergative languages encode the sole argument of intransitive verbs morphosyntactically the same as the patient-like argument of transitive verbs. If English were ergative, than rather than *She killed him — he died*, we would have *She killed him — him died*.

[4] e.g. semantically void *it* in constructions like *It's raining*.

The information from different cues is combined according to their respective weights to yield a probabilistic interpretation. In German, for example, word order generally drives interpretation in the presence of ambiguous case marking with a preference for subject-initial word orders. However, animacy can change this interpretation such that sentences like *Die Gabel leckte die Kuh*, (lit. 'the fork licked the cow' but morphosyntactically ambiguous) are instead interpreted correctly without requiring forks to have tongues. Using such sentences, MacWhinney, Bates, and Kliegl (1984) and Kempe and MacWhinney (1999) were able to suggest a partial ordering for the core cues animacy, morphology, and word order in several languages.

By separating itself from strict notions of "syntax", the Competition Model presents a parsimonious account of language processing and cross-linguistic variation. Crucially, the Competition Model is a purely psycholinguistic one based on behavioral measurements involving explicit offline judgements. As such, it is susceptible to bias from metalinguistic knowledge and cannot inform us about the precise time course nor the underlying neural architecture of sentence processing.

1.1.2. Beyond Behavior: Neurocognition

The extended Argument Dependency Model (eADM) is a neurocognitive, and more recently, neurobiologically grounded model of cross-linguistic language comprehension, which, like the Competition model, eschews traditional notions of syntax and linguistic domains and uses holistic, cue-based processing (Bornkessel and Schlesewsky 2006; Bornkessel-Schlesewsky and Schlesewsky 2009, 2013, 2014). In particular, the cues are used to drive an interpretation based on the "actor" participant, i.e. the participant primarily responsible for the state of affairs being described (cf. Van Valin 2005; and "Proto-Agent", Dowty 1991; Primus 1999). Similar to the cues of the Competition Model, individual actor-related features will be more important for actor identification in certain languages as opposed to others (e.g. case marking in German, Japanese or Hindi versus English) and, within a particular language, some actor-related features will be weighted more strongly than others (Bates, McNew, et al. 1982; MacWhinney, Bates, and Kliegl 1984; Kempe and MacWhinney 1999; Bornkessel-Schlesewsky and Schlesewsky 2009).

Based on the results of electrophysiological studies across a range of typologically diverse languages, the eADM posits that the human language comprehension system endeavors to identify the actor participant as quickly and unambiguously as possible while comprehending a sentence. Accordingly, if several candidates are available, they compete for the actor role with measurable neurophysiological repercussions (Bornkessel-Schlesewsky and Schlesewsky 2009; Alday, Schlesewsky, and Bornkessel-Schlesewsky 2014). Moreover, non-prototypical actors — even in intransitive constructions — are more difficult to identify, which is reflected accordingly in electrophysiology (Bornkessel-Schlesewsky and Schlesewsky 2009).

Actor identification is based to some extent on the self as the prototype of the ideal actor (Bornkessel-Schlesewsky and Schlesewsky 2009). Cues are thus *features* of this prototype and

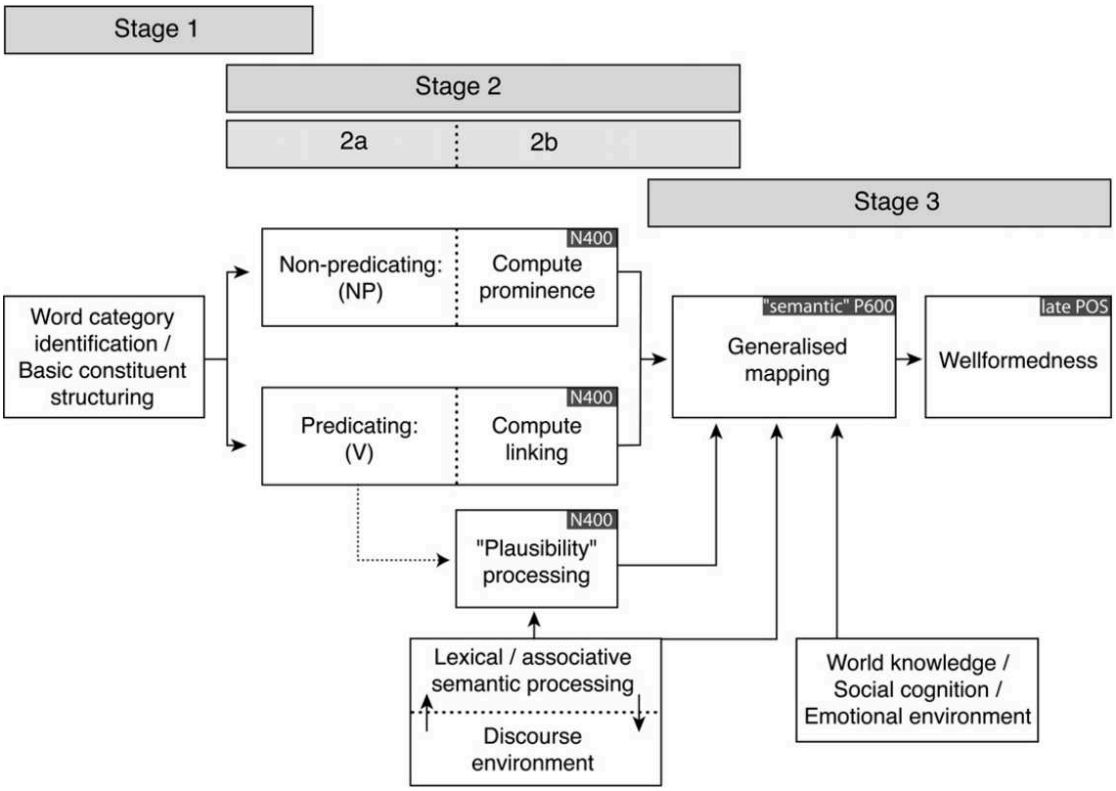


Figure 1.1.: The eADM before its neuroanatomical formulation. Reproduced with permission from Bornkessel-Schlesewsky and Schlesewsky (2008)

encompass both the language-specific cues of the Competition Model (morphology, word order) as well as domain-general features (e.g. animacy, certain movement parameters related to biological motion, first-person and singular) (Bornkessel-Schlesewsky and Schlesewsky 2009, 2013). We call these cues *prominence features* both because they render a referent more salient in a linguistic sense but also because they are more distinguishable from the background in a perceptual sense (e.g. a lone, animate entity like a wolf stands out from a naturalistic scene like a field). We distinguish between *actor prototypicality*, which is in some sense an innate property of the referent, and *prominence* which is a property of an argument's realization in a given context (cf. Frenzel, Schlesewsky, and Bornkessel-Schlesewsky 2015).

Earlier versions of the eADM proposed a three-stage cascading architecture, broadly breaking processing up into (1) initial chunking and morphological analysis, (2) computation of prominence on roughly the lexical level (but including dependency computations between arguments), and (3) sentence scale processing and evaluation, including wellformedness, pragmatic and world knowledge (Figure 1.1, cf. Bornkessel and Schlesewsky 2006; Bornkessel-Schlesewsky and Schlesewsky 2008; Bornkessel-Schlesewsky and Schle-

sewsky 2009). Information flows in a cascade from one stage to the next as soon as it becomes available. Thus even at the start of an initially morphologically ambiguous phrase, the ambiguous morphological information is passed along the processing pipeline. In other words, processing is incremental at every step, with incrementally updated results from the previous processing stage being passed along to the next stage. Information is thus propagated in a continuous manner along sequential yet overlapping processes, which reflects the continuity of processing in the brain. The cornerstone of the actor strategy was found in Stage 2 and named *Compute Prominence*, where the core arguments competed for the actor role.

Later revisions of the eADM have focused on neurobiology and moved away somewhat from the stage-based architecture of the earlier forms (Figure 1.2, cf. Bornkessel-Schlesewsky and Schlesewsky 2013). Nonetheless, the actor strategy remains central to the model. Likewise, previous computational assumptions are now grounded in neuroanatomy, such as hierarchical organization, which leads to a time-space correspondence (i.e. the later processing steps are performed further away from the primary perceptual regions). New to the model are the neuroanatomical streams (based on major fiber-tracts in the brain) for processing flow and information propagation. Expanding upon work in non-human primates that found a division of labor between the dorsal and ventral streams into roughly "what" and "where" (or "how"), the dorsal and ventral streams are proposed to perform different types of manipulations (DeWitt and Rauschecker 2012; Rauschecker and Scott 2009; Bornkessel-Schlesewsky and Schlesewsky 2013; Bornkessel-Schlesewsky, Schlesewsky, et al. in press). The dorsal stream is order-sensitive and thus performs non-commutative operations, such as those based on word oder, while the ventral stream is order-insensitive and thus performs commutative operations such as relative animacy rankings between arguments or the formation of dependencies, e.g. associating modifiers with the modified (Bornkessel-Schlesewsky and Schlesewsky 2013; Bornkessel-Schlesewsky, Schlesewsky, et al. in press; Bornkessel-Schlesewsky and Schlesewsky in press). Both streams are assumed to be bidirectional (i.e. both bottom-up and top-down influences are possible) and some cross-talk is possible, but the data necessary for a fully developed theory on these latter points are not yet available. The streams converge in the frontal cortex, where a final integrative/evaluative/control step is performed in and near the area traditionally known as Broca's Area.

In all of its iterations, the eADM depends heavily on the actor heuristic, and this is the part of the model where the theory is most developed. Indeed, early formulations were perhaps accurately described by Brouwer, Fitz, and Hoeks (2012) as "a model of core argument interpretation (rather than a fully fledged model of sentence comprehension)", although newer revisions provide a parsimonious if somewhat underspecified model of language comprehension. A precise, quantitative formulation of the actor strategy and prominence is thus imperative, and that is the central task of the modeling work presented in this dissertation.

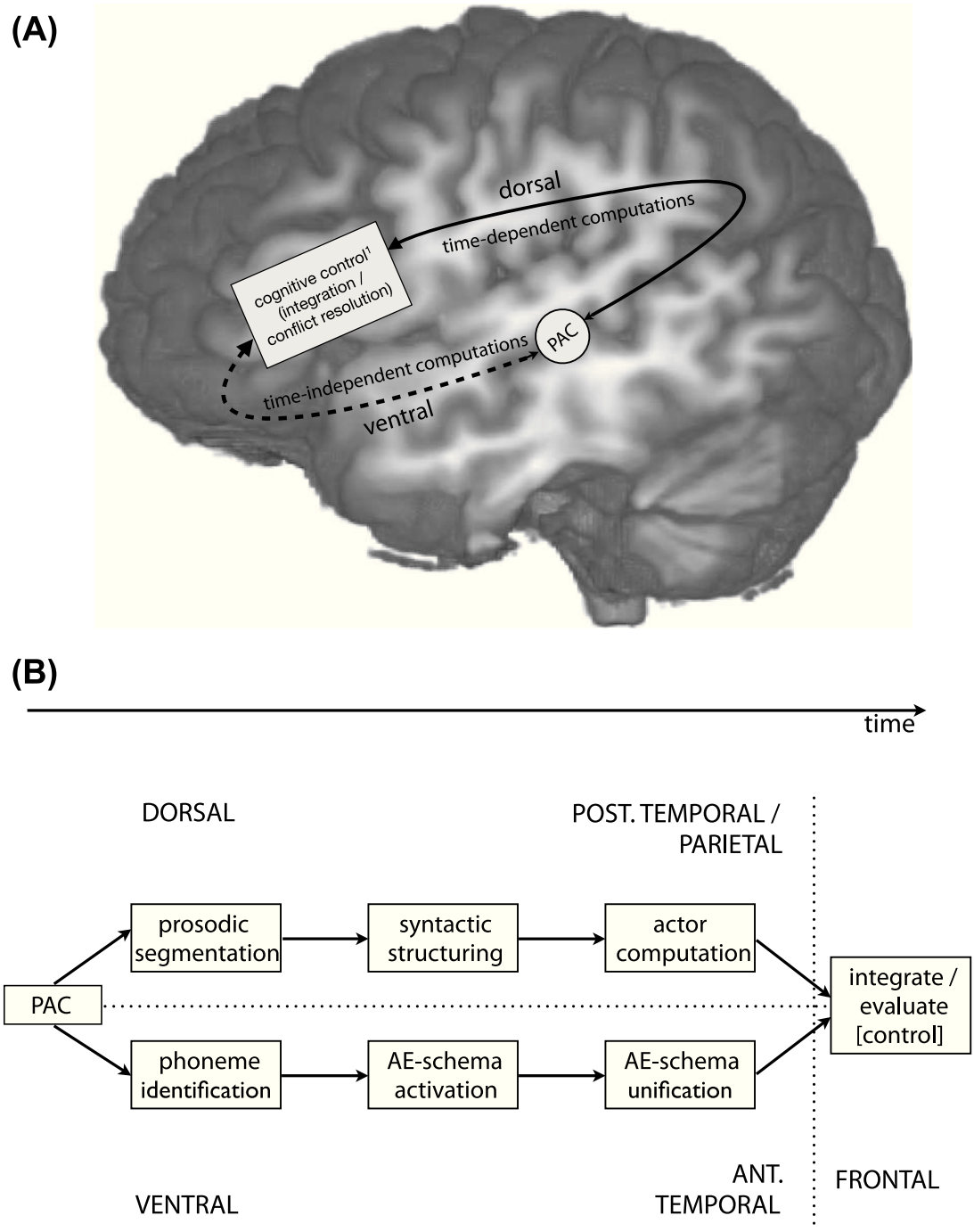


Figure 1.2.: The newer eADM with neuroanatomical streams. Reproduced with permission from Bornkessel-Schlesewsky and Schlesewsky (2013)

1.2. The Goal of Modeling

The evolution of the Competition Model and the eADM provide an opportunity to consider the role of theory and modeling — they both feature a similar core mechanism (cue-based processing, competition), yet differ in significant ways, both in the details of that mechanism and in the level of description sought. Popper (1934) suggested that scientific methodology should be based on falsifiability, as it is in general impossible to verify a theory because it is impossible to test every single possible datum. (This problem had been the bane of the Vienna Circle a few years before.) Lakatos (Lakatos and Musgrave 1970; Lakatos 1978) expanded this notion into research programmes[5] consisting of an immutable "hard core" of non-revisable beliefs, whose falsification renders the entire theory falsified, and a "protective belt" of revisable beliefs, or mutable details. (Modifying the hard core leads to a new research programme.) While the protective belt should be modified to accommodate new data, it should nonetheless be well defined and as specific as possible. Just-so stories and other post-hoc flexibility are often symptomatic of a lack of specificity (cf. Simmons, Nelson, and Simonsohn 2011; Wagenmakers et al. 2012; Gelman and Loken 2013).

Similarly, implementation details for computationally modeled theories should belong to the protective belt and not the hard core. For example, a major problem in the so-called "past-tense debate" of 1990s psycholinguistics was that the debate often focused on implementation details and other parts of the protective belt without testing the claims of the hard core (cf. Alday 2010). A computational implementation is not a theory; however, a theory is often underspecified without a computational implementation due to free parameters (cf. Seidenberg and Joanisse 2003). And even computationally implemented models often show surprising flexibility via free parameters (cf. Howes, Lewis, and Vera 2009).

From this perspective, the Competition Model and the eADM have parts of their respective hard cores in common and thus could potentially belong to the same broader research programme, yet they belong to distinct sub research-programmes on account of their different explanatory goals. The eADM strives to put forth a parsimonious model of the neurobiology of language (albeit currently restricted to language comprehension), while the Competition Model restricts itself to a psychobehavioral model of language comprehension. Nonetheless, the overlap allows for many results to be transferred from one sub research-programme to the other (cf. Marr and Poggio 1976).

1.2.1. Mechanisms and Properties

In this vein, it makes sense to distinguish between mechanisms and properties. A *mechanism* is a process or operating principle which defines the "how" of a model, while a *property* is more (epi)phenomenological. Both mechanisms and properties may be central to the research programme, belonging to its hard core, or auxiliary hypotheses, belonging to its

[5] The British spelling *research programme* is traditionally used even in texts with American orthography, when used in the narrow sense meant by Lakatos.

flexible belt. A well-defined mechanism is sufficient for many properties and can thus be falsified by falsifying those properties. However, a theory may begin by postulating a set of properties and attempting to derive a mechanism for explaining those properties. The Competition Model and the eADM thus have many similar properties but differ in the specification of their (admittedly, related) mechanisms.

A model that differs in its mechanism from the proposed theory can still be useful, if it can tell us something about the properties proposed by the theory.[6] In this sense, it is reasonable to use models that are clearly not "true"[7] because they can still be useful.[8] The use of grammatical relations like "subject" and "object" to describe language comprehension are thus models which have several useful properties (many of which make them sufficient for use in language education), but whose usefulness is much more limited from the perspective of a single cross-linguistically valid account of human language processing.

1.2.2. Levels of Abstraction

In discussing which models are useful for exploring the properties proposed by a research programme, it often makes sense to distinguish between levels of description. Many times the difference in fundamental mechanism offered by a model arises from a different level of description. In neuroscience, Marr and Poggio (1976) proposed four levels of description for an information-processing system, presented here from lowest to highest:[9]

1. **Hardware:** The concrete, physical realization and implementation of the system and the functioning of its atomic components (e.g. neurons, synapses in neurobiology).
2. **Mechanisms:** The low-level mechanisms and fundamental operations of information processing (e.g. memory, executive function, etc.), which map onto complex combinations of the physical primitives (e.g. brain networks, neuroanatomy).
3. **Algorithms:** The algorithms used to solve the information processing problem, which are carried out with the low-level mechanisms.
4. **Computation:** The nature of the information processing problem to be solved ("what" and "why"), somewhat similar to traditional black-box and behavioral approaches.

Previous versions of the eADM primarily addressed the upper (latter) two levels, but with each iteration, the eADM has attempted to address additional issues of the mechanistic level. (Cognitive neuroscience as a whole is still a long way from solving the mapping between the mechanistic and physical levels, in part due to the complexity of the systems involved and in part due to ethical constraints.) Despite their surface similarities, the eADM and the

[6] "[A] theory can be valuable even if it doesn't correspond to the real world because of what it can teach us about theories that do correspond to the real world." Source: http://4gravitons.wordpress.com/2013/03/29/in-defense-of-pure-theory/.

[7] http://4gravitons.wordpress.com/2012/11/26/why-i-study-a-theory-that-isnt-true/

[8] Or, as the saying attributed to George Box goes, "All models are wrong, but some are useful."

[9] It is interesting to note that the original text explicitly defines four levels, although it is often stated as Marr's Tri-Level Hypothesis, with the middle two-levels collapsed (cf. Dawson 1998). This is perhaps due to the somewhat confusing example involving three levels of processing for vision.

Competition Model begin to diverge at the computational level — the eADM frames language comprehension as deeply intertwined with another domain-general computational problem, namely identifying actors, while the Competition Model focuses on the single computational problem of language comprehension.

In this sense, the actor heuristic is a description at a computational and algorithmic level, yet one that seems to map well onto lower levels. While it is certainly important that the proposed algorithms are actually implementable on the mechanisms and hardware available (i.e. that a neurolinguistic theory of language comprehension actually be neurobiologically possible), Marr and Poggio (1976) emphasize the importance of a precise characterization at both the computational and algorithmic levels. In the work presented here, we focused on computational and algorithmic aspects, which are best revealed by the high temporal precision of EEG. As such, EEG, and in particular ERP, serves as the primary online measure, while offline measures such as judgements and reaction times provide a complementary black-box description of the computational problem.

1.3. The New Statistics, Parameter Estimation and Mixed-Effects Models

Precise characterizations have traditionally been a problem in the psychological and language sciences. Human behavior is notoriously complex and language especially so. Moreover, traditional statistical methods based on null-hypothesis significance testing (NHST), e.g. ANOVA, deliver a binary — and hence qualitative — decision. To counter the limits of this approach (which deserve — and have received — books of their own), an approach based on parameter estimation has been proposed under the banner "The New Statistics" (Fidler et al. 2004; Cumming 2013, 2014). Oddly enough, parameter estimation was a necessary part of some of the oldest statistical techniques, e.g. regression; however, many of these were computationally intractable with the numerical techniques and computers of the middle twentieth century, especially when adapted for large datasets or repeated-measure designs. Thus ANOVA, a computationally easy special case of regression yielding only F-tests, came to be preferred to explicit regression yielding both F-tests and parameter estimates.

Recently, linear mixed-effects models have become popular in the psychological and linguistic sciences, due in part to new, faster and easier-to-use implementations as well as increased computing power (Baayen, Davidson, and Bates 2008; Jaeger 2008; Barr 2008; Kliegl, Wei, et al. 2010; Kliegl, Masson, and Richter 2010; Barr et al. 2013; Barr 2013). Mixed effect models allow hierarchical modeling of subject and item effects in addition to the experimental manipulation, or, equivalently, extend linear regression to accomodate for sampling variation along multiple axes simultaneously (subjects, items, etc. in addition to residual error). The work presented here is based extensively on the application of regression techniques — both traditional (generalized) linear models and their mixed-effects extension. In addition to supporting parameter estimation, regression models allow for the inclusion of additional

covariates, which in turn enables modeling more of the variation present in everyday language use.

1.4. Towards a More Precise Formulation of the Actor Heuristic

Parameter estimation thus gives us a method of hardening the core of a research programme as well as firming up its protective belt so that its parts may one day be moved into the core. Exploration of the parameter and property space requires specification through the development of models, yet many models are woefully underspecified. In the following chapters, a number of quantitative modeling approaches are explored, all with the goal of improving the quality of the theory. In Chapter 2, we examine the statistical learnability and global optimality of the actor heuristic compared to the subject heuristic by examining the behavior of a dependency parser, and justify the use of dependency parsers as a model sharing many properties with human language comprehension. Chapter 3 examines an initial quantification of prominence and distinctness with EEG data using weights derived *a priori* from qualitative results in the literature, while Chapter 4 demonstrates the feasability of reducing the free parameters in the eADM by estimating the feature weights at a single subject level. Subsequently, the feasibility of studying the electrophysiology of language and of the actor heuristic in a natural story context is demonstrated in Chapter 5. Chapters 2 and 6 present new, unpublished material, while Chapters 3, 4 and 5 introduce material already published or currently under peer review. This difference is also reflected in the formats of the respective chapters. Review chapters consist of a brief summary of the publication, a description of its relevance as well as a short entry on my contribution to the published work, followed by the manuscript in its entirety. Finally, Chapter 7 reviews the major insights garnered by this comprehensive approach and presents an outlook.

2. Looking for Arguments That Get Stuff Done: Greediness and the Actor Strategy

> All models are wrong, but some are useful.
>
> George Box

As mentioned in the Introduction, the traditional grammar-school notion of "subject" is cross-linguistically a tenuous concept at best. Nonetheless, it has achieved some manner of success as a model, both in terms of linguistic description and as a driving principle of parsing and comprehension strategies. Clearly, the subject model has some useful properties that capture certain aspects of linguistic reality — many languages have something like subject-verb agreement and there seems to be even cross-linguistically a preference for subject-initial word order (cf. Comrie 2011).

The preference for subject-initial configurations can be viewed from a computational perspective as a *greedy* algorithm, or an algorithm which makes locally optimal choices at each increment. However, global optimality, i.e. the correct interpretation of a sentence in its entirety, does not necessarily follow from local optimality. Consider for example the partial sentence:

(1) Peter sah ...
Peter.ambiguous saw.1st-3rd ...

At this point, the input consists of a noun with ambiguous case marking and a verb, which agrees in number with the noun. There are thus two possible interpretations: Peter is either the seer (2) or the seen (3).

(2) Peter sah mich.
Peter.ambiguous saw.1st-3rd me.

(3) Peter sah ich.
Peter.ambiguous saw.1st-3rd I.

The input is thus locally underspecified. We can avoid a potentially false interpretation by weakening the requirement for incremental processing, but the evidence does not seem to support this — a wide range of studies report an N400 on the second noun for such ambiguous-verb-nominative configurations (for a particularly recent one, see Alday, Schlesewsky, and Bornkessel-Schlesewsky 2014). Yet object-initial orders seem to be preferred with object-experiencer verbs like *gefallen*:[1]

(4) Peter gefällt die Sängerin.
Peter.ambiguous pleases.3rd the singer.f.NOM.

(5) Peter gefällt der Sängerin.
Peter.ambiguous pleases.3rd the singer.f.DAT.

This is still broadly compatible with a subject-based heuristic under a slightly weakened incrementality requirement: if the initial argument is followed by a "normal" verb, then interpret it as a subject, otherwise if the initial argument is followed by an object-experiencer verb, interpret it as an object. Although more weakly incremental than a word-for-word account, it is nonetheless reasonable in its assumptions, especially under traditional (generative) syntactic theory, where the subject concept only makes sense in a structural relation to the verb. However, when we consider that verb-final constructions are also possible, this proposal of "wait for the verb" quickly becomes untenable. Moreover, electrophysiological evidence also indicates that even unambiguous object-first constructions are at least locally dispreferred (e.g. increased N400 amplitude on the second noun in verb-final constructions Frisch and Schlesewsky 2001; cf. Bornkessel 2002). A greedy, subject-based heuristic can thus fail to be globally optimal in a number of situations. As incrementality — greediness — seems to be a fundamental property of the human language system, we are forced to reconsider the assumption of subject-centricity, despite its usefulness as a model in other contexts.

Within the framework of the eADM, the actor-heuristic, or more generally, prominence-driven processing, was introduced in Chapter 1 as an alternative to subject-centric processing. If we consider the previous examples, then we see a more consistent pattern, namely a decreasing prominence ranking across the sentence, the quantification of which is the subject of Chapter 3.

For the object-experiencer verbs,[2] it is difficult to find a clear "actor" in a strict sense of agency — arguably the stimulus has causal properties, but does not actually *do* anything, while the experiencer suffers or endures an emotional state. The properties of the actor are modeled on the prototype of the self and its role in the world (i.e. as part of the perception-action-loop, cf. Bornkessel-Schlesewsky, Schlesewsky, et al. in press; Bornkessel-Schlesewsky and Schlesewsky 2013, in press; see also Bornkessel-Schlesewsky

[1] lit. 'to please', but used in the sense of 'to like', only with the liker as the object and the likee as the subject.
[2] i.e. verbs, whose objects are subject to a mental state. Examples in English include *excite*, *please*, and *amuse*.

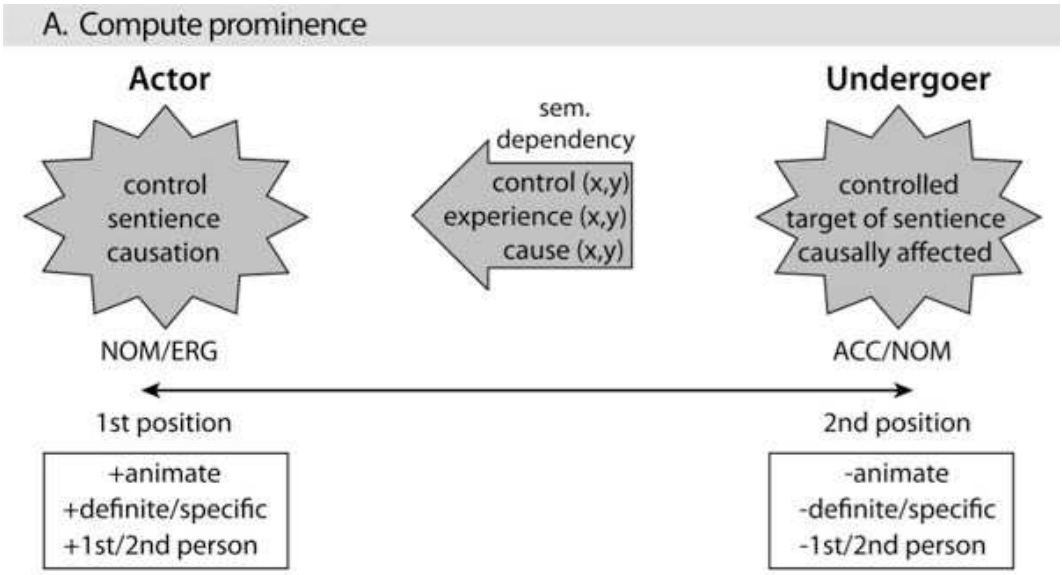


Figure 2.1.: Reproduced with permission from Bornkessel-Schlesewsky and Schlesewsky (2009)

and Schlesewsky 2009; and Primus 1999). In this sense, emotional states, a mark of sentience, are more prominent than the stimulus (cf. Figure 2.1, Primus 1999).

The label "actor strategy" is thus a convenient shorthand for prominence processing, based on its most common variant (and potentially its evolutionary roots as part of the perception-action-loop, cf. Bornkessel-Schlesewsky, Schlesewsky, et al. in press; Wolpert 1997). Earlier formulations proposed a complementary role for the actor, which they termed "undergoer", but ontologically, this role is completely dependent on and largely defined by the contrast to the actor (cf. Bornkessel-Schlesewsky and Schlesewsky 2009; Primus 1999). Later formulations of the eADM removed the explicit undergoer category (Bornkessel-Schlesewsky and Schlesewsky 2013), but we can still use the term as a convenience. We can thus view the actor strategy as trying to establish the dependency relationship between arguments (see Figure 2.1). This predicts that this dependency relationship may potentially be more reliably determined in a greedy fashion in verb-final constructions, even with ambiguous case morphology, than the subject-centric constructions.

The central goal of this dissertation is specification through modeling, especially quantitative modeling. The claimed global optimality of the actor strategy compared to the subject strategy in a greedy system must then be demonstrated. Dependency parsing, although differing greatly in its mechanism from human language comprehension, nonetheless provides a useful model for exploring the properties of the proposed processing strategies. In the words of Seidenberg and Joanisse (2003), it's time to show a(n implemented) model. For that, we turn to dependency parsing.

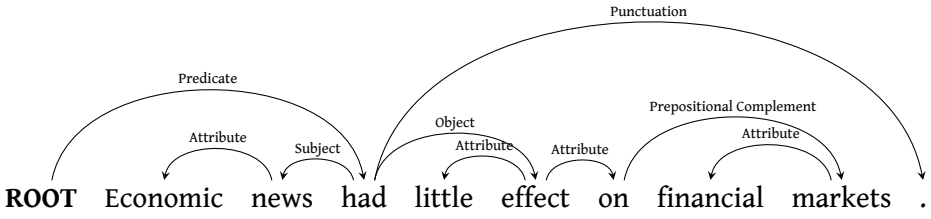


Figure 2.2.: Dependency structure for an English sentence (adapted from Nivre 2010)

2.1. Dependency Parsing

In the computational linguistics literature, "dependency parsing" refers to systems and techniques for mapping a sentence to a hierarchical structure based on the principles of dependency grammar, a syntactic tradition going back centuries, which has primarily found application in descriptive linguistics (Nivre 2010). Much like generative grammar, there are several different theoretical frameworks with different assumptions about the finer details, but all share a core assumption: syntactic structure consists of *words* and directed, assymetrical relationships between them called *dependencies* (Nivre 2010). As is usual in modern linguistics, this structure can be represented as a rooted[3] tree (see Figure 2.2), with the node closer to the root called the *head* and the node closer to the periphery called the *dependent*. We represent this tree with arrows from each head to its dependents,[4] i.e. flowing from the root towards the leaves. Each arc is given a label indicating the dependency type (see Figure 2.2), which generally describes functional relationships between a head and its dependents, e.g. "subject", "object" or "attribute". This differs from the more common phrase structure representation, which uses *phrases* and *structural* categories instead of *words* and *functional* categories (Nivre 2010, cf. Figure 2.3). Although the type of information represented directly differs, it is in general possible to convert between the two types of representations (Nivre 2010).

For our purposes, the representation offered by dependency parsing offers several advantages. Functional relationships, such as subject and actor (see below) more reliably align with a form-to-meaning mapping than do structural ones. As determining this mapping is the goal of language comprehension, an explicit encoding of functional relationships is advantageous for a model of human language processing. In the following, we present an initial account of dependency parsing as a model of language comprehension.

[3] For computational reasons, we introduce an artifical root, which prevents elements from lacking a syntactic head.

[4] N.B. There is a rival tradition which draws the arrows in the other direction (Nivre 2010).

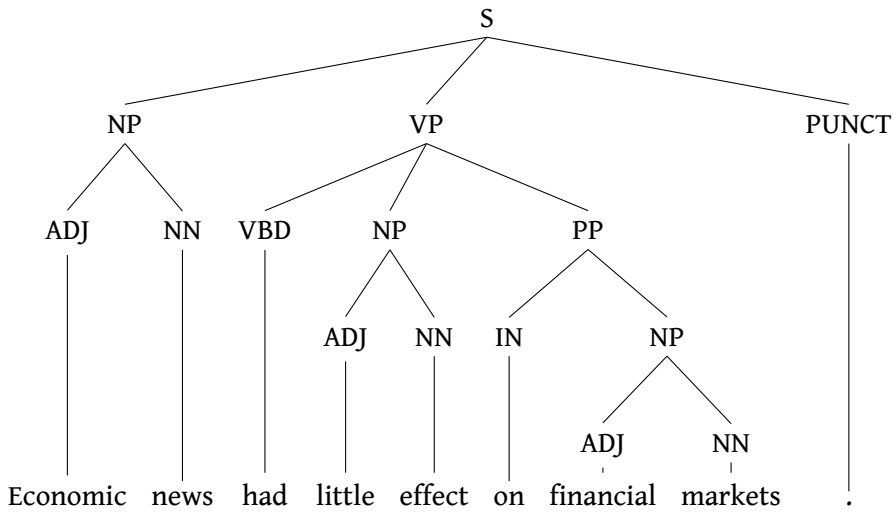


Figure 2.3.: Phrase structure for an English sentence (adapted from Nivre 2010)

2.2. Dependency Parsing as a Cognitive Representation

A further advantage of dependency relationships is that they can be interpreted as encoding parameterized, hierarchical evaluation of language units.[5] For example, the representation in Figure 2.2 directly encodes which words modify which others: modifiers are directly adjoined to the modified, verbal complements (subjects, objects and potentially prepositional complements) are directly adjoined to the verb complex. We can view modifiers as parameters (in a functional-computational sense) to a noun phrase, modifying its computational behavior (meaning), and similarly verbal complements parameterize the verb via its argument structure.

This computational perspective, while simple, has several subtle and desirable implications. First, there is no need for a strong separation of verbs and nouns (or word categories in general): the word category emerges from a given parameterization. A verb-like word with subject and object or actor and undergoer parameters is functionally a verb. A noun-like word being used as a parameter of a functional verb is a noun. This matches well with the assumptions of the eADM regarding the flexibility of word categories (Bornkessel-Schlesewsky and Schlesewsky 2009, in press) as well as neurolinguistic findings indicating that the brain does not necessarily distinguish between word categories (for fMRI evidence, see Tyler et al. 2004; Vigliocco et al. 2011; for EEG, Federmeier et al. 2000; for a review, see Crepaldi et al. 2011; for a dissenting opinion, Shapiro and Caramazza 2003).

[5] It seems a bit redundant to say "[syntactic] dependencies encode computational dependencies", but this is a rather important point.

Moreover, this representation captures the notion of "level of detail": each subsequent embedding is a level of detail that can optionally be removed. (In phrase structure representation, the adjectives and the nouns are at the same level of embedding.) This allows us to focus on core aspects of comprehension with natural language by simply trimming the trees so that deeper embeddings are ignored, and successively reduce the amount of trimming as the theory becomes more developed.

Finally, the combination of parameterization and level of detail yields a model of hierarchical encoding or evaluation[6] with each head representing the evaluation of itself as parameterized by its dependents. In other words, subtrees represent a model of *chunking* in a way that captures the most essential details as being most accessible, i.e. higher up in the tree and thus more recently evaluated. This chunking model is also compatible with models of language processing based on content-addressable memory, such as Lewis, Vasishth, and Van Dyke (2006), with the growing chunks representing accumulating feature bundles and attachment depth related to decay *of individual features*. Anti-locality effects, where increasing distance between a noun and its verb can increase processing speed, then reflect reactivation via re-evaluation during new attachment.[7]

The dependency representation is thus especially well suited to modeling the cognitive representation of language. For a model of language processing, however, we require that the parser also construct its representations in a comparable way, i.e. incrementally, and, ideally that the parser be able to learn and not require the input of an explicit grammar, i.e. be data-driven. In the dependency tradition, transition-based parsing meets both of these requirements.

2.3. Transition-based Parsing

A *transition-based parser* is a data-driven shift-reduce parser. In other words, transition-based parsers take a local perspective to parsing, examining each input and deciding whether to use it immediately, e.g. by combining it with previous input (*reduce*), or to read in additional input and see if the situation improves any (*shift*). In dependency parsing, there are two possible reductions: *left arc* and *right arc* attachment, which correspond respectively to attaching the new input to existing input as either a head or a dependent. These state *transitions* are learned in a supervised way by training the decision mechanism (*oracle*) on annotated data. Although language learning in humans is arguably not explicitly supervised, human learners do receive implicit feedback as to whether or not they arrived at the correct interpretation.

[6] Here and in the following, "evaluate" and its derivatives are meant in a computational sense as in "function evaluation", whereby a computation is carried out.

[7] Note that this does provide for a certain correlation with the notion of "dependency" and its effect on language processing as put forth in the Syntactic Prediction Locality Theory (SPLT) and its successor Dependency Locality Theory (DLT) (Gibson 1998, 2000); however, the dependencies in dependency parsing are inherently different from the combined syntactic-referential dependencies put forth by Gibson, not the least because Gibson defines locality via phrase-structural distance.

Crucially, although there have been some recent efforts to implement back-tracking or other repair strategies (Honnibal, Goldberg, and Johnson 2013), the majority of transition-based parsers are forced to follow through on their decisions. In this way, they can also be tricked into garden-pathing in the same way humans are — a locally optimal decision to reduce turns out to be wrong. Neither transition-based parsers nor the human-processing mechanism can escape when they have gone far enough along the wrong path. With smaller violations, however, the human language processer is capable of recovering, while a transition-based parser cannot escape its previous choices (even if it can make correct choices later on that lead to self contradiction, such as two subjects). Local parsing errors thus remain in the final output as well as potential consequences of initial errors. Taken together, these errors should correlate with the combined local and global optimality of certain comprehension strategies, such as the subject and actor strategies. In particular, we can explore and compare the optimality of the subject and actor strategies by examining the performance of a transition-based dependency parser trained on label sets containing either actors and undergoers or subjects and objects.

2.4. A First Attempt

In the following, we present an initial exploration of the performance characteristics of a dependency parser trained on traditional subject-object labels compared to one trained on prominence-based labels. We used the MaltParser (Nivre et al. 2007), which is a completely data-driven transition-based parser with flexible specification of *feature models*, the set of input features that the oracle is trained and operates on.

For the training and test sets, we used stimuli from the experiments presented in Chapter 4 of Bornkessel (2002), which followed broadly speaking a 2x2x2 design for WORD-ORDER x AMBIGUITY x VERBTYPE. Subject-object and object-subject word orders were presented in a verb-final configuration. All sentences were globally unambiguous with only one argument agreeing with the verb in number; however, the sentences were also manipulated so as to be either locally case ambiguous or locally unambiguous. In other words, in unambiguous sentences, it was possible to deterministically assign grammatical relations based on local input, but in ambiguous sentences this was not possible until the verb. Both active-dative verbs and dative-experiencer verbs were used; thus thematic role assignment was only aligned with syntax in half of the sentences. An example of the four ambiguous conditions is presented in (6).

(6) Gestern wurde erzählt, dass
Yesterday was told that

a. Maria Sängerinnen folgt.
Maria singers follow.SG
'Maria follows singers.'

b. Maria Sängerinnen folgen.
Maria singers follow.PL
'Singers follow Maria.'

c. Maria Sängerinnen gefällt.
Maria singers please.SG
'Maria pleases singers.', i.e. 'Singers like Maria.'

d. Maria Sängerinnen gefallen.
Maria singers please.PL
'Singers please Maria.', i.e. 'Maria likes singers.'

For this first attempt, the parser was trained and tested using a single experimental item (particular lexical realization of the experimental conditions) with a delexicalized feature model (see below). Experimental stimuli were used because no appropriate treebank was available, and a single item sufficed because psycholinguistic stimuli are strictly controlled to be feature matched and thus each additional item introduces no variation to the training set. A larger training set can be emulated in a restricted sense by increasing the number of training iterations.

2.4.1. Properties of Appropriate Feature Models

Features in MaltParser are specified in terms of the input buffer (FIFO[8]) and a stack of partially processed tokens (LIFO[9]), which corresponds to memory in a cognitive architecture with most recent elements most easily accessible. Items in both the stack and input buffer may additionally be addressed by index or by their dependency relationship (i.e. by both linear and graph position), but arcs may only be made between the top/front most element of the stack and input buffer. In terms of cognitive architecture, this is equivalent to only being able to integrate the new input with the most recent representation, which can be either the directly preceding element or a more complex hierarchical structure computed from recent input.

Although MaltParser supports look-ahead, look-ahead *must not* play a role in a model of natural language processing because language is inherently sequential.[10] Look-back corresponds to memory and is currently an underspecified portion of the model. A reasonable assumption seems to be the availability of all features for the top of the stack and some features

[8] first in, first out
[9] last in, first out
[10] Parafoveal preview in reading somewhat violates this principle by providing limited access to upcoming input, and this is reflected in different electrophysiological effects in reading (cf. Kretzschmar, Bornkessel-Schlesewsky, and Schlesewsky 2009; Kretzschmar 2010). For a model of comprehension during reading, it may therefore make sense to allow a limited look-ahead with reduction in the number of features available.

for both the next element in the stack and the dependencies of the topmost element. This corresponds to full availability for the most recently processed element as well as some availability for recently processed elements and co-activated elements as in the Lewis, Vasishth, and Van Dyke (2006) model.

For the current experiment, no features were specified beyond those of the topmost element as well as the current input. While this arguably crippled the parser, it prevented a more serious problem for such short, simple sentences, namely that the parser shifted until the verb and then performed globally informed processing.

2.4.2. Non-Traditional Dependency Structures

In addition to the restrictions to the feature model, incrementality in human language processing also presents another set of constraints for the graph representation of the core grammatical or prominence relations, i.e the exact dependency relationship between the verb and nominal arguments. In particular, it is unlikely that the human language system waits for the verb to process the nouns, which implies that the language system attempts to categorize the nominal arguments immediately. This can be inferred by either a further nominal argument or the verb in verb-medial constructions. As such, a dependency structure that reflects this and allows for immediate, incremental arc attachment should be preferred.

The solution used here, which we call *Argument-Verb Chain*, attaches the subject or more prominent argument to the verb and the object or less prominent argument to the other argument. Thus, when a verb is encountered, a subject or more prominent argument can be immediately integrated, which roughly corresponds to Compute Linking in earlier versions of the eADM (Bornkessel-Schlesewsky and Schlesewsky 2009). Alternatively, when an additional noun is encountered, the relationship between the two can be immediately established, which roughly corresponds to Compute Prominence in earlier versions of the eADM (Bornkessel-Schlesewsky and Schlesewsky 2009). The correct parse under this scheme for the sentences in 6 is presented in Figure 2.4.

2.4.3. Preliminary Results

Testing was conducted using a mixture of the leave-one-out cross-validation procedure and self-verification. In both cross- and self-validation, the parser failed to achieve accuracy above 85% on either label set, despite highly constrained input. Initial results showed better performance for the subject label with an extreme sensitivity to the exact feature specification for both relative and absolute performance. For example, encoding case ambiguity as having no case marking increased accuracy compared to encoding case ambiguity as being compatible with both nominative and dative cases (depending on the exact feature model, in excess of 50 percentage point difference).

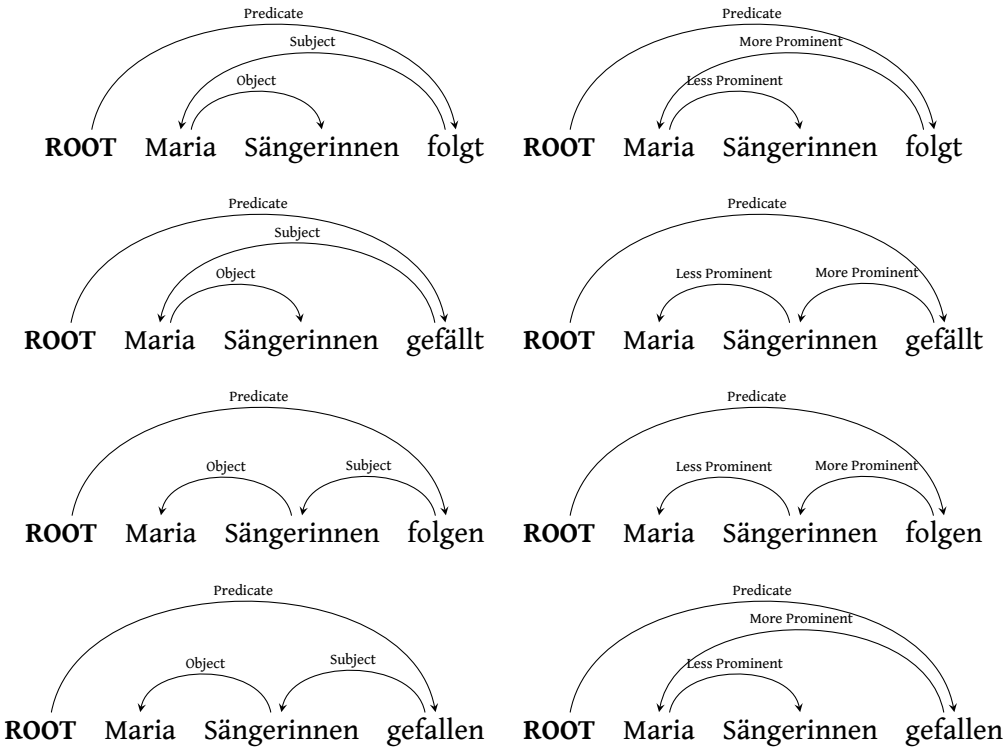


Figure 2.4.: Argument-verb-chain dependency structure for the sentences in (6). The left column uses the traditional grammatical-relations label set, while the right column uses the prominence label set.

This sensitivity to feature specification also revealed a more problematic aspect of the experimental design. The parser was trained and tested on the critical conditions of a psycholinguistic experiment, *without any distractor stimuli.* Much as human test subjects do, the parser developed an optimal strategy for the experiment instead of for general use. Sentence interpretation for the subject-label was possible by only examining number agreement, while sentence interpretation for the actor-label additionally required examination of the verb type. The disadvantage of this additional complexity is also reflected in improved performance for additional training iterations with the actor strategy, but not for the subject strategy.

This suggests that the advantage of the actor strategy, at least for traditional nominative-accusative languages, arises in part from distributional statistics of constructions. When all combinations are equally likely, then there is no purely linguistic advantage to the actor strategy and there may even be a slight disadvantage. From a psychological perspective, the actor strategy does have a few advantages: (1) it is a domain-general mechanism and (2) it has cross-linguistic applicability in a way that "subject" does not. Initial work on applying dependency parsing to languages such as Hindi, which are split-ergative, show

high unlabeled accuracy, but a somewhat disappointing label accuracy (Nivre 2009).[11] While a data-driven parser may be able to learn ergative alignment with as much success as nominative-accusative alignment, learning both systems in a single language seems difficult. Prominence-based labels (even if applied to a more traditional graph structure than used here) may offer a way to resolve this problem.

Finally, one training-testing run based on a feature set with access to the second element in the stack (i.e. with additional memory) yielded more promising results for the actor strategy, with the actor strategy achieving an aggregate accuracy of 67%, while the subject strategy scored only 45.8%. More interestingly, the parser tended to make mistakes and generalizations as humans do; the results from leave-one-out testing for object-initial word orders are in Appendix A. Using subject-labels, the parser does indeed develop a subject preference, while using actor-labels, it does develop a preference for more prominence. Crucially, the number of mistakes for a parse with actor-labels follows more closely the behavioral preferences of human test subjects than does the number for a parser with subject-labels.

Unfortunately, these interesting results must be regarded for the time being with caution. Using a snapshot of the testing environment (including intermediate build and training products) taken after these tests consistently yields the same results; however, running the testing again using a clean checkout from version control yields different results. Thus far, we have not been able to find the source of the discrepancy, but we have reason to believe that such results may again be possible: namely the mistakes in the test sets seem to more closely match the biases found in the leave-one-out training sets than the results from the replication attempts do. For example, in the non-replicable environment, training sets missing an object-initial construction resulted in a preference for subject-initial constructions in testing, but this type of expected bias is not always found in the replication attempts. The replication attempts also demonstrated a suspiciously low accuracy for the actor-label (lower than chance), which was insensitive to removing number and verb type from the feature model, suggesting a problem with the replication attempt.

2.5. Review and Outlook

In this chapter, we examined the use of transition-based dependency parsers as a computational model of human language processing. Following a discussion of the useful properties of these parsers, we used them to test the learnability and optimality of greedy strategies based on traditional grammatical relations (subject and object) and prominence (actor and undergoer) with mixed results. The performance of each strategy depended on the exact feature model specification the oracle was trained on, and much like human test subjects in the absence of distractors, the parser developed an experiment-specific strategy on several occasions. While this challenges the supremacy of the traditional subject heuristic, it does

[11] i.e. the correct dependency relationships (arcs) were established, but often were the wrong type of relationship (labels).

not establish the optimality of the actor heuristic either. Finally, prominence-based parsing may provide a solution to issues related to parsing split-ergative languages.

Further work must focus on developing an adequate training treebank so that the statistics of natural language can be taken into account, as a "fully balanced" training set actually introduces bias for rare constructions and thus reduces their difficulty in the testing set. Additionally, while some necessary properties of the oracle's feature model were established, there is still a great deal left underspecified. Exploring the behavior of different feature models on larger training corpora will provide insight into how different cues are integrated to optimize decision making.

3. Distinctness as a Numerical Quantity

> Anyone who cannot cope with mathematics is not fully human. At best he is a tolerable sub-human who has learned to wear shoes, bathe, and not make messes in the house.
>
> Robert Heinlein, *Time Enough For Love*

A key element of the proposed actor heuristic is competition between the candidates for the actor role. In metaphor and in practice, competition is much more difficult when the competitors are closely matched, which leads to the proposition that the actor heuristic is most efficient when the competitors are maximally *distinct* (Bornkessel-Schlesewsky and Schlesewsky 2009, 2013, 2014). Quantifying distinctness is thus a necessary step in quantifying the actor heuristic.

In Alday, Schlesewsky, and Bornkessel-Schlesewsky (2014), we presented an initial exploration of different possibilities for quantifying prominence and distinctness based on two basic concepts from linear algebra: *metrics* and *dot products*. In this chapter, we introduce the formalisms used in that publication, summarize the results and present some considerations for future research.

3.1. Brief Summary of Methods and Results

3.1.1. Mathematical Formalisms

In developing a computational model for prominence, we follow the usual practice and encode the set of prominence features for a given participant as a *feature vector*, i.e. an ordered list of numerical encodings for prominence features. In particular, we view each feature as a dimension in Euclidean n-space, where n is the number of features we are trying to model. Distinctness can then be conceived of as "distance" in this space, which in general corresponds to the idea of a *metric*[1] in mathematics.

[1] For a vector space V defined over the real numbers $\mathbb{R}$, and vectors $u, v, w \in V$, a metric is a function $d : V \times V \rightarrow \mathbb{R}$ satisfying the following properties:

The Manhattan Metric as Interference

In particular, the Manhattan metric, $d(u, v) = \sum_i |u_i - v_i|$, seems to reflect a natural notion of distance in prominence space, as it provides a measure of feature overlap. Feature overlap correlates inversely with the notion of *interference* found in the literature on working memory (cf. McElree 2006; Jonides et al. 2008; Lewis and Vasishth 2005; Lewis, Vasishth, and Van Dyke 2006), and thus Manhattan distance yields a measure of distinctness comparable (but inverted from) interference.

Dot Product as Prominence

A key feature of the eADM (and the Competition Model) is the language-specific weighting of prominence features. We can conceptualize this as a distortion of prominence space with small differences in heavily-weighted features stretched to become greater in relation to their less strongly weighted features. This distortion is similar in spirit to the distortion of space-time by heavy objects put forth by general relativity, which yields a useful metaphor: the distortion along the feature weighting works as an attractor basin for the actor role, which fits in well with recent suggestions from computational neuroscience that suggest attractor networks provide a neurobiologically plausible means of modeling decision-making processes (Deco, Rolls, Albantakis, et al. 2013; Deco, Rolls, and Romo 2009; Basten et al. 2010; Heekeren et al. 2004).

This notion can be extended to "repulsor (hills)", which describe dispreferred or unstable configurations. An example is the state produced by an initial accusative before the presentation of a subsequent nominative — the actor heuristic is temporally forced to consider an untenable actor assignment. The sign of a feature's weight indicates whether it is an attractor or a repulsor, while its magnitude indicates its strength.

The dot product of the weight vector with the feature vector, i.e. the sum of individually weighted features, is thus a measure of the strength of attraction or repulsion. This corresponds to a measure of prominence or compatibility with the actor role. Moreover, this measure has the interesting property of being equivalent to a weighted, signed Manhattan distance:[2]

1. (Identity) $d(u, v) = 0$ if and only if $u = v$
2. (Symmetry) $d(u, v) = d(v, u)$
3. (Triangle Inequality) $d(u, w) \leq d(u, v) + d(v, w)$

   Together, these imply a fourth property:

4. (Non-negativity) $d(u, v) \geq 0$

[2] As a metric is non-negative by definition, it is no longer accurate to call a signed variant a "metric". For such near misses, we use the everyday term "distance" and reserve the term "metric" for those meeting the formal criteria.

$$\begin{aligned}\mathbf{w} \cdot \mathbf{NP2} - \mathbf{w} \cdot \mathbf{NP1} &= \sum_i w_i \cdot \text{NP2}_i - \sum_i w_i \cdot \text{NP1}_i \\ &= \sum_i w_i \cdot (\text{NP2}_i - \text{NP1}_i) \\ &= \mathbf{w} \cdot (\mathbf{NP2} - \mathbf{NP1})\end{aligned}$$

Adding in appropriate absolute value signs to the last line would yield a weighted Manhattan metric.

Elementwise Difference as Net Gain in Actor Features

Comparing the Manhattan metric, which lacks any notion of "directionality", to a signed difference is somewhat unfair when both measures are supposed to describe a greedy, i.e. order-dependent, process. As such, we use the signed Manhattan distance proposed above without weights as a third distinctness measure. By setting all the weights equal to one, we have a way to establish the effect of weighting. This measure is equal to the summed elementwise difference and thus measures the net gain in actor-compatible features.

3.1.2. Results

With the above mathematical formalisms, we were able to provide a computational implementation of Compute Prominence (see Chapter 1), which we combined with a simple shift-reduce parser to provide measures for the stimuli used in an EEG experiment. The EEG experiment used verb-medial sentences manipulating word order (SO vs. OS), local case ambiguity on the first NP, and the type of both NPs (pronoun vs. full NP). For the first NP, the pronoun was third-person singular, while for the second NP the pronoun was first-person singular. Disambiguation for the ambiguous condition always occurred on the second NP and never on the verb. This experimental manipulation is known to elicit a biphasic N400-late positivity effect, which allows us to test the proposed measures in both time windows.

Using mixed-effects models, we examined the predictive power of the three measures in the N400 and late-positivity time windows. In both time windows, the difference-in-prominence measure (signed difference of dot products) provided the best fitting model, the net actor-feature change (signed elementwise difference) again the second best, and the interference model (Manhattan metric) the worst. Model fits in the N400 time window were generally better than model fits in the late-positivity time window, which fits well with results indicating that the N400 is a more direct measure of actor competition (for a review, see Bornkessel-Schlesewsky and Schlesewsky 2009). Moreover, there is strong evidence that the late-positivity effect is in part task-related (Haupt et al. 2008; Sassenhagen, Schlesewsky, and Bornkessel-Schlesewsky 2014), which was supported by the much larger improvement

in model fit with the inclusion of reaction time for the late-positivity model compared to the N400 model.

3.1.3. Future Directions

Recently, Frenzel, Schlesewsky, and Bornkessel-Schlesewsky (2015) showed that there is an electrophysiological effect for the *actor prototypicality* of a given lexeme or referent, which is distinct from the electrophysiological effect resulting from the *prominence* of its realization in a sentence context. Incorporating such lexical data into the model proposed here should then improve its fit to the EEG data. This could be done either at the population level (incorporating actor prototypicality values derived from another group of test subjects) or at the individual level (allowing each test subject to rate the stimuli at a later date).

Although we are able to ground our choice of metric and signed distance measures based on their abstract properties, it may nonetheless be worthwhile to consider other metrics for future work. In particular, the Mahalanobis distance (Mahalanobis 1936), which can be thought of as the generalization of z-scores to higher dimensions, shows some conceptual promise. The Mahalanobis distance can be conceptualized geometrically as expressing distance from the center of mass of an object, or equivalently, a distribution.[3] In terms of actor space, this could be used to measure the distance from a particular argument to the distribution representing all actor arguments. (Neurocognitively, this would be sum total of an individual's experience with actor arguments. In terms of modeling, this could be calculated from a corpus.) As a standardized measure, the Mahalanobis distance would then provide a measure of actor prototypicality that accounts for the distributional differences of the various prominence scales. As a corollary, the Mahalanobis distance could also be used as a dissimilarity measure between two arguments.

The Mahalanobis distance has two potential problems: lack of directionality and weighting based purely on distributional characteristics instead of informedness (i.e. cue availability instead of cue validity). As the Mahalanobis distance is related to the Euclidean distance (and indeed, standardized Euclidean distance is a special case of Mahalanobis distance), adding directionality is not as trivial as simply removing absolute value signs. Squaring changes not only the sign (preventing issues with non-negativity, especially given the subsequent square root) but also many characteristics of the metric.[4] (The subsequent square root is less of a problem, as we could simply interpret directionality via the imaginary component.) Additionally, many of the desirable properties of the Mahalanobis distance, e.g. its geometric interpretation, arise from a weighting that is incompatible with the principles of the Competition Model and eADM.

[3] We are using the term "distance" here because we are being somewhat loose in our terminology, e.g. measuring the distance between a distribution and a point, but the Mahalanobis distance can be used as a metric in the rigorous sense.

[4] A clear example of this is the difference between the mean and the median in statistics, which can be thought of the center of a distribution based on the Euclidean and Manhattan metrics, respectively. See also http://www.johnmyleswhite.com/notebook/2013/03/22/modes-medians-and-means-an-unifying-perspective/.

3.2. Relevance

In this paper, we presented for the first time a computational implementation of the actor-heuristic. We demonstrated that, as predicted by the eADM, a weighted feature model outperforms unweighted interference models such as those proposed in the working memory literature. On the basis of the mathematical formalism developed here and evidence from computational neuroscience, we developed the notion of attractor basins as an additional useful formalism for the actor heuristic, which has since been integrated into the eADM (cf. Bornkessel-Schlesewsky and Schlesewsky 2014), and suggested that this mechanism may subserve a number of sub-computations necessary for language comprehension. Finally, this work serves as a framework for testing actor prototypicality effects, which complements the recent work from Frenzel, Schlesewsky, and Bornkessel-Schlesewsky (2015).

3.3. Publication

Peer-Reviewed Article P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (2014). "Towards a Computational Model of Actor-based Language Comprehension". In: *Neuroinformatics* 12.1, pp. 143–179. DOI: 10.1007/s12021-013-9198-x

Conference P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (2012). *Towards a Computational Model of Actor-based Language Comprehension*. Poster presented at the Neurobiology of Language Conference. San Sebastian

My Contribution For this paper, I developed a mathematical model of distinctness based on dot products and metric spaces and provided a computational implementation, including a basic shift-reduce parser for applying the theory to existing experimental data. Using that implementation, I analyzed an existing experiment using traditional, factorial methods as well as using mixed-effects models to compare different numerical realizations of distinctness and wrote large portions of the introduction and discussion as well as all of experimental methods, results and conclusions.

Neuroinform
DOI 10.1007/s12021-013-9198-x

ORIGINAL ARTICLE

Towards a Computational Model of Actor-Based Language Comprehension

Phillip M. Alday · Matthias Schlesewsky · Ina Bornkessel-Schlesewsky

Abstract Neurophysiological data from a range of typologically diverse languages provide evidence for a cross-linguistically valid, actor-based strategy of understanding sentence-level meaning. This strategy seeks to identify the participant primarily responsible for the state of affairs (the actor) as quickly and unambiguously as possible, thus resulting in competition for the actor role when there are multiple candidates. Due to its applicability across languages with vastly different characteristics, we have proposed that the actor strategy may derive from more basic cognitive or neurobiological organizational principles, though it is also shaped by distributional properties of the linguistic input (e.g. the morphosyntactic coding strategies for actors in a given language). Here, we describe an initial computational model of the actor strategy and how it interacts with language-specific properties. Specifically, we contrast two distance metrics derived from the output of the computational model (one weighted and one unweighted) as potential measures of the degree of competition for actorhood by testing how well they predict modulations of electrophysiological activity engendered by language processing. To this end, we present an EEG study on word order processing in German and use linear mixed-effects models to assess the effect of the various distance metrics. Our results show that a weighted metric, which takes into account the weighting of an actor-identifying feature in the language under consideration outperforms an unweighted distance measure. We conclude that actor competition effects cannot be reduced to feature overlap between multiple sentence participants and thereby to the notion of similarity-based interference, which is prominent in current memory-based models of language processing. Finally, we argue that, in addition to illuminating the underlying neurocognitive mechanisms of actor competition, the present model can form the basis for a more comprehensive, neurobiologically plausible computational model of constructing sentence-level meaning.

Keywords Computational model · Language processing · Emergence · Ambiguity resolution · Actor identification

Parts of the research reported here were supported by the German Research Foundation (grant BO 2471/3-2) and the EEG study was performed while IBS was at the Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

P. M. Alday (✉) · I. Bornkessel-Schlesewsky
Department of Germanic Linguistics,
University of Marburg,
Deutschhausstr. 3, 35032 Marburg, Germany
e-mail: phillip.alday@staff.uni-marburg.de

M. Schlesewsky
Department of English and Linguistics,
University of Mainz,
Jakob-Welder-Weg 18, 55099 Mainz, Germany
e-mail: schlesew@uni-mainz.de

Introduction

The roughly 6000 languages of the world present a diverse set of grammars and input forms for the single processing mechanism of the human brain. Fundamental differences in word order, different means of encoding different parts of morphosyntax and broad variation in dropped / elided elements are just some of the variation with which the brain must cope; the complexity of a language is matched and exceeded by the complexity of language diversity. In light of the extreme variance between the languages of the world and their respective grammars, less syntax-bound language processing strategies have been proposed.

Published online: 04 August 2013

Neurophysiological data from a range of typologically diverse languages provides evidence for a comprehension and processing heuristic based on the notion of "actor", the participant primarily responsible for the state of affairs (Bornkessel-Schlesewsky and Schlesewsky 2009). The role of actor, while correlating strongly with certain parts of morphosyntax in some languages, is a language independent construct and is orthogonal to traditional notions of grammar (Bornkessel-Schlesewsky and Schlesewsky 2013a). Here, we present a computational implementation of the heuristic as well as a quantitative comparison with EEG data from an experiment primarily manipulating word order and its related ambiguities. We show that the actor heuristic is not just an interesting, qualitative theoretical construct, but rather a quantifiable and testable model. Indeed, we show that the quantification of the actor heuristic is a reliable, effective predictor of ERP data.

Neurophysiological Model and Language Processing Strategy

Before turning to the computational model that is the focus of the present paper, we will briefly describe the empirical neurocognitive model on which it is based. This discussion will serve primarily to introduce the critical notion of competition for the actor role, which will be central to the computational model to be introduced later. Having introduced actor competition, we will briefly summarize the empirical evidence in support of it.

The Extended Argument Dependency Model (eADM) and Actor-Centered Comprehension

The extended Argument Dependency Model ((e)ADM; Bornkessel 2002; Schlesewsky and Bornkessel 2004; Bornkessel and Schlesewsky 2006; Bornkessel-Schlesewsky and Schlesewsky 2008, 2009, 2013b) is a neurobiologically motivated, neurocognitive model of language comprehension with an explicit focus on cross-linguistic diversity. In other words, the model aims to account for language processing in typologically diverse languages and to explain which aspects of the processing architecture are universal and which are language-specific.

The eADM posits that language processing is organized in a cascaded, hierarchical fashion and proceeds along two major functional-neuroanatomical streams in the brain. One of these, the postero-dorsal stream, engages in time-dependent computations, while the other, the antero-ventral stream, engages in time-independent computations (Bornkessel-Schlesewsky and Schlesewsky 2013b). Time-dependent computation refers to the notion that, in the combination of two elements, A and B, the order in which they are encountered is crucial for the way in which they are combined. For example, in German, the order in which two noun phrases are encountered in an NP-NP-V sequence changes the likelihood for one being interpreted as the actor argument as opposed to the other. In time-independent computation, by contrast, the order of encountering two elements A and B will not influence the way in which they are combined. For example, a plausibility-based heuristic which computes the most plausible combination of arguments and the verb (e.g. given "apple" and "eat", the reading that the apple is the undergoer of the eating event rather than the actor) is independent of which element is encountered first. These time-independent computations are implemented in terms of schema unification (see below for a brief description of schemata and the ventral stream). Thus, time-dependent versus time-independent computations could also be described as "sequence-dependent" versus "sequence-independent" operations.

Processing in both streams is organized in a hierarchical manner in accordance with the neurobiological principle of hierarchical processing (Felleman and Van Essen 1991; Rauschecker 1998; Rauschecker and Scott 2009; DeWitt and Rauschecker 2012) and classic assumptions regarding the structure of complex cognitive models (Simon 1962; Newell 1990). This means that, as information flows along the streams, the representations that are processed are assumed to become increasingly complex.[1] In the following, we will refer to the successive points of information processing within the hierarchy as "processing steps" for convenience, though this is clearly a cognitive term that does not directly reflect the underlying neurobiological organization.

In a first step (ignoring preceding aspects of phonological processing and segmentation), the processing system identifies word categories and uses these to build a constituent structure ("syntactic structuring" within the postero-dorsal stream). Crucially, and in contrast to the assumptions of other comprehension models (Friederici 2002; Hagoort 2005; Vosse and Kempen 2000), this structure does not determine sentence interpretation: this is accomplished via a separate mechanism, as we shall see shortly. A second function of category processing in this step is to classify the current input element in terms of its function, e.g. whether it is referential ("nouny") or predicating ("verby").

[1] Note that, though the model is hierarchically organized, it is not modular in the traditional Fodorian sense (Fodor 1983). Firstly, due to the cascaded nature of processing, a particular processing step need not be fully complete before the next step is initiated. Secondly, from a neurobiological perspective, connections within each pathway are inherently bidirectional such that top-down modulations of information processing are always possible. Nevertheless, we assume that there is an asymmetry in the directionality of information flow based on the tenet of hierarchical organization.

In a second step, sentence-level interpretive mechanisms set in. In the postero-dorsal stream, the system determines sentence meaning from an action-based perspective by assessing who or what is primarily responsible for the state of affairs being described, i.e. here, the actor heuristic mentioned above comes into play. The antero-ventral stream, by contrast, constructs a schema-based representation of sentence-level meaning via the unification of "actor-event schemata". For reasons of brevity, we will not go into details with respect to the properties of these schemata or their unification, and focus instead on the "actor computation" step posited as part of the postero-dorsal stream. For a detailed discussion of actor-event schemata, the interested reader is referred to Bornkessel-Schlesewsky and Schlesewsky (2013b).

The notion of "actor computation" within the postero-dorsal stream is based on the assumption that a linguistic actor is a "stable, language-independent category, possibly rooted in the human ability to understand goal-directed action" (p. 250) (Bornkessel-Schlesewsky and Schlesewsky 2013a). The fact that humans are generally attuned to this category as opposed to others could be due to basic evolutionary demands. In the words of Leslie (1995): "Agents are a class of objects possessing sets of causal properties that distinguish them from other physical objects" and "as a result of evolution, we have become adapted to track these sets of properties and to efficiently learn to interpret the behaviour of these objects in specific ways" (p. 122). By tracking (potential) actors, i.e. those entities that appear suited to bringing about changes in the environment (e.g. warranting a fight-or-flight response), we can interpret the world around us and make predictions about upcoming events (see also Frith and Frith 2010). In accordance with this assumption, it has been demonstrated that the human attention system appears to have developed a special sensitivity towards humans and non-human animals (i.e. good potential actors) as opposed to other categories (New et al. 2007). In this way, the actor-centered comprehension strategy posited by the eADM essentially views a sentence as an instruction to conceptualize a particular scenario in which an actor is engaged in a certain event or state of affairs.

How are actor participants identified during language comprehension? In this regard, we have proposed that the prototypical actor may be modeled on the first person (i.e. the self as an acting agent, see Tomasello 2003, Haggard 2008). According to Dahl, this "self-as-actor" perspective is tied to humans perceiving conspecifics as being "like myself, individuals who can perceive the world and act upon it" (Dahl 2008, p. 149). Thus, in order to understand the environment around us, we use the self as a model for other animate entities (and particularly other humans), which in turn serve as a model for inanimate entities. In view of these considerations, the language comprehension system uses the features +self and +animate as cues to the identification of actor participants (see below for a summary list of actor features). Furthermore, in line with the notion that the self-as-actor perspective involves seeing others as individuals (i.e. other "selves"), an optimal actor is individuated (i.e. definite and specific). Finally, actorhood correlates with particular morphosyntactic features, which are partly cross-linguistically applicable (in particular: occurring as the first argument in a sentence) and partly language-specific (e.g. nominative case marking). Thus, the different features vary in applicability across languages; they also vary in their language-specific weighting, i.e. their importance to identifying actor participants in a particular language (cf. also MacWhinney and Bates 1989; Bates et al. 2001). We have posited that language-particular cues to actorhood (e.g. the importance of morphological case marking in a language such as German) are acquired via their high degree of co-occurrence with the universal actor features based on the first person model (Bornkessel-Schlesewsky and Schlesewsky 2013a). Thus, prototypical actor features derived from the self-as-actor perspective are used to bootstrap other, language-specific (morphosyntactic) features of actor participants—a view that is similar to that adopted by emergentist models of other linguistic categories such as parts of speech (Croft 2001).

Linguistic prominence features related to actor identification.

1. +self
2. +animate/+human
3. +definite/+specific
4. +1st position (correlates with actorhood cross-linguistically; (Tomlin 1986))
5. +nominative (correlates with actorhood in nominative-accusative languages with morphological case)

The degree to which arguments are good competitors for the actor role is defined by two points: (a) their own prototypicality in terms of the defining actor features and the correlating prominence features (see above), and (b) the existence and prototypicality of further competitors. Thus, an initial argument is preferentially analyzed as an actor even if it is not highly prototypical (e.g. if it is inanimate).[2]

[2]In this regard, the assumptions of the eADM differ from those of the Competition Model (e.g. Bates et al. 1982, 2001, MacWhinney and Bates 1989), which assumes that a strong cue for the undergoer role (e.g. accusative case in a language such as German or Hungarian) can, to all intents and purposes, exclude an argument from being considered a potential actor. The eADM, by contrast, posits that a sole argument is always considered for the actor role no matter how bad a candidate it is—unless there is a second, more optimal candidate (for discussion, see Bornkessel-Schlesewsky and Schlesewsky to appear).

Once a second argument—a competitor—is encountered, however, the relative actor prototypicality of the arguments is crucial in determining whether the actor preference for the first argument can be maintained or whether it needs to be revised.

In accordance with the model architecture in Fig. 1, increased competition for the actor role (including the need to revise a previous actor choice) correlates with increased activation in the posterior superior temporal sulcus and the temporo-parietal junction as part of the postero-dorsal stream (Bornkessel et al. 2005; Grewe et al. 2006; Bornkessel-Schlesewsky and Schlesewsky 2009) and the anterior temporal lobe as part of the antero-ventral stream (Magnusdottir et al. 2012; for discussion, see Bornkessel-Schlesewsky and Schlesewsky 2013b). In neurophysiological terms, it is reflected in increased amplitude of the N400 event-related brain potential (Bornkessel-Schlesewsky and Schlesewsky 2009, 2013a). We assume that the N400—and negative ERP deflections in general—result from a mismatch between top-down and bottom-up information sources within the two processing streams (for proposals that the N400 depends on an integration of top-down expectations and bottom-up input, see (Federmeier 2007; Lotze et al. 2011)). Crucially, as scalp ERPs are macroscopic responses which typically result from the mixing of multiple underlying sources, the claim is not that an N400 effect elicited by actor competition results from activation changes in only a single locus within a stream. Rather, it is likely due to the summed reaction of top-down/bottom-up integration within multiple processing steps within both streams (e.g. "actor computation" within the postero-dorsal stream and "actor-event schema unification" within the antero-ventral stream). In addition, since we posit that the notion of top-down/bottom-up integration can be generalized to other language-related negativities (e.g. left-anterior negativity (LAN) effects, which result, for example, from

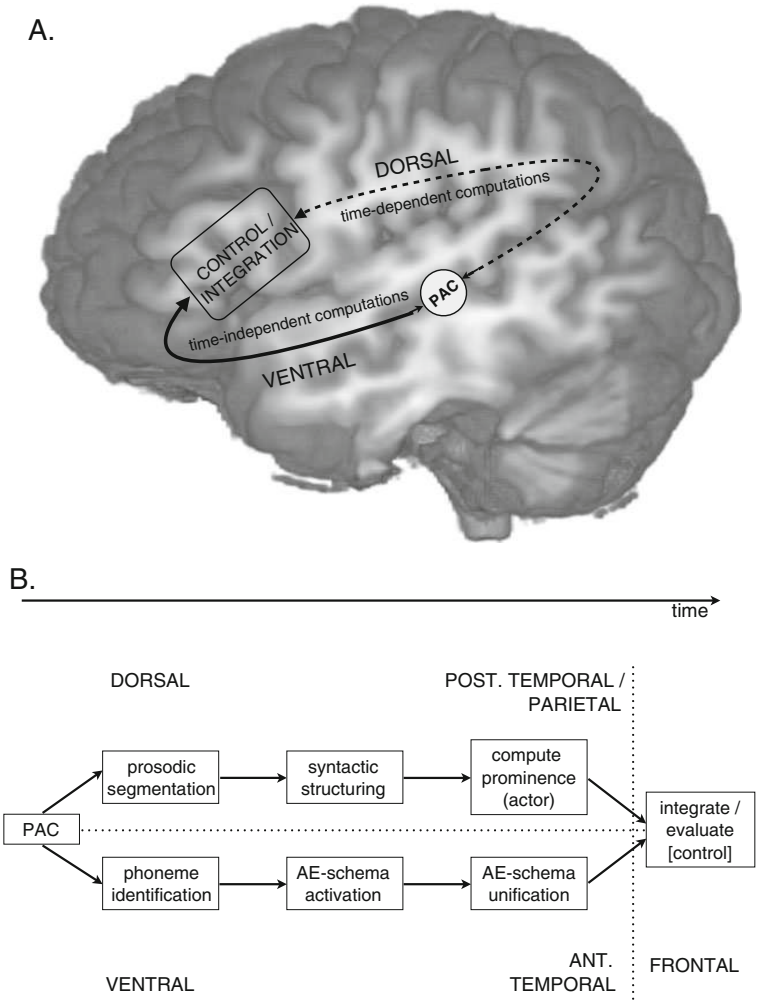


Fig. 1 Model architecture for the latest version of the extended Argument Dependency Model, eADM (adapted from (Bornkessel-Schlesewsky and Schlesewsky 2013b)). Panel A provides a basic overview of the model's neuroanatomical assumptions: the ventral (*solid line*) and dorsal (*dashed line*) streams are assumed to emanate from primary auditory cortex (PAC) and to perform information processing in a hierarchically organized manner. Thus, in spite of the fact that the streams are inherently bidirectional, there is an asymmetry in the directionality of information flow on account of the hierarchical organization. Panel B shows the assumed structure of hierarchical processing within the two streams

subject-verb agreement errors in which the expectation for a particular agreement morpheme is not met), latency and topography of the negativity response are assumed to vary depending on the loci within the streams giving rise to the mismatch and the timing of their activation (Sassenhagen et al. 2013).

When actor competition is behaviorally relevant (e.g. when participants perform an acceptability judgment task and actor competition affects how acceptable a sentence is deemed to be), it additionally engenders a late positive ERP response. In this view, late positivities in language processing ("P600" effects) are viewed as members of the domain-general P300 family (Coulson et al. 1998; Roehm et al. 2007; Kretzschmar 2010). Recently, Sassenhagen et al. (2013) linked this idea to a neurobiological model of the P300, the Locus Coeruleus-Norepinephrine (LC/NE-P3) model (Nieuwenhuis et al. 2005). According to this model, the P300 results from activation of the Locus Coeruleus (i.e. the brain stem source for noradrenergic projections to the cortex) following the detection of subjectively significant events. This results in a release of norepinephrine, thereby increasing neural responsivity to a particular stimulus and influencing the behavioral response to it. From this perspective, late positivities in language processing reflect a systemic neuromodulator release facilitating the application of decision processes rather than linguistic processing per se. In support of this view, Sassenhagen et al. (2013) found that an N400 – late positivity scalp ERP pattern engendered by semantically anomalous versus plausible words in a highly constrained sentence context could be decomposed using independent component analysis (ICA) and single-trial analyses. Results revealed that, while N400 effects were timelocked to critical stimulus onset, positivity effects were response-locked.

In summary, actor competition is reflected in N400–late positivity patterns in electrophysiological studies, though, as described above, the two components of this biphasic response are functionally distinct and, in principle, independent of one another. In addition, the presence or absence of the late positivity effect is conditioned by the experimental environment and task.

Evidence for the Actor Heuristic and for Competition for the Actor Role

The eADM's notion of competition for the actor role is supported by a range of cross-linguistic studies on sentence comprehension, which have provided evidence for the following generalization regarding online-processing:

> *Cross-linguistic generalization regarding actor identification in online language processing* (Bornkessel-Schlesewsky and Schlesewsky 2009):
>
> The processing system attempts to identify the actor role as quickly and unambiguously as possible.
>
> Corollaries:
>
> The processing system prefers actor-initial orders
>
> The processing system prefers prototypical actors.

Evidence for this generalization stems from electrophysiological studies in a number of typologically varied languages, including Turkish (Demiral et al. 2008), Chinese (Wang et al. 2009) and Hindi (Choudhary et al. 2010), thus corroborating previous findings of a "subject-first preference" in European languages (e.g. Dutch: Frazier (1987), German: Schriefers et al. (1995), Schlesewsky et al. (2000), Bornkessel et al. (2004b), amongst others; Italian: de Vincenzi (1991), Penolazzi et al. (2005)). Importantly, the empirical findings from non-Indo-European languages (Chinese and Turkish) support the assumption of an actor-first rather than a subject-first preference, since they rule out explanations based solely on formal subject features such as agreement. They further suggest that the actor-first preference cannot be reduced to structural simplicity or frequency (see Wang et al. (2009) for a summary). The finding of an actor-first preference even in an ergative language (Hindi) further demonstrates the need to assume an actor-first as opposed to a subject-first preference.[3]

The preference for prototypical actors shows up in a similarly ubiquitous way: When an argument that is unambiguously the actor in a transitive (two participant) relation is non-prototypical because it is inanimate, different languages consistently show an N400 effect (for a comprehensive review, see (Bornkessel-Schlesewsky and Schlesewsky 2009)). For an illustration, consider the following example from Frisch and Schlesewsky (2001):

(1) a. Paul fragt sich, welchen Förster
Paul asks himself, [which forester]$_{\text{ACC}}$
der Zweig ...
[the twig]$_{\text{NOM}}$ ...

b. Paul fragt sich, welchen Angler
Paul asks himself, [which angler]$_{\text{ACC}}$
der Jäger ...
[the hunter]$_{\text{NOM}}$ ...

[3] In an ergative language such as Hindi, the actor argument in a transitive (two-participant) event is not morphosyntactically "privileged" in the sense that it does not agree with the verb, for example. Thus, it does not qualify for grammatical subjecthood in the same way as a transitive actor in a non-ergative language such as German, Dutch or Italian. The results from Hindi thus provide strong converging support for the assumption that the actor preference is interpretive rather than grammatical in nature.

In example (1), the initial accusative—as a very poor actor candidate—leads the processing system to expect to encounter a better actor candidate as a second argument. When this expectation is contradicted by the features of an inanimate second argument (1a), which is also an atypical actor, an N400 effect arises in comparison to a control condition with an animate second argument (1b). Thus, as described in the preceding section, the N400 effects arises from a mismatch between top-down information (the expectation) and bottom-up information (the features of the second argument). In addition to German (Frisch and Schlesewsky 2001; Roehm et al. 2004) and English (Weckerly and Kutas 1999), this effect has been shown in Mandarin Chinese (Philipp et al. 2008) and Tamil (Muralikrishnan et al. 2008).

Why a Computational Model of Actor Competition?

As is apparent from the preceding section, the actor-based comprehension strategy is well described in qualitative terms. In formalizing this strategy with an implemented model, we pursue a twofold aim. Firstly, from the computational implementation of the actor strategy, we aim to gain quantitative predictions that can be tested against empirical data. This will allow for the predictions of the eADM to be tested in a more stringent manner and for cross-linguistic similarities and differences to be expressed in more explicit terms.

Secondly, these quantitative predictions can be used to illuminate the basic processing mechanisms underlying the actor strategy. In particular, we aimed to compare two alternative conceptualizations of actor competition: unweighted similarity-based interference and weighted competition. Similarity-based interference is a notion that features prominently in contemporary approaches to working memory (WM), which emphasize the status of WM as the activated portion of long-term memory rather than as a separate buffer (for overviews, see McElree (2006), Jonides et al. (2008); for approaches to language-processing based on this notion, see Lewis et al. (2006), Lewis and Vasishth (2005)). Accordingly, memory retrieval is conceptualized not as the result of a (serial) search, but of a content-addressable pointer mechanism based on so-called retrieval cues. These cues (for example, case, number or other features) provide the relevant information required to access the item in question. Retrieval becomes more effortful when cues overlap (i.e. apply to several items in memory), a phenomenon termed "similarity-based interference".

Similarity-based interference appears well suited as a potential mechanism underlying actor competition effects (Bornkessel-Schlesewsky and Schlesewsky 2013a), which, as described in detail above, arise when multiple candidates within a clause bear actor features. Since interference in its typical form (i.e. as conceptualized within the WM literature) is based exclusively on feature overlap, it predicts that the degree of competition for the actor role should be a function of the number of actor features shared by the arguments, while the specific weighting of a feature within a language should be irrelevant. More directly: feature overlap is an all-or-nothing measure for individual features because either two entities overlap in a given feature or not. By contrast, a second potential conceptualization of the degree of actor competition is that it goes beyond similarity-based interference in the classical sense and rather also takes into account the importance of a particular feature for actor recognition in the language under consideration. From this perspective, the degree of actor competition should be proportional to the difference in prominence (i.e. goodness of fit to the actor role) between arguments with individual features weighted according to their language-specific importance.

By means of the computational implementation introduced in the next section, we will calculate explicit metrics for the two alternative conceptions of actor competition outlined above and will test these against data from a neurophysiological experiment on sentence processing.

In this way, we endeavor to use the computational implementation of the actor strategy not only as a means of deriving more precise (quantifiable) predictions, but also to shed further light on the how the strategy is neurocognitively implemented.

Computational Implementation

The present implementation focuses on the core calculation of actor competition for referential elements, called Compute Prominence in previous versions of the eADM (Bornkessel-Schlesewsky and Schlesewsky 2009; Bornkessel and Schlesewsky 2006). For convenience, we similarly use the existing terminology "Stage 1" to refer to the initial chunking and analysis step and "Stage 2" to refer to the sentence-level interpretative mechanisms of the second step. A brief summary of the current software implementation can be found in the Technical Notes at the end of this article (p. 29).

The computational implementation does impose one restriction that the neurocognitive model upon which it is based does not: Stage 1 completes in full before Stage 2 begins. This is however not as detrimental to the approximation as it may initially seem because Stage 2 processes each constituent incrementally in the original sequential

order. Furthermore, both incremental and final full processing results for Compute Prominence are computed and optionally displayed.[4]

The completion of Stage 1 in its entirety before Stage 2 is unfortunately not capable of modeling cases where additional, disambiguating information becomes available. In German, this primarily happens in noun phrases via gender (indefinite NPs) and number (definite NPs) information available on the first non-article adjective or head noun. (Disambiguation via verb agreement is also possible, but this is an interaction with the computation for predicating elements—Compute Linking in previous versions of the eADM—and is not currently modeled for non verb-final word orders.) However, none of these forms of disambiguiation occur in the present experimental manipulation. Nonetheless, processing of ambiguities remains a major focus of present and future research.

Stage 1

The initial chunking and morphological analysis in Stage 1 is performed here only in a restricted fashion. The full complexity of German phrase structure would be a non-trivial undertaking and lies outside the scope of this paper and its focus on the actor heuristic (Stage 2) and an appropriate computational implementation. However, a sufficient implementation of Stage 1 to parse the stimuli from an EEG/ERP experiment with their relatively rigid structure is possible.

In the present experiment (see section "EEG Experiment"), it suffices to process inflection carried via pronouns and articles. In German, the article carries the majority of the morphological burden in noun phrases. The head noun inflects for number and can carry an additional marker for dative in the plural; however, this information was redundant in the present experiment, where neither dative nor plurals were used.[5] In the pronominal system, there exist a few ambiguities, especially between the nominative and accusative 3rd person for neutra and feminina as well as in the plural. This ambiguity could potentially be resolved by agreement with the verb; however, it also presents a general test case for the heuristic implemented by Compute Prominence. The ambiguity is thus marked as such and otherwise not further processed by Stage 1. The further ambiguity in the pronominal system between the 3rd person feminine dative and the second person plural nominative is always resolved by verbal agreement, but as there are no datives in this experiment, this special case is not processed further and is implemented by `pass` (a syntactic placeholder in Python similar to `void`) in the branch construct.

Stage 2

Implementation of Stage 2 was restricted to the function Compute Prominence, which provides in non-headmarking languages the most important parts of the actor heuristic in single-sentence processing.[6] Furthermore, Compute Prominence remains largely unchanged in recent and further planned updates to the eADM.

In implementing Compute Prominence, we view the hierarchies as dimensions in (a subspace of) Euclidean n-space, where n is the number of hierarchies. The prominence of an individual argument is thus a vector, with each component being a scalar representing the prominence with respect to a particular hierarchy. The hierarchies with the corresponding values for various linguistic features in the current implementation are given in Table 1. The "additional" feature NUMBER derives from another prototypical feature of the self-as-actor view: singular correlates with stronger individuation. Negative values are used to actively penalize a particular prominence component in the next calculation. That is, negative values indicate a feature that strongly correlates with a poor actor candidate (designated in the computational model by the feature +dep).[7]

For now, we make the a priori assumption that case is a singular feature with multiple levels cf. (Kempe and MacWhinney 1999). However, it is possible that "case" is merely a convenient moniker for a set of strongly correlating binary features such as ±nom and ±acc. Ambiguity would then be encoded by setting all individual case features to the same value, e.g. +acc, +nom for an ambiguity between nominative and accusative. This latter approach is the typical one found in NLP and has the interesting feature that individual cases can carry different, individual weights (see below). For example, +acc may be a much stronger indicator of a particular role assignment than +nom.

[4]This restriction exists primarily to simplify the implementation in a single Python program (see Technical Notes); to better model the waterfall data flow, coroutines could be used or Stage 1 and Stage 2 could be split into two programs connected by Unix pipes.

[5]Furthermore, the article also carries number information, albeit with a small ambiguity that is resolved through further adjectives or the marking on the head noun.

[6]Of course, contextual effects also play a role in normal language use.

[7]For the purposes of the present paper,+dep may simply be considered a convenient label for a poor actor candidate. For an in-depth discussion of ±dep and a motivation in terms of a previous version of the eADM, see (Bornkessel and Schlesewsky 2006)

Table 1 Prominence hierarchies and the corresponding scalar values for the various features as used in the current implementation of Compute Prominence

Feature	Hierarchy
Person:	First = 1 > Other = 0
Case:	Nominative = 1 > Dative = 0 > Accusative = −1
Animacy:	Animate = 1 > Inanimate = 0
Position:	Early = 1 > Late = 0
Number:	Singular = 1 > Plural = 0
Definiteness:	Definite = 1 > Indefinite = 0

Similarly, other features with multiple levels are encoded binarily according to their most prominent tendency.

±PERSON is actually ±1ST. PERSON and not a multi-tiered variable. Extending this to accommodate the second person would only require the addition of a further field in the prominence vector. As all functions in the implementation are written to handle vectors of arbitrary length, this would require no changes to the core code. However, the question remains open for the model development, whether representation as multiple fields or as a variable with more degrees of freedom is the sensible choice. Multiple variables allow for learning the weights (and hence the impact) of distinct levels separately; however, this potentially allows for unlikely combinations of multiple levels.[8] A final consideration in the weight encoding is the use of fuzzy logic for the boolean values. For example, an ambiguous noun phrase could be assigned values between zero and one to indicate some form of probability for a given analysis. A plant could be assigned an animacy value of 0.25 (alive and able to die but largely not capable of independent action), an animal could be assigned a value of 0.8 (alive and capable of independent action, but not sentient) and a human a value of 1.0 (alive, willful and sentient). Furthermore, this corresponds with animacy hierarchies seen in the languages of the world, with similar ordering, but language-specific cut-off points between levels (Silverstein 1976). Such gradience adds a flexibility to the use of binary features at the cost of making the prominence encoding somewhat less sparse.

The language-specific relative weights are also stored in a vector in the same space. The scalar (dot) product of the weight vector with the prominence vector yields a scalar value for the total prominence. This value is then compared with a threshold value to determine if +dep is assigned immediately. Compute Prominence is applied to both arguments and the values are compared, with the more prominent argument being assigned −dep (i.e. designated as the actor argument) and the less prominent argument +dep.[9] In the special case of an Object-Experiencer verb, prominence values for the case hierarchy are inverted: the entire hierarchy is multiplied by -1, thus reversing the orientation such that the accusative and dative outrank the nominative.

A sample sentence set for the EEG experiment used here as well as an analysis for a single condition of that set are given in Tables 2 and 3.

Distinctness/Actor Competition

The use of vectors to represent prominence data also allows for several other calculations to be made. The magnitude of the projection of an argument's prominence vector on the prominence vector for an idealized actor or undergoer is an index for the prototypicality of a particular argument. Similarly, the scalar product of the two argument vectors corresponds inversely with distinctness.

Distinctness is more broadly the distance between two arguments in actor-space. It thus provides a measure for the degree of competition for the actor role: when distinctness is low, multiple arguments bear actor features and competition for the actor role is high; when distinctness is high, actor features accumulate on only a single argument and competition for the actor role is low. Various metrics are provided (selectable as command line options in the implementation here) for the distance measurement. The Manhattan metric[10] reflects the summed distance between individual features. This is the default in the model and reflects an intuitive notion of distinctness. Furthermore, the Manhattan metric provides a general measurement of feature overlap and thus correlates inversely with traditional notions of interference—the fewer features that overlap/interfere, the larger the Manhattan distance. The Euclidean metric[11] is also provided and reflects a more continuous notion of distinctness. Finally, the difference in Euclidean magnitude of the two vectors is provided as a metric reflecting the difference in "absolute" (unweighted scalar) prominence. This magnitude difference also reflects the directionality of the prominence shift—the prominence of NP1 is subtracted from the prominence of NP2. Thus, NP2 is more prominent if and only if the magnitude difference is

[8] This is perhaps an advantage—in languages where inclusive and exclusive first person are morphologically distinct, this could be represented by the interaction of ±1st. Person and ±2nd. Person. This added complexity nonetheless introduces its own cost and brings language specific features deeper into the model.

[9] In the case of a single argument, e.g. intransitivity, the distinction measure is not performed and the model depends solely on the threshold comparison. The present experiment included only monotransitive sentences.

[10] $d(x, y) = \sum_i |y_i - x_i|$

[11] $d(x, y) = \sqrt{\sum_i (y_i - x_i)^2}$

positive. The selected metric is outputted (in batch mode; see Technical Notes) as the field `dist`. For the data here, this is the Manhattan metric, and so `dist` is equal to $\sum_i |NP2_i - NP1_i|$.

Distinctness as distance in actor space is calculated on the raw prominence vectors without the weight distortion. This has the simultaneous advantage and disadvantage that the distance is language independent.[12] Weighted distance is given by the field `sdiff` (in the batch mode output) and is calculated as the difference of the weighted, scalar prominences: $\vec{w} \cdot \vec{NP2} - \vec{w} \cdot \vec{NP1} = \sum_i w_i \cdot \text{NP2}_i - \sum_i w_i \cdot \text{NP1}_i$, where $\vec{w}$ is the language specific weight vector. Since the weighted scalar prominences are calculated by the dot product of the weight and feature vectors, this is equivalent to the weighted, signed Manhattan distance.[13] Baseline weights were based on previous work done in German in the framework of the Competition Model cf.(Kempe and MacWhinney 1999; MacWhinney et al. 1984). The qualitative ordinal scales were converted to quantitative interval scales via a simple order of magnitude mapping: for a feature f_1 ranked more strongly than another feature f_2, $f_1 = 10 f_2$ (see Stevens (1951) for the classification of scales). If a feature is considered a much stronger cue than another, then two orders of magnitude of separation was assumed: $f_1 = 100 f_2$. Thus, we have the following ordering: $case = 1000 > position$, $person = 100 > animacy$, $number = 10 > definiteness = 1$.[14] The exact numerical values have no empirical meaning in their own right. Rather their relationship to one another is central.[15] Changing the precise values will of course change the coefficients in the fitted mixed linear model, but will not change the properties of the model as a whole.[16]

In terms of the research questions introduced above, the main aim of the present paper is to compare the distinctness measures `dist` and `sdiff` as predictors of empirical neurophysiological data. While `dist` provides a good measure of similarity-based interference, `sdiff` implements the alternative, weighted notion of distinctness.

EEG Experiment

We tested the effectiveness of the model parameters as predictors of neurophysiological activity using data from an EEG experiment on word order processing in German.

The experiment manipulated actor competition by varying actor-undergoer order and case-marking ambiguity in transitive sentences with a noun phrase (NP1) – verb – noun phrase (NP2) structure. In particular, it examined sentences which—due to locally ambiguous case information—were initially compatible with an actor-first reading but subsequently required a reinterpretation as undergoer-initial. These are cases where actor competition is particularly high. They were compared with locally ambiguous sentences in which the actor-first preference was borne out and competition for the actor role was thus considerably less pronounced as well as with unambiguously case-marked sentences. Crucially for present purposes, the relative prominence of the two arguments—and hence their relative degree of actor prototypicality—was also manipulated in order to induce more subtle variations of actor competition. To this end, NP1 was either realized as a non-pronominal NP or as a 3rd person pronoun and NP2 was either realized as a non-pronominal NP or a 1st person pronoun. Recall from section "The Extended Argument Dependency Model (eADM) and Actor-Centered Comprehension" that optimal actorhood is assumed to be modeled on the first person within

[12] In as far as all features are treated equally—some languages may not take advantage of certain features, e.g. English largely does not use case.

[13] This follows very straightforwardly from the definition and standard properties of the dot product:

$$\begin{aligned}\vec{w} \cdot \vec{NP2} - \vec{w} \cdot \vec{NP1} &= \sum_i w_i \cdot \text{NP2}_i - \sum_i w_i \cdot \text{NP1}_i \\ &= \sum_i w_i \cdot (\text{NP2}_i - \text{NP1}_i) \\ &= \vec{w} \cdot (\vec{NP2} - \vec{NP1})\end{aligned}$$

[14] As noted by an anonymous reviewer, the EEG experiment presented below does not include any number or animacy contrasts. This however does not present any great problem for the data at hand: due to properties of the metrics at hand, the non contrasting features simply cancel out and do not even introduce additional parametric levels into the respective prominence metrics. These features remain in the models present because their presence does not detract from the comparisons in question and avoids an experiment-specific model. One subtle disadvantage does come into play here though: the fit of the weights for these two features is not tested. Especially our ranking of position relative to animacy may prove problematic and, as such, more explicit testing, manipulation and determination of model weights is planned for future research.

[15] This follows from the notion of an actor space—we can expand or contract the space by a constant multiple without changing the inherent properties of it. Specifically, $c\vec{v} \cdot c\vec{w} = c(\vec{v} \cdot \vec{w})$.

[16] Subject to the constraints of the effects this has on precision and representation on the computing machine in question. Theoretically, we could divide all of these values by 1000 (the maximum weight given here), giving us coefficients on [0, 1], which would reflect their impact in the notation of probability theory. This is a very interesting approach, as the deterministic impact of case would receive a (probability) coefficient of one—certainty. However, this all too easily leads to the assumption that there is necessarily a single feature which, when unambiguous, is singularly deterministic in its influence. Or, in the particular case of German, that the impact of case is always deterministic—clearly, this is not the case as all too often, the morphological marking is ambiguous: 0×1000 is still 0.

Table 2 Stimulus design in the EEG experiment. Every condition appeared for each lexical item

Initial	Ambiguous	NP1-Type	NP2-Type	Sentence(Example)
Actor	Yes	Noun	Noun	Die Bettlerin bedrängte den Kommissar auf der Straße.
Undergoer	Yes	Noun	Noun	Die Bettlerin bedrängte der Kommissar auf der Straße.
Actor	No	Noun	Noun	Der Bettler bedrängte den Kommissar auf der Straße.
Undergoer	No	Noun	Noun	Den Bettler bedrängte der Kommissar auf der Straße.
Actor	Yes	Noun	Pronoun (1sg)	Die Bettlerin bedrängte mich auf der Straße.
Undergoer	Yes	Noun	Pronoun (1sg)	Die Bettlerin bedrängte ich auf der Straße.
Actor	No	Noun	Pronoun (1sg)	Der Bettler bedrängte mich auf der Straße.
Undergoer	No	Noun	Pronoun (1sg)	Den Bettler bedrängte ich auf der Straße.
Actor	Yes	Pronoun (3sg)	Noun	Sie bedrängte den Kommissar auf der Straße.
Undergoer	Yes	Pronoun (3sg)	Noun	Sie bedrängte der Kommissar auf der Straße.
Actor	No	Pronoun (3sg)	Noun	Er bedrängte den Kommissar auf der Straße.
Undergoer	No	Pronoun (3sg)	Noun	Ihn bedrängte der Kommissar auf der Straße.
Actor	Yes	Pronoun (3sg)	Pronoun (1sg)	Sie bedrängte mich auf der Straße.
Undergoer	Yes	Pronoun (3sg)	Pronoun (1sg)	Sie bedrängte ich auf der Straße.
Actor	No	Pronoun (3sg)	Pronoun (1sg)	Er bedrängte mich auf der Straße.
Undergoer	No	Pronoun (3sg)	Pronoun (1sg)	Ihn bedrängte ich auf der Straße.

The base sentence (first example) translates to "The beggar hassled the commissioner in the street." The gender of NP was varied for the ambiguity condition; the person of NP2 was varied for the NP2-type condition. Abbreviations: 3sg = third person singular, 1sg = first person singular

the eADM. Accordingly, 1st person pronouns are optimal actors, 3rd person pronouns (which are not 1st person, but nevertheless highly individuated) are somewhat less optimal actors and non-pronominal noun phrases are somewhat less optimal again. By manipulating person rather than more commonly examined actor features such as animacy, the present study therefore allowed us to test the effectiveness of our computational implementation of actor computation (Compute Prominence) as well as the self-as-actor perspective.

Participants

Thirty-seven monolingually raised native speakers of German (20 women; mean age: 25.9 years, range: 20–40 years) participated in the EEG study after giving written informed consent. Participants were right-handed as assessed by a German version of the Edinburgh handedness inventory (Oldfield 1971). The majority of the participants were students at the Free University Berlin at the time of the experiment. Two additional participants were excluded due to technical problems or a failure to complete both experimental sessions.

Materials

The critical sentence types used in this study are shown in Table 2. Sixty sets of the conditions shown in Table 2 were constructed, thus resulting in a total of 960 critical sentences. These were subdivided into two lists of 480 sentences each (30 from each condition and 8 from each lexical set). The critical sentences for each list were pseudo-randomly interspersed with 240 filler sentences. Fillers

Table 3 Summarized analysis for the sentence *Die Bettlerin bedrängte den Kommissar auf der Straße* "The beggar hassled the commissioner in the street"

Feature	NP1	NP2	Weight
Case	0	−1	1000
Animacy	1	1	10
Person	0	0	100
Number	1	1	10
Definiteness	1	1	1
Position	1	0	100
Prominence			
Simple	5	2	
Weighted	121.0	−979.0	
Metrics			
`dist`	2		
`signdist`	+2		
`sdiff`	−1100.0		

Please note that the order of operation in computing the metrics matters: sum the pairwise differences (with absolute values, no weighting, or weighting, for `dist`, `signdist` and `sdiff`, respectively)

were also declarative main clauses of German but did not contain case or word order ambiguities. Eighty of the filler sentences were ungrammatical due to a case or agreement violation and 60 were semantically implausible, thus ensuring that participants needed to take into account both the grammaticality of the sentences and their plausibility when performing the acceptability judgement task (see below). The filler sentences were the same across the two lists. List presentation was counterbalanced across participants, with each participant reading the sentences from one list once.

Procedure

Participants were seated in a dimly lit, sound-attenuated booth, approximately 1 meter in front of a 17 inch computer screen. Sentences were presented visually in a phrase-by-phrase manner (i.e. noun phrases were presented together as chunks). Each trial began with the presentation of a fixation asterisk (presentation time: 300 ms, followed by an inter-stimulus-interval, ISI, of 200 ms). Single words were presented for 400 ms and phrases for 500 ms, with an ISI of 100 ms in each case. Following the presentation of the sentence-final word or phrase, there were 500 ms of blank screen, after which a question mark signalled to participants that they should judge the acceptability of the preceding sentence using two hand-held push-buttons. They were instructed that their judgement should be based both on form and content (i.e. also take into account the plausibility of the sentence). Assignments of the left and right buttons to "yes" and "no" responses were counterbalanced across participants. Following the judgement or after the maximal reaction time of 2000 ms had expired, there was an inter-trial interval (blank screen) of 1000 ms before presentation of the next sentence began. The experiment was conducted in two sessions, separated by approximately a week. In each session, a participant read 8 blocks of 45 sentences each, with blocks separated by short breaks. Sessions lasted approximately 3 hours including electrode preparation.

EEG Recording and Preprocessing

The EEG was recorded from 25 Ag/AgCl electrodes fixed at the scalp by means of an elastic cap (Easy Cap GmbH, Herrsching, Germany). AFZ served as ground. Electrodes were positioned according to the international 10-10 system. The electrooculogram was monitored by means of electrodes placed at the outer canthi of both eyes (horizontal EOG) and above and below the participant's right eye (vertical EOG). EEG and EOG channels were amplified by means of a Refa amplifier (Twente Medical Systems, Enschede, The Netherlands) and digitized with a sampling rate of 250 Hz. Channels were referenced to the left mastoid but rereferenced to linked mastoids offline. In order to eliminate slow signal drifts, a 0.3–20 Hz band-pass filter was applied to the raw EEG data. Trials containing EEG or EOG artifacts were excluded from the final data analysis (the EOG rejection criterion was 40 μV). For display purposes only, the grand average ERPs were smoothed with an 8Hz low-pass filter.

EEG Data Analysis

In an initial step, we performed a standard data analysis for language-related event-related brain potential (ERP) studies. Thus, average ERPs were calculated per condition, electrode and participant from the onset of the critical second noun phrase to 1000 ms post onset, before grand averages were computed over all participants. We then computed a repeated-measures ANOVA with the factors word order (actor-initial versus undergoer-initial), ambiguity (NP1 ambiguous between actor and undergoer versus unambiguously marked), NP1-Type (definite noun phrase versus 3rd person pronoun), NP2-Type (definite noun phrase versus 1st person pronoun) and region of interest (ROI). Lateral regions of interest were defined as follows: left-anterior (F3, F7, FC1, FC5); left-posterior (CP1,

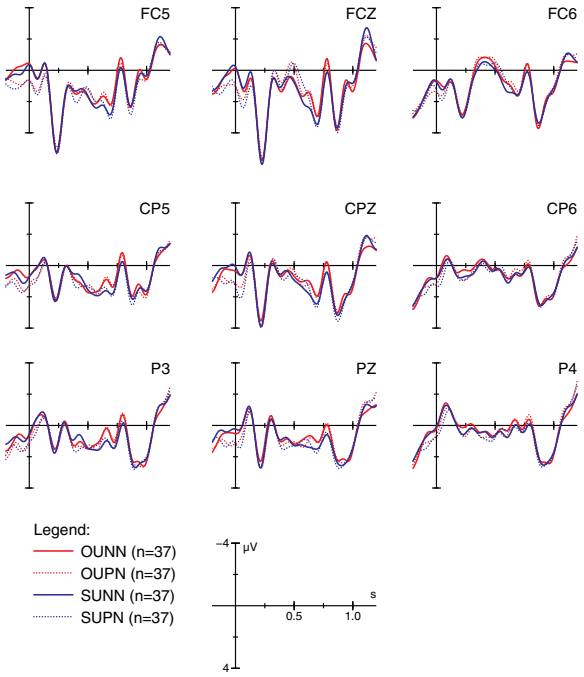


Fig. 2 Grand average ERPs triggered at the onset of NP2 for the unambiguous condition and NP2 a noun with definite article. The condition codes reflect the $2 \times 2 \times 2 \times 2$ design: S = subject (actor) initial word order, O = object (undergoer) initial; U = unambiguous, A = ambiguous; N = Noun, P = pronoun, for NP1 & NP2 respectively

CP5, P3, P7); right-anterior (F4, F8, FC2, FC6); right-posterior (CP2, CP6, P4, P8). A single ROI was used for the midline sites (FZ, FCZ, CZ, CPZ, PZ). For analyses involving more than one degree of freedom in the numerator, significance values were corrected when sphericity was violated (Huynh and Feldt 1970). This analysis was used to identify regions in which the effects were most pronounced for the subsequent analysis using linear mixed effects models, in which we tested the effectivity of the distinctness metrics as predictors of language-related electrophysisological activity.

Linear mixed effects models provide a tool capable of handling the random variation introduced by intersubject differences and lexical effects (Baayen et al. 2008), which are not modeled in the current implementation. Furthermore, they allow continuous predictors such as the actor metrics here, while ANOVA-based analyses do not. Using the R package `lme4` (Bates et al. 2013), we calculated models using subject and item as random factors, and the various distinctness measures as fixed factors. For the random factors, we used the maximal random-effect structure common to all models, i.e. random slopes grouped per distinctness measure, as models without random slopes are anti-conversative (Barr et al. 2013).[17,18] As an exact estimation of p-values in mixed effects models is not straightforwardly possible due to difficulties in estimating the degrees of freedom, we follow Baayen et al. (2008) in considering an absolute t-value exceeding 2 as an indication of significance at the 5 %- level.

Intermodel comparisons are also not completely straightforward, especially in the case of non-nested models. Most importantly, log-likelihood tests and the associated χ^2-statistic (i.e. the parallel to traditional ANOVA, even called via the function `anova()` in R) are only valid for nested models.[19] To compare non-nested models, we turn to information-theoretic criteria (cf. Burnham and Anderson 2002, p. 88). In particular, Akaike Information Criterion (AIC, (Akaike 1974)) and Bayesian Information Criterion (BIC, (Schwarz 1978)) provide further tools for comparing models, based on log-likelihood (fit) penalized by the number of parameters (overfitting). The absolute value of these scores is not meaningful in itself, but the general rule when comparing two models is "smaller is better". In the following, AIC and BIC are shown in the model summaries, while only AIC with log-likelihood and degrees of freedom for the fixed factor is shown in the model comparisons, since the comparison of non-nested models always involved models with the same number of parameters. For nested models, χ^2-statistics (based on likelihood ratio tests) are also shown.

[17] The models resolved for ambiguity in the P600 time window have only random intercepts, as models with random slopes failed to converge.

[18] Higher order interactions were excluded for three reasons. First, comparing models which differ in random-effect structure is less straightforward than those which differ in only fixed-effect structure. (Even for the fixed effects, the comparison between non nested models requires information-theoretic criteria, see main text.) Second, models with higher order interactions in the random-effects structure did not always converge and due to the aforementioned complexities of comparing random-effects structures, it is not clear which of several higher-order models to choose from. Finally, computational complexity increases extremely quickly with random effect complexity. Limiting the random-effects structure to the maximal common one provides an acceptable balance between estimation accuracy, ease of comparison, and computer time.

[19] Models in which the parameters for one model form a proper subset for the parameters of the other.

Model Performance and Prediction: Results

Behavioral Data

The results of the rating task showed that participants judged all conditions to be highly acceptable (lowest mean acceptability ratings were 86 % for the ambiguous, undergoer-initial condition with two non-pronominal noun

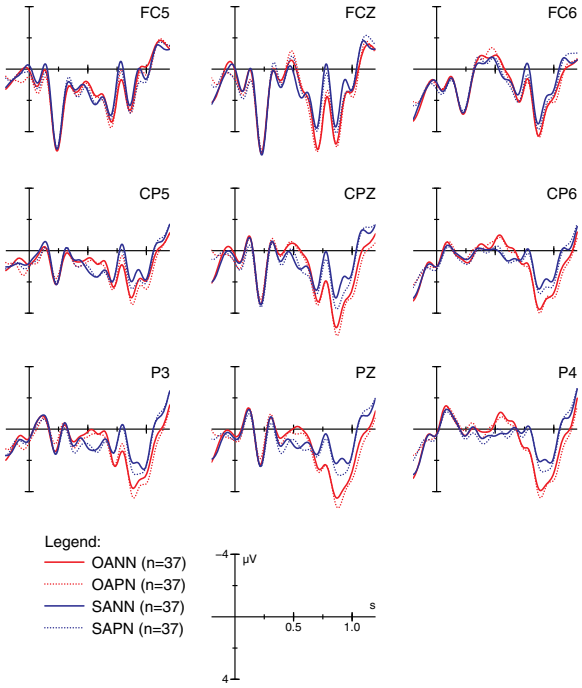


Fig. 3 Grand average ERPs triggered at the onset of NP2 for the ambiguous condition and NP2 a noun with definite article. The condition codes reflect the 2 × 2 × 2 × 2 design: S = subject (actor) initial word order, O = object (undergoer) initial; U = unambiguous, A = ambiguous; N = Noun, P = pronoun, for NP1 & NP2 respectively

phrases and 89 % for the ambiguous, undergoer-initial condition with NP1 a pronoun and NP2 a non-pronominal noun phrase; all other conditions showed an acceptability of 93 % or higher). We refrain from analyzing the ratings statistically in order to avoid interpreting ceiling effects. Most importantly for present purposes, they demonstrate that participants found the sentences acceptable and that they were able to correctly reanalyze the ambiguous undergoer-initial sentences (which should otherwise have been judged as unacceptable).

The analysis of the reaction times (restricted to sentences correctly judged as "acceptable") revealed an interaction of AMBIGUITY, WORD-ORDER and NP2-TYPE ($F(1, 36) = 4.95, p < 0.03$) and an interaction of AMBIGUITY and NP1-TYPE ($F(1, 36) = 8.45, p < 0.006$). Resolving both interactions by ambiguity showed an WORD-ORDER x NP2-TYPE interaction only for locally ambiguous ($F(1, 36) = 11.58, p < 0.002$) but not for unambiguous sentences ($p > 0.15$). For the ambiguous sentences, the interaction was due to longer reaction times for undergoer-initial as opposed to actor-initial sentences when the second noun phrase was non-pronominal / third person (mean RTs of 473 ms for undergoer-initial and 439 ms for actor-initial sentences; $F(1, 36) = 16, 77, p < 0.0003$), while there was no effect of actor-undergoer order when the second noun phrase was a first person pronoun ($p > 0.22$). The interaction of AMBIGUITY and NP1-TYPE was due to longer RTs for unambiguous sentences in which NP1 was realized as a non-pronominal NP as opposed to a first person pronoun (mean RTs of 443 ms and 431 ms, respectively; $F(1, 36) = 8.53, p < 0.006$).

In summary, reaction times were longer when sentences required a reanalysis towards an undergoer-initial order—as expected from the perspective of an actor-first preference—but this effect was only observable when the disambiguating second noun phrase was a non-pronominal third person, not when it was a first person pronoun. This finding provides converging support for the assumption that first person is a strong cue for actorhood, which can attenuate the behavioral reanalysis effect (for previous findings showing that strong cues for the target reading can attenuate reanalysis effects in behavioral data, though they are still observable electrophysiologically, see Bornkessel et al. 2004b).

Measures

The output from the implementation includes the selected distinctness metric as well as the scalar (weighted) difference in prominence for each item and condition (i.e. for each experimental stimulus). Additionally, the

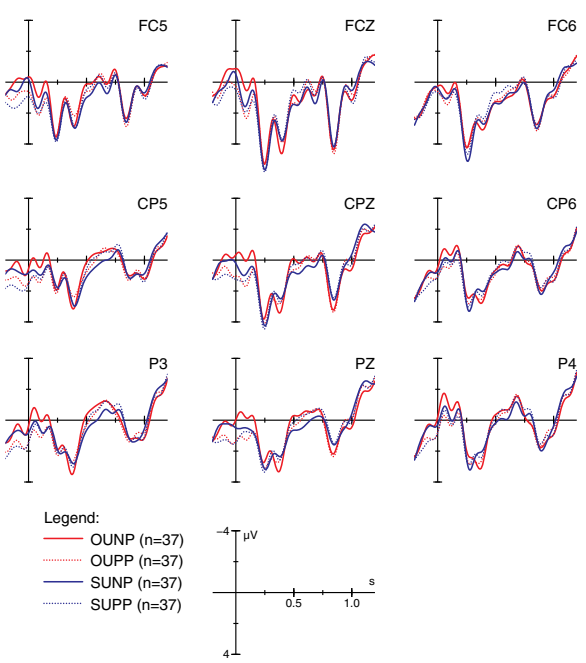


Fig. 4 Grand average ERPs triggered at the onset of NP2 for the unambiguous condition and NP2 a first person pronoun. The condition codes reflect the 2 × 2 × 2 × 2 design: S = subject (actor) initial word order, O = object (undergoer) initial; U = unambiguous, A = ambiguous; N = Noun, P = pronoun, for NP1 & NP2 respectively

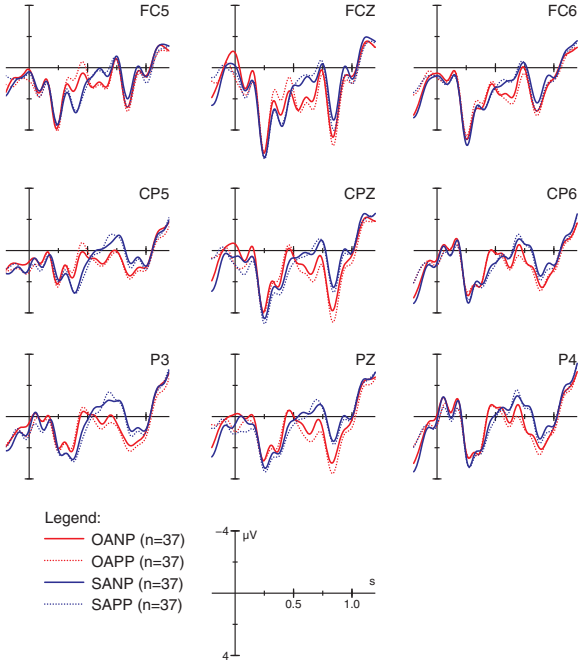


Fig. 5 Grand average ERPs triggered at the onset of NP2 for the ambiguous condition and NP2 a first person pronoun. The condition codes reflect the 2 × 2 × 2 × 2 design: S = subject (actor) initial word order, O = object (undergoer) initial; U = unambiguous, A = ambiguous; N = Noun, P = pronoun, for NP1 & NP2 respectively

implementation also outputs prominence scores calculated for NP1 and NP2. We duplicate this data for all subjects and enter it into the dataframe for EEG data. Based on visual inspection and significance testing via repeated measures analyses of variance (ANOVAs), we restricted the analysis to a subset of the data.

First, the time window for the N400 was found to be about 300-500ms post stimulus onset for the pronouns, and about 100ms later (400–600ms post onset) for the nouns (cf. Fig. 2 & 3 vs. Fig. 4 & 5). A similar effect was found for the late positivity (P600) time window: 600-800ms post onset for the pronouns and 700-900ms post onset for the nouns. As such, the time windows were parameterized in the model: "N400" vs "P600", with the exact time window reflecting whether the target stimulus was a noun or a pronoun. This difference in latency is not unexpected: both frequency and length are known to influence the latency of exogenous ERP components. Pronouns being highly frequent and short (a classic example of Zipf's Law; (Zipf 1935, 1949; Manning and Schütze 2000)) thus elicit a somewhat earlier effect. This is predicted by the cascaded architecture of the eADM: for shorter words, the information that is necessary for processing to proceed to the next step accrues more quickly.

Our choice of relatively traditional windows for the N400 and late positivity should thus not be taken as reflections of an a priori assumption about the ontological latency of these components. As previously mentioned, the decisive attribute of a component is its polarity; latency is to some extent an indication of the amount of processing necessary to reach the computational step reflected by a particular component. Amplitude is meaningful as a vague correlate of processing power needed at a particular step; however, due to well-known issues with equivalent dipoles, cancellation, etc., amplitude of scalp EEG is not a monotonic function of processing effort.

Furthermore, the ANOVA performed across five regions of interest (four quadrants and midline) revealed the strongest effects and interactions in the left posterior ROI, and for simplicity and computability, we restrict our model fitting to this ROI. The relevant ANOVA results are summarized in the Appendix (Tables 32 and 33).[20]

[20]More rigorous methods are available for dynamically determining the time window and topographical distribution of components. Maris (2004) and Maris and Oostenveld (2007) propose the necessary methods for non parametric method testing and determination of the effects in time and space (topography). Issues of computational tractability as well as data set size (different (sub)sets of data have to be used for determining the spatiotemporal distribution and testing it) reaffirmed our decision against introducing too many non-traditional methods for this initial computational model.

N400

We begin with the parametric time window "N400". Here, we first generated the most basic models using only the distinctness measures as fixed effects. However, we note that neither `sdiff` nor `dist` explicitly encode the experimental parameter ambiguity, i.e. the degree of evidence for an actor or non-actor analysis of the first argument based on morphological case marking, the strongest cue to actorhood in German. In German, there are two possible ways of deterministically resolving locally ambiguous case marking: (a) the elimination of possibilities by another unambiguously marked argument and (b) agreement with the verb. It is, however, generally accepted in the psycholinguistic and neurolinguistic literature that sentence processing proceeds incrementally, i.e. the processing system uses strategies to resolve local ambiguities even in the absence of clear evidence for one or the other reading in the input (Marslen-Wilson 1973; Crocker 1994). In this experiment, ambiguity resolution was provided via (a) on the second argument, which means that the initial ambiguity affects both which predictions the language system is able to make initially and how much new information becomes available at NP2. It thus makes sense to see how the parameter ambiguity interacts with our distinctness measures. For `sdiff` (Table 4), a likelihood ratio test reveals a significant improvement (Table 5). Similarly, we find an improvement for `dist` (Table 6), albeit a smaller one (Table 7). The comparison between the models with ambiguity, which are shown in

Table 4 Summary of model fit for `sdiff` (weighted distinctness) and ambiguity in the N400 window

Linear mixed model fit by maximum likelihood			
AIC	BIC	logLik	deviance
395191	395291	−197584	395169
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.18	0.42
	c.(sdiff)	2.3e−08	0.00015
subj	(Intercept)	1.2	1.1
	c.(sdiff)	2.6e−08	0.00016
Residual		17	4.1
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	1.2	0.19	6
ambiguityunambig	0.5	0.031	16
c.(sdiff)	−0.00039	4e−05	−9.8
ambiguityunambig:c.(sdiff)	0.00038	2.5e−05	15

Table 5 Statistics for models in the N400 time window based on the `sdiff` metric, showing the effect of ambiguity with interaction

	Df	AIC	logLik	Chisq	Chi Df	Pr(>Chisq)
sdiff: mean ~ c.(sdiff) + ...	9	395675	−197828.97			
sdiff.ambiguity: mean ~ ambiguity * c.(sdiff) + ...	11	395190	−197584.42	489.11	2	<2.2e−16 ***

(Random effect structure elided. See page 12)

Tables 4 and 6 for `sdiff` and `dist`, respectively, show that `sdiff` provides a better fit to the data.

Examining the models more closely, we see that the interaction between ambiguity and `dist` was not significant. In light of this missing interaction with `dist`, we can also consider using ambiguity as a simple model parameter that does not interact with our distinctness measures. In this case, we find a significant improvement for `dist` (Table 8) over the model without any ambiguity, and, moreover, the model with interaction does not differ significantly from the one without (Tables 9 and 10).

It thus appears that the interaction with ambiguity was particular to `sdiff`.

At this point, it is important to note that `sdiff` differs from `dist` not only in its weighting, but also in its use of "directionality" by being a signed value. We can also calculate a signed version of `dist`, termed `signdist`, by the sum of the pairwise differences: $\sum_i (\mathrm{NP2}_i - \mathrm{NP1}_i)$. This is the same as the Manhattan metric without absolute value signs or `sdiff` with all weights equal to one (via associativity of addition and subtraction, see Footnote 13, p. 9). Intuitively, this measurement is the net change in prominence features—a negative value indicates fewer prominence features, while a positive value indicates more positive features. As with the other distance measures, the parameter ambiguity improves model fit significantly. Tellingly, the minimally adequate model for `signdist` with ambiguity (Table 11) does not differ from the minimally adequate model for `dist` (Table 12); see Table 13 for a direct comparison between the minimally adequate models for all 3 predictors.

Table 6 Summary of model fit for `dist` (feature overlap) and ambiguity in the N400 window

Linear mixed model fit by maximum likelihood			
AIC	BIC	logLik	deviance
395140	395240	−197559	395118
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.18	0.43
	c.(dist)	0.057	0.24
subj	(Intercept)	1.2	1.1
	c.(dist)	0.12	0.35
Residual		17	4.1
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	1.3	0.19	6.6
ambiguityunambig	0.26	0.036	7.2
c.(dist)	0.24	0.07	3.5
ambiguityunambig:c.(dist)	−0.012	0.036	−0.33

Holding ambiguity constant to examine the interaction in more depth, we can again compare `sdiff` and `dist`. We find that they do not differ for unambiguous sentences; however, `sdiff` performs substantially better than `dist` and even `signdist` as a model predictor for ambiguous sentences (Table 14). This is immediately apparent in Figs. 6–10.

It is clear that `dist` behaves roughly the same, regardless of ambiguity, while `signdist` and `sdiff` interact with ambiguity—directionality clearly plays a role in the ambiguous condition. We even see that the slope in the ambiguous condition for `dist` is actually in the opposite direction of the other two predictors.

In Figs. 7 and 8, we can observe some difference between `signdist` and `sdiff`, in particular that the confidence interval is broader for `signdist`, which indicates a poorer fit. When we visualize the data in three dimensions as contour plots instead of as two subplots, the difference becomes even clearer (Figs. 9 and 10). Color indicates height, variation in color thus means variation in height, i.e. slope. Level curves, like in a topography map, indicate the overall shape of the landscape. The flat coloring in the unambiguous conditions for `signdist` and `sdiff` is indicative of the amount of variation being very small in comparison to the variation by the ambiguous condition. Moreover, `sdiff` shows a much more nuanced behavior in the ambiguous condition than `signdist`—this is clearly visible in the spacing between contour lines and their respective heights (difference between neighboring colors in the figures). The combination of weightedness and direction is much more telling about the processing of ambiguities than direction alone.

The `sdiff` model for the N400 time window reveals a strong negative correlation between `sdiff` and mean ERP

Table 7 Statistics for models in the N400 time window based on the `dist` metric, showing the effect of ambiguity with interaction

	Df	AIC	logLik	Chisq	Chi Df	Pr(>Chisq)
dist: mean ~ c.(dist) + ...	9	395188	−197585.02			
dist.ambiguity: mean ~ ambiguity * c.(dist) + ...	11	395139	−197558.87	52.31	2	4.37e−12 ***

(Random effect structure elided. See page 12)

amplitude, especially in the ambiguous condition (cf. sign of the *t*-statistic in Table 15, gradient direction in Fig. 8).

The decrease in the mean reflects the negativity in the ERP response, while the increase in prominence reflects an undergoer-first word order. For `signdist`, we see a weaker, yet similar effect. For `dist`, we see a positive correlation with ERP amplitude (`dist` is non negative per definition, but the models used centered values): the more features that don't overlap, the greater the mean, and hence, the smaller the negativity. In the unambiguous conditions, the correlation between mean ERP amplitude and `sdiff` is not significant (Table 16); however, we again see that the sign remains negative.

Additionally, the interaction of signedness and weightedness in `sdiff` expresses itself twofold. Signedness is a form of directionality and leads to a better gradient structure, and this becomes especially important when the correct directionality is not initially clear, namely in the ambiguous condition. This provides for the similarity in structure between `sdiff` and `signdist` that we see in Figs. 7 and 8 in contrast to the level structure of `dist` (Fig. 6). The weightedness of `sdiff` then contributes a further, necessary granularity to the model. Directionality provides information about the direction of change, while weightedness contributes additional information about the amount of change.

From the models tested, the best is then the one using `sdiff` interacting with ambiguity for its fixed effects (Table 13).

Finally, in an additional post-hoc test, we can compare `sdiff` to a traditional, unweighted syntactic measure—i.e. a subject as opposed to an actor strategy—using new metrics, `syndist` and `synsigndist`, which are restricted to the features PERSON, NUMBER and CASE. The unweightedness follows from the all-or-nothing principles of agreement and case marking—either a verb and a noun agree or they don't / a noun is either nominative or it isn't. As is evident in Table 17, even under these experimental circumstances, without global ambiguity, where the deterministic case marking of German provides for a clear syntactic analysis, the prominence-based model fares better. Nevertheless, since cues to actorhood and subjecthood show considerable overlap in the present experimental design, this result can only be taken as a tentative initial indication that an actor-based strategy outperforms a subject-based strategy when the two are tested against each other via computational modelling.

Table 8 Summary of model fit for `dist` (feature overlap) and ambiguity in the N400 window (without interaction)

Linear mixed model fit by maximum likelihood			
AIC	BIC	logLik	deviance
395138	395229	−197559	395118
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.18	0.43
	c.(dist)	0.057	0.24
subj	(Intercept)	1.2	1.1
	c.(dist)	0.12	0.35
Residual		17	4.1
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	1.3	0.19	6.6
ambiguityunambig	0.26	0.036	7.2
c.(dist)	0.24	0.067	3.5

Late Positivity (P600)

Visual inspection of the ERP data (Figs. 2–5) suggests a secondary effect in the form of a late positivity, which is line with previous findings on undergoer-initial ambiguous sentences in German (Haupt et al. 2008). As in the N400 time window, the models including ambiguity as an additional fixed factor perform far better than those without ambiguity. Similarly, `dist` shows no interaction with ambiguity (Tables 18 and 19), while `sdiff` and `signdist` do (Tables 20, 21, 22 and 23).

In accordance with the reverse in polarity over a biphasic reaction, we also see a reverse in effect direction: whereas a more highly positive `sdiff` correlated with a decreased mean (negativity) in the N400 time window, it correlates with an increased mean (positivity) in the P600 time window. Similarly, `dist` now correlates with a decreased mean (negativity, or here, lack of a positivity).

Table 9 Statistics for models in the N400 time window based on `dist` metric, showing the effect of ambiguity without interaction

	Df	AIC	logLik	Chisq	Chi Df	Pr(>Chisq)
dist: mean ~ c.(dist) + ...	9	395188	−197585.02			
dist.ambiguity.no_int: mean ~ ambiguity + c.(dist) + ...	10	395137	−197558.92	52.20	1	5.00e−13 ***

(Random effect structure elided. See page 12)

This is reflected in the respective *t*-statistics (Tables 18, 20 and 22): their signs have reversed.

A direct comparison of the minimally adequate models for each predictor can be found in Table 24.

Following the smaller effect size of the late positivity, the models do not differ by much. (The apparent trivial advantage for `dist` in AIC stems from it having fewer degrees of freedom, i.e. a smaller overfitting penalty.) However, upon resolving the interaction, we again see a greater differentiation in the ambiguous but not in the unambiguous condition[21] (Tables 25–28). In the ambiguous conditions, `sdiff` outperforms `dist`—as in the N400 time window.

This is also clearly reflected in Figs. 11, 12 and 13 where the gradience reflected in the unambiguous condition differs from the unambiguous condition for `sdiff` but not `dist` (see also Figs. 14 and 15 for a comparison of `signdist` and `sdiff` split by ambiguity). Interestingly, the difference in variance in the mean is overall less in both conditions: this is reflected by the narrower confidence intervals (Figs. 11–13) and in the visibility of the color gradient for the unambiguous condition for `sdiff` (Fig. 15). The latter is indicative of the variance in the ambiguous condition being comparable enough to the unambiguous condition that the same scale provides the necessary resolution for both conditions. The reversal in effect direction is also apparent in the reversal of the color schemes for the contour plots.

While the effect in the P600 time window is smaller than in the N400, the general trend is nonetheless clear: increased prominence of the second argument compared to the first leads to an increase in the mean amplitude in the later time window. More succinctly, we see a positivity in the P600 time window for an object-initial word order. Taken together with the N400 for the object-initial word order, we have a biphasic pattern for object-initial sentences.

As discussed in section "The Extended Argument Dependency Model (eADM) and Actor-Centered Comprehension", the eADM posits a functional distinction between the two components comprising the biphasic pattern observed here. While the N400 is assumed to reflect actor competition per se (including its resolution), the late positivity is assumed to index the behavioral reorientation induced by subjectively significant (task-relevant) events. In the present study, sentences requiring a reanalysis of the actor-first preference entailed such a reorientation since the degree of actor competition was relevant for participants' completion of the judgement task. As this explanation presupposes that (in contrast to the N400 effect) the positivity effect is reaction-locked rather than stimulus-locked, we computed an additional analysis in which we included logarithmically transformed mean reaction times per participant and condition into the mixed effects models as continuous predictors. While both models are greatly improved by including average reaction time by subject for each condition (i.e. single-subject averages) as a factor (Tables 29 and 30), the improvement is much greater (many orders of magnitude) for the late positivity window, as would be expected for an effect of task.

Discussion

We have presented a computational model that implements the actor strategy in language comprehension. The predictions of the model were tested against the results of an empirical study using event-related brain potentials (ERPs). Specifically, we examined the predictive capacity of two metrics for computing argument distinctness (i.e. degree of competition for the actor role): the unweighted distance measure `dist` (the Manhattan metric) and the weighted scalar difference measure `sdiff`. While both measures proved to be statistically significant predictors of N400 - late positivity amplitude, `sdiff` provided better model fits than `dist`. This was apparent particularly in ambiguous sentences, which, in some cases, called for a reanalysis towards an undergoer-initial order. Moreover, though this was not the primary focus of the present study, the current results provide an initial indication that the `sdiff` metric of actor computation provides a better fit to the electrophysiological data than a metric based purely on cues to syntactic subjecthood. They further show that the N400 and late positivity responses can be dissociated in that the latter is

[21] The slightly better performance of `dist` in this comparison of the unambiguous conditions is twofold: (1) it has fewer degrees of freedom and hence a smaller overfitting penalty in the AIC measure, and (2) the positive-only nature of `dist` lines up with the directionality of the positivity (but not the negativity).

Table 10 Statistics for models in the N400 time window based on `dist` metric, comparing the modelling of ambiguity with and without interaction

	Df	AIC	logLik	Chisq	Chi Df	Pr(>Chisq)
dist.ambiguity.no_int: mean ~ ambiguity + c.(dist) + ...	10	395137	−197558.92			
dist.ambiguity: mean ~ ambiguity * c.(dist) + ...	11	395139	−197558.87	0.11	1	7.41e−01

(Random effect structure elided. See page 12)

tied more closely to participants' behavioral reactions than the former. In the following, we will first discuss the evidence supporting a weighted as opposed to an unweighted distance metric and the architectural consequences arising from this result, before turning to implications for the functional interpretation of the N400 and late positivity in language processing tasks. We will then describe how this initial computational model of the actor strategy might serve to advance the development of a neurobiologically plausible model of actorhood computation. Finally, we will describe some future directions resulting from this work.

Evidence for and Consequences of a Weighted Distance Metric

As mentioned in section "Distinctness/Actor Competition", `dist` (the Manhattan metric) roughly corresponds with "feature overlap" or traditional notions of similarity-based interference, such as those in memory-based models of language processing (e.g. Lewis et al. 2006; Lewis 2000; Lewis and Vasishth 2005; McElree et al. 2003). By contrast, `sdiff` takes into account the language-specific weighting of the actor-related features. For German, this weighting places particular emphasis on the function of unambiguous case marking, which is the dominant cue to actor assignment when it is available (MacWhinney et al. 1984). In addition, it allows for coalitions of features to form to overcome "deficient" (ambiguous) case marking (Bates et al. 1982). The strong role of case as well as well the supplementary role of coalitions of weaker cues is most obvious in comparing the ambiguous and unambiguous conditions.

The predictive power of `sdiff` comes not only from its weighting, but also from its directionality, which serves to model the incremental demands of language processing, including the development (and possible fulfillment) of expectations. In relation to the current experiment, the directionality of the `sdiff` measure (negative or positive) essentially reflected the degree to which the parser's expectations about the prominence of the second argument were met. When `sdiff` was positive, the second argument was more prominent than the first, thus requiring a revision of the initial actor analysis of the first argument. This was reflected in a biphasic N400 - late positivity pattern, as was already observed in previous studies on actor-reanalysis in German (Haupt et al. 2008). As noted in section "Evidence for the Actor Heuristic and for Competition for the Actor Role" we interpret the negativity as an index of actor competition (leading to reanalysis of the initial actor-first preference in this case) and the late positivity as reflecting a behaviorally significant categorization of the sentences as less well-formed. This categorization reflects

Table 11 Summary of model fit for `signdist` (directed, net change in prominence features) and ambiguity in the N400 window

Linear mixed model fit by maximum likelihood			
AIC	BIC	logLik	deviance
395396	395496	−197687	395374
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.17	0.42
	c.(signdist)	0.015	0.12
subj	(Intercept)	1.2	1.1
	c.(signdist)	0.015	0.12
Residual		17	4.1
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	1.2	0.19	6.1
ambiguityunambig	0.5	0.031	16
c.(signdist)	−0.2	0.03	−6.8
ambiguityunambig: c.(signdist)	0.2	0.019	11

Table 12 Statistics for the minimally adequate models for each unweighted predictor in the N400 time window

	Df	AIC	logLik
dist.ambiguity.no_int: mean ~ ambiguity + c.(dist) + ...	10	395137	−197558.92
signdist.ambiguity: mean ~ ambiguity * c.(signdist) + ...	11	395395	−197686.89

(Random effect structure elided. See page 12)

Table 13 Statistics for the minimally adequate models for each predictor in the N400 time window: `sdiff` and `signdist` interact with ambiguity, `dist` does not

	Df	AIC	logLik
dist.ambiguity.no_int:	10	395137	−197558.92
mean ~ ambiguity + c.(dist) + ...			
sdiff.ambiguity:	11	395190	−197584.42
mean ~ ambiguity * c.(sdiff) + ...			
signdist.ambiguity:	11	395395	−197686.89
mean ~ ambiguity * c.(signdist) + ...			

(Random effect structure elided. See page 12)

the unmotivated positioning of the undergoer argument in a position that linearly precedes that of the actor.

Crucially, however, the advantage of `sdiff` over `dist` as a predictor of language-processing related neurophysiological activity cannot be reduced to the directionality of the `sdiff` metric. This was shown by the comparison of the two basic metrics with a directional (signed) version of `dist`, `signdist`. For locally ambiguous sentences, model fits involving `sdiff` were better than those involving both `dist` and `signdist`, thus attesting to the fact that both directionality and feature weighting contribute to the advantage for `sdiff`. Both N400 and late positivity amplitude for a revision of the actor-first preference were modulated by NP2 prominence (i.e. depending on whether NP2 was a first person pronoun or a definite noun phrase) and the magnitude of this modulation was predicted more accurately by the more fine-grained, weighted `sdiff` metric than by the unweighted `dist` and `signdist` metrics.

An additional divergence between the `sdiff` and `dist` measures is apparent in the model fits for the late positivity time window. Here, `sdiff` showed directionally opposite effects for ambiguous and unambiguous sentences: in the ambiguous cases, more positive `sdiff` correlated with higher positivity amplitude (as described above), while, for unambiguous sentences, more positive `sdiff` correlates with decreased positivity amplitude. By contrast, `dist` does not differentiate between ambiguous and unambiguous sentences, as demonstrated by the fact that the interaction of

Table 14 Statistics for models in the N400 time window with NP1 ambiguous

	Df	AIC	logLik
dist: mean ~ c.(dist) + ...	9	197520	−98751.07
sdiff: mean ~ c.(sdiff) + ...	9	197353	−98667.50
signdist: mean ~ c.(signdist) + ...	9	197477	−98729.61

(Random effect structure elided. See page 12)

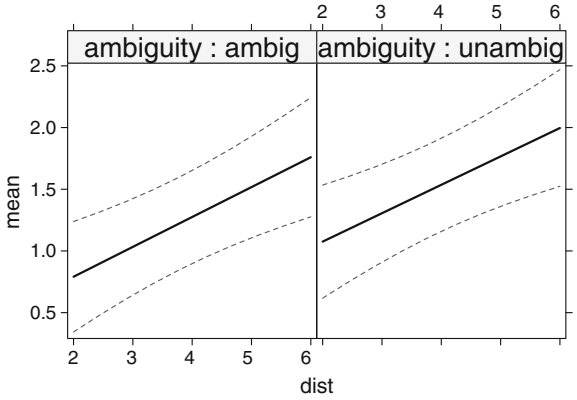


Fig. 6 Mean EEG in the N400 time window as a function of `dist` (*centered*) and its interaction with ambiguity. *Dashed lines* indicate the 95 % confidence interval

`dist` and ambiguity can be removed from the model without affecting model fit. Thus, in unambiguous sentences, the data show a small, late positivity for actor- as opposed to undergoer-initial sentences. We posit that this could again be explained in terms of predictability in online processing. Specifically, unambiguous initial accusative marking (sentences with very strongly negative `sdiff`) allows for comparatively more prediction: in contrast to an initial nominative, it is apparent that the construction is transitive and that a second argument is required (Bornkessel et al. 2004a; Wolff et al. 2008). Accordingly, unambiguous nominative-initial (actor-initial) engender a slightly increased late positivity at the less predictable NP2 in comparison to their accusative-initial counterparts. Interestingly, we observed no such effect in the earlier time window. This supports the perspective that the N400 reflects competition for the

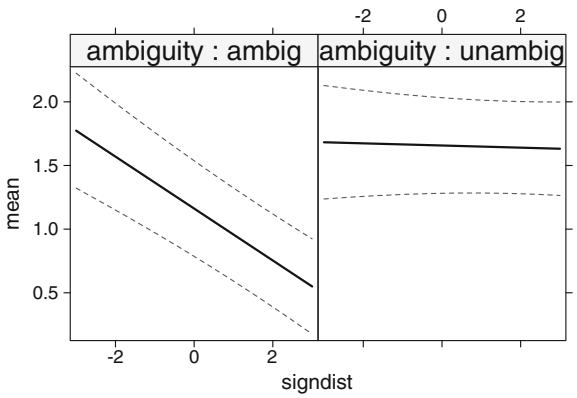


Fig. 7 Mean EEG in the N400 time window as a function of `signdist` (*centered*) and its interaction with ambiguity. *Dashed lines* indicate the 95% confidence interval

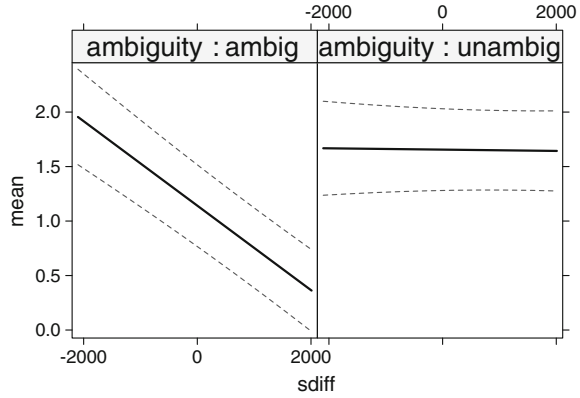


Fig. 8 Mean EEG in the N400 time window as a function of `sdiff` (*centered*) and its interaction with ambiguity. *Dashed lines* indicate the 95 % confidence interval

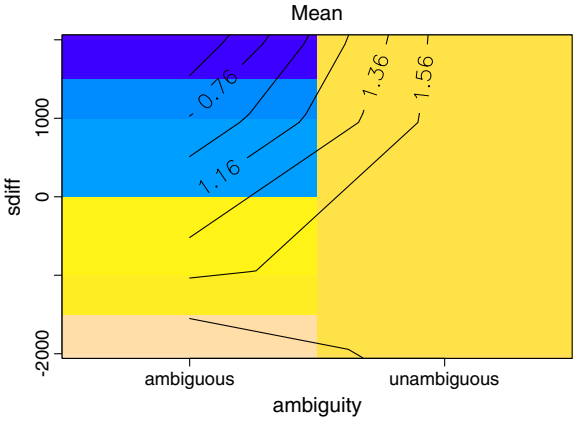


Fig. 10 Mean EEG in the N400 time window as a function of `sdiff` (*centered*) and its interaction with ambiguity. The colors indicate the "height", i.e., a range of (predicted) values of the mean EEG; the value is given by the contour curves. Colors that are closer together (e.g. *light* vs. *dark blue*) indicate finer differences. More color indicates more variation

actor role more directly than the late positivity. In unambiguous sentences, competition is relatively low due to the strong weighting of unambiguous case marking information in German. Hence, no effects on N400 amplitude were observed in these sentence types. The late positivity, by contrast, reflects a behaviorally relevant well-formedness categorization, which can, in part, be envisaged as dependent on how predictable a particular element is within a given sentence context. This result emphasizes the qualitative difference between the N400 and late positivity effects, in spite of their tight interrelationship within the overall biphasic response. Such a difference is further supported by the finding that the amplitude of the late positivity effect showed a substantially stronger correlation with reaction times for the behavioral task than the amplitude of the N400.

Overall, our findings suggest that the neural implementation of actor competition is best modeled by a weighted— rather than an unweighted—measure of the distance between the arguments in terms of actor features. This indicates that actor competition cannot be wholly reduced to similarity-based interference—at least in the sense of similarity-based interference as it is currently assumed in existing memory-based models of sentence processing. Crucially, similarity-based interference is a property of memory

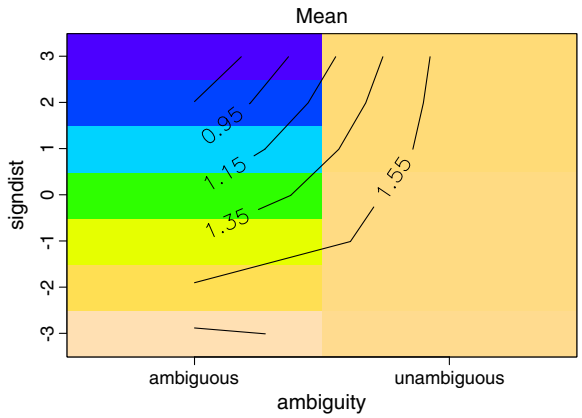


Fig. 9 Mean EEG in the N400 time window as a function of `signdist` (*centered*) and its interaction with ambiguity. The colors indicate the "height", i.e., a range of (predicted) values of the mean EEG; the value is given by the contour curves. Colors that are closer together (e.g. light vs. dark blue) indicate finer differences. More color indicates more variation

Table 15 Summary of model fit for `sdiff` (weighted distinctness) in the N400 window, with NP1 ambiguous

Linear mixed model fit by maximum likelihood			
AIC	BIC	logLik	deviance
197353	197429	−98668	197335
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.21	0.46
	c.(sdiff)	9.3e−08	0.0003
subj	(Intercept)	1.1	1.1
	c.(sdiff)	1.1e−07	0.00033
Residual		17	4.1
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	1.2	0.19	6.2
c.(sdiff)	−0.00039	7e−05	−5.6

Table 16 Summary of model fit for `sdiff` (weighted distinctness) in the N400 window, with NP1 unambiguous

Linear mixed model fit by maximum likelihood			
AIC	BIC	logLik	deviance
197470	197546	−98726	197452
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.21	0.46
	c.(sdiff)	2.5e−08	0.00016
subj	(Intercept)	1.5	1.2
	c.(sdiff)	3.3e−08	0.00018
Residual		17	4.1
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	1.7	0.21	7.9
c.(sdiff)	−5.9e−06	3.8e−05	−0.16

Table 18 Summary of model fit for `dist` (feature overlap) and ambiguity in the P600 window

Linear mixed model fit by maximum likelihood			
AIC	BIC	logLik	deviance
403101	403193	−201541	403081
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.066	0.26
	c.(dist)	0.068	0.26
subj	(Intercept)	1.2	1.1
	c.(dist)	0.15	0.39
Residual		19	4.4
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	0.47	0.18	2.6
ambiguityunambig	−0.093	0.038	−2.4
c.(dist)	−0.51	0.075	−6.9

models assuming a direct access to memory representations via a content-addressable pointer mechanism rather than a memory search: "The defining property of a content-addressable retrieval process is that information (cues) in the retrieval context enables direct access to relevant memory representations, without the need to search through extraneous memory representations" (McElree 2006, p. 163). Thus, since different types of cues serve to specify the "parts" making up the pointer address, they are not weighted—just as in a street address the name of the street, say, is not weighted differently to the house number or the postcode. It is therefore the qualitative overlap between cues that leads to similarity-based interference and weighting of the cues has no obvious role in a memory retrieval mechanism of this type. While, to the best of our knowledge, these characteristics apply to all existing models of language processing drawing on the assumption of direct memory access and similarity-based interference (McElree 2006; Lewis et al. 2006; Lewis and Vasishth 2005; Martin and McElree 2008), we cannot exclude that it may in principle be possible to assume a weighting of retrieval cues. If this were the case, the current findings could potentially also be subsumed under models based on the notion of similarity-based interference in memory retrieval. Nevertheless, we would like to stress once again that this assumption of feature weighting is not incorporated in any current models of this type.

Table 17 Statistics comparing the predictors for a (unweighted) syntactic subject and (weighted) actor-prominence features in the N400 time window

	Df	AIC	logLik
synsigndist.ambiguity: mean ~ ambiguity * c.(synsigndist) + ...	11	395292	−197635.41
sdiff.ambiguity: mean ~ ambiguity * c.(sdiff) + ...	11	395190	−197584.42

(Random effect structure elided. See page 12)

Relation to Previous Work on Computer-Implemented Models

Vosse and Kempen (2008) conducted a similar computer-supported study using experimental stimuli from a previous experiment on scrambling (non canonical word order) effects with different verb classes in German, e.g. sentences where actorhood features do not as clearly align with syntactic subjecthood (Bornkessel and Schlesewsky 2006). The model presented is based primarily on morphosyntactic features—sentence topology (especially important for German data, but also used to model information structural alternations in word order), the related linear word order, and lexical features (including word category) and frames (related to valency in traditional linguistics). Input is processed via "unification" (see also Kempen, this issue), whereby trees are successively assembled and attached to one-another to build a single, unified tree / representation for the sentence. Processing difficulty is represented by cycles required to attach items in the correct configuration, e.g. a single nominative argument in agreement with the verb is readily attached, whereas a non agreeing noun-verb pair requires more effort to attach.

Table 19 Statistics for models in the P600 time window based on `dist` metric, comparing the modelling of ambiguity with and without interaction

	Df	AIC	logLik	Chisq	Chi Df	Pr(>Chisq)
dist.ambiguity.no_int: mean ~ ambiguity + c.(dist) + ...	10	403101	−201540.73			
dist.ambiguity: mean ~ ambiguity * c.(dist) + ...	11	403101	−201539.69	2.09	1	1.48e−01

(Random effect structure elided. See page 12)

In this way, Vosse and Kempen's model is similar to the model presented here: both allow for a particular type of competition for attachment to a representation / role.[22] The models differ however in which features are modeled as well as their ability to model the entire time-course. Crucially, Vosse and Kempen's model primarily models late positivities (although they acknowledge that "certain negativities might find their origin in parser dynamics as well") and fails to predict that subsequent studies have consistently shown a biphasic N400 - late positivity pattern following a reanalysis towards an undergoer-initial order in sentences with accusative verbs rather than only a positivity (Haupt et al. 2008).

However, it is important to point out that the aim of the present study was not to pit an actor-based interpretation strategy against a subject-centered interpretation strategy. Rather, based on the empirical motivation for an actor strategy in our own previous research (see section "Evidence for the Actor Heuristic and for Competition for the Actor Role"), it sought to examine the predictive capacity of various computational metrics designed to implement the actor heuristic. Thus, while the current findings provide an initial indication that a computational model based on an an actor-centered rather than subject-centered interpretation strategy shows a superior fit to electrophysiological findings on human sentence comprehension (see the improvement of `sdiff` over `synsigndiff` in both time windows (Tables 17, and 31), the present experimental design included a considerable degree of overlap between the features relevant to the two strategies. Thus, a direct computational test of an actor-first strategy against a subject-first strategy in situations where the two diverge more strongly remains to be carried out in future research.

Implications for the Interpretation of the N400 and Late Positivity ERP Responses

The present findings have interesting and potentially important implications for the interpretation of language-related ERP responses. In this section, we will therefore relate our results to current approaches to the N400 and late positivity in turn.

With regard to the N400, many researchers have recently come to favor a lexically-based interpretation of this component. According to this perspective, modulations of N400 amplitude do not reflect the computation of message-level meaning, but can rather be reduced to the effort required to retrieve a word from semantic memory (Kutas and Federmeier 2000). Effort is conditioned, in part, by intrinsic properties of the word such as its frequency, but also by its degree of preactivation given the preceding sentence and discourse context (Lau et al. 2008; Brouwer et al. 2012; Stroud and Phillips 2012). These assumptions can explain why N400 amplitude is modulated by single-word predictability (e.g. DeLong et al. 2005) and also why, in English and Dutch, "semantic reversal anomalies" (i.e. sentences such as "The hearty meals were devouring ...", Kim and Osterhout 2005) engender only late positivity effects but not N400 effects.

Table 20 Summary of model fit for `sdiff` (weighted distinctness) and ambiguity in the P600 window

Linear mixed model fit by maximum likelihood			
AIC	BIC	logLik	deviance
403740	403841	−201859	403718
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.06	0.24
	c.(sdiff)	1.5e−08	0.00012
subj	(Intercept)	1.2	1.1
	c.(sdiff)	1.3e−08	0.00011
Residual		19	4.4
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	0.72	0.18	4
ambiguityunambig	−0.6	0.033	−18
c.(sdiff)	0.00053	3.4e−05	15
ambiguityunambig:c.(sdiff)	−0.00059	2.6e−05	−23

[22] Vosse and Kempen (2009) describe their parsing framework as a "dynamic model of syntactic parsing based on activation and inhibitory competition."

Table 21 Statistics for models in the P600 time window based on the `sdiff`, comparing the modelling of ambiguity with and without interaction

	Df	AIC	logLik	Chisq	Chi Df	Pr(>Chisq)
sdiff.ambiguity.no_int: mean ~ ambiguity + c.(sdiff) + ...	10	404244	−202112.37			
sdiff.ambiguity: mean ~ ambiguity * c.(`sdiff`) + ...	11	403740	−201859.03	506.68	1	< 2.2e−16 ***

(Random effect structure elided. See page 12)

Table 22 Summary of model fit for `signdist` (directed, net change in prominent features) and ambiguity in the P600 window

Linear mixed model fit by maximum likelihood			
AIC	BIC	logLik	deviance
403875	403976	−201927	403853
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.059	0.24
	c.(signdist)	0.015	0.12
subj	(Intercept)	1.2	1.1
	c.(signdist)	0.0082	0.091
Residual		19	4.4
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	0.72	0.18	4
ambiguityunambig	−0.6	0.033	−18
c.(signdist)	0.31	0.028	11
ambiguityunambig:c.(signdist)	−0.36	0.02	−18

While this lexical view of the N400 is rather appealing and is able to account for a wide range of findings in the language-related ERP literature, it does not suffice to explain the present findings. Firstly, consider the basic finding of an increased N400 whenever a reanalysis towards an undergoer-initial order was required. This could be explained by the lexical view under the assumption that, following the actor interpretation of the first noun phrase and the subsequently encountered transitive verb, the processing system expects to encounter a second noun phrase marked for (or at least compatible with) accusative (rather than nominative) case. In terms of preactivation, this would entail preactivating accusative case forms—either in terms of full-form lexical entries or of abstract, but nevertheless lexically stored, grammatical information. (But note that this explanation presupposes a rather specific view of lexical organization.) Crucially, however, it is not clear how this explanation might extend to the additional modulation of the actor-reanalysis effect via person / pronominality. The system has no way of predicting whether the second noun phrase will be a first person pronoun or a non-pronominal NP (since there is no expectation to encounter an actor argument at this point, one could not make the argument that a first person argument is more highly expected since it is a more prototypical instantiation of an actor argument). Thus, it is not clear how a purely lexically-based account of the N400 might account for the present findings (for further examples of problematic results for this class of N400 models, see Lotze et al. 2011; Bornkessel-Schlesewsky et al. 2011; Bourguignon et al. 2012). Rather, our data suggest that the N400—as one instance of a broader class of negativity responses—reflects at least certain aspects of integration between the current input and the input previously encountered. While top-down factors such as predictability, which can plausibly be translated into the notion of lexical preactivation, play an important part in determining N400 amplitude, bottom-up properties of the current input item must also be taken into account.

With regard to the late positivity, the close relationship between positivity amplitude and behavioral responses (reaction times) provides converging support for accounts of this component which posit a general (task-related) explanation rather than a specific linguistic function (e.g. reanalysis or effortful combinatorial analysis, Hagoort 2003; Kuperberg 2007). In addition to the account advocated here,

Table 23 Statistics for models in the P600 time window based on the `signdist`, comparing the modelling of ambiguity with and without interaction

	Df	AIC	logLik	Chisq	Chi Df	Pr(>Chisq)
signdist.ambiguity.no_int: mean ~ ambiguity + c.(signdist) + ...	10	404211	−202095.55			
signdist.ambiguity: mean ~ ambiguity * c.(signdist) + ...	11	403875	−201926.73	337.66	1	< 2.2e−16 ***

(Random effect structure elided. See page 12)

Table 24 Statistics for the minimally adequate models for each predictor in the P600 time window: the models do not differ by much

	Df	AIC	logLik
dist.ambiguity.no_int: mean ~ ambiguity + c.(dist) + ...	10	403101	−201540.73
sdiff.ambiguity: mean ~ ambiguity * c.(sdiff) + ...	11	403740	−201859.03
signdist.ambiguity: mean ~ ambiguity * c.(signdist) + ...	11	403875	−201926.73

(Random effect structure elided. See page 12)

such a view has been proposed most prominently from the perspective of the conflict monitoring hypothesis (e.g. Kolk et al. 2003; van de Meerendonk et al. 2009). According to this proposal, late positivity effects in language processing reflect the detection of conflicting information and an ensuing check of the input for errors in previous processing steps. Evidence for this perspective stems, for example, from the finding that late positivities can be observed in response to various types of conflicts including orthographic errors (Vissers et al. 2006) and that, while both weak and strong semantic conflicts induce N400 effects, only strong conflicts engender an additional late positivity (van de Meerendonk et al. 2010). The conflict monitoring hypothesis can therefore also account for the observation that a reanalysis of the actor-first preference engenders late positivity effects (in addition to N400 modulations): here, conflict is high in comparison to sentences with an actor-initial word order. More precisely, in contrast to the N400, which is observable for all visually presented words (cf. the description in the very first study Kutas and Hillyard 1980), but can be described as an effect in certain contexts, the late positivity belongs more to the class of relative effects, occurring primarily in contrast to a condition with less (resolvable) conflict in experiments with a conflict-focused task (e.g. acceptability judgments) (Sassenhagen et al. 2013; Frenzel et al. 2011; Hahne and Friederici 2002).

However, a crucial difference between the conflict monitoring account and the present approach is that, according to the conflict monitoring view, the late positivity

Table 25 Statistics for models in the P600 time window with NP1 ambiguous

	Df	AIC	logLik
dist: mean ~ c.(dist) + ...	9	202377	−101179.75
sdiff: mean ~ c.(sdiff) + ...	9	202102	−101042.26
signdist: mean ~ c.(signdist) + ...	9	202396	−101189.33

(Random effect structure elided. See page 12)

Table 26 Statistics for models in the P600 time window with NP1 unambiguous

	Df	AIC	logLik
dist: mean ~ c.(dist) + ...	5	200678	−100334.17
sdiff: mean ~ c.(sdiff) + ...	5	200976	−100483.28
signdist: mean ~ c.(signdist) + ...	5	200981	−100485.95

(Random effect structure elided. See page 12)

Table 27 Summary of model fit for `sdiff` (weighted distinctness) in the P600 window, with NP1 ambiguous

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance
202103	202179	−101042	202085
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.1	0.32
	c.(sdiff)	1.1e−07	0.00034
subj	(Intercept)	1.1	1
	c.(sdiff)	2.2e−07	0.00047
Residual		19	4.4
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	0.72	0.18	4
c.(sdiff)	0.00053	9.2e−05	5.7

Table 28 Summary of model fit for `sdiff` (weighted distinctness) in the P600 window, with NP1 unambiguous

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance
200977	201019	−100483	200967
Random effects:			
Groups	Name	Variance	Std.Dev.
item	(Intercept)	0.1	0.32
subj	(Intercept)	1.4	1.2
Residual		19	4.3
Fixed effects:			
	Estimate	Std. Error	t value
(Intercept)	0.12	0.2	0.58
c.(sdiff)	−6.4e−05	1.2e−05	−5.6

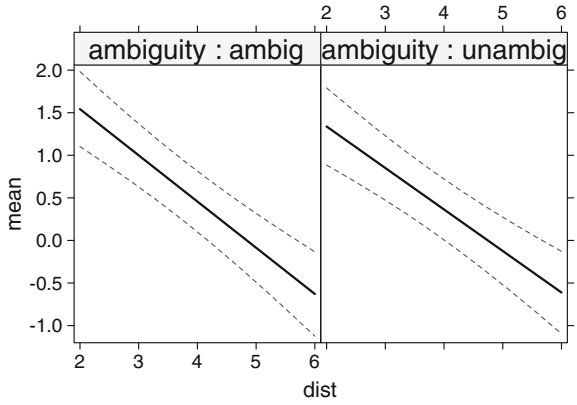


Fig. 11 Mean EEG in the P600 time window as a function of `dist` (*centered*) and its interaction with ambiguity. *Dashed lines* indicate the 95 % confidence interval

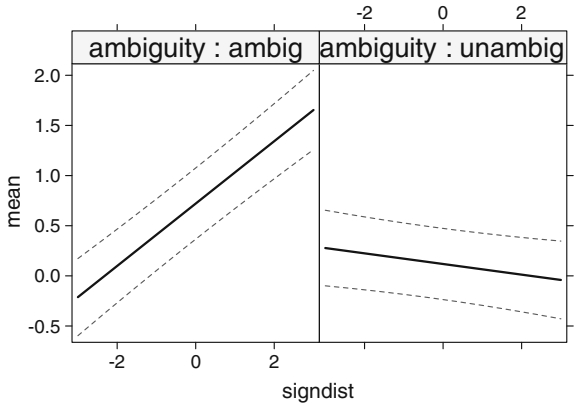


Fig. 12 Mean EEG in the P600 time window as a function of `signdist` (*centered*) and its interaction with ambiguity. *Dashed lines* indicate the 95 % confidence interval

reflects a (domain-general) reanalysis of the input (conflict resolution) (van de Meerendonk et al. 2009) while we posit that conflict resolution is reflected in the N400. Evidence for the N400 as the locus of conflict resolution stems from the observation that recent studies examining reanalyses of the actor-first preference have consistently found N400 effects, with additional late positivities depending on the behavioral relevance of the object-initial order (Bornkessel et al. 2004b; Haupt et al. 2008). Specifically, when the object-initial order was licensed by the presence of an object-experiencer verb and therefore did not call for a behavioral reorientation (i.e. judgement of the sentence as unacceptable), only an N400 effect was observed but no late positivity. The present results provide converging support for this perspective, since conflict resolution in the sense of a reanalysis should be more closely tied to the conflict-inducing feature in the input rather than to the behavioral response. Thus, the observation that the amplitude of the late positivity correlated considerably more strongly with the reaction times for the judgment task is expected under the assumption that the N400 reflects (input-related) conflict resolution, while the late positivity reflects the behavioral consequences of the conflict (and its resolution) in the given task environment.[23]

[23]Note that an explanation along these lines can also account for the dissociation between mild and strong conflicts observed by van de Meerendonk et al. (2010). As it appears plausible to assume that only the strong conflicts were registered as behaviorally significant, our account derives the finding of a late positivity for these conflicts, while no such effect was observed for mild conflicts. This explanation leads to the testable prediction that, with different task instructions (e.g. a judgment task emphasizing that even mild implausibilities should be classified as such), van de Meerendonk et al. (2010)'s mild conflict stimuli should also engender a late positivity.

Towards a Neurobiologically Realistic Computational Model of Actor computation

The present results demonstrate that the `sdiff` measure is a promising candidate for a neurocognitively plausible formalization of actor competition, as it is a valid predictor of neurophysiological activity related to sentence comprehension. In addition, we propose that this metric can be viewed as a first step towards a computational formalization of the neurobiological model described in section "The Extended Argument Dependency Model (eADM) and Actor-Centered Comprehension". Specifically, we suggest that the insights gleaned from the present work may further our understanding of how linguistic categories posited within the eADM—such as the actor role—are recognized and processed in a neurobiologically plausible manner and how we might

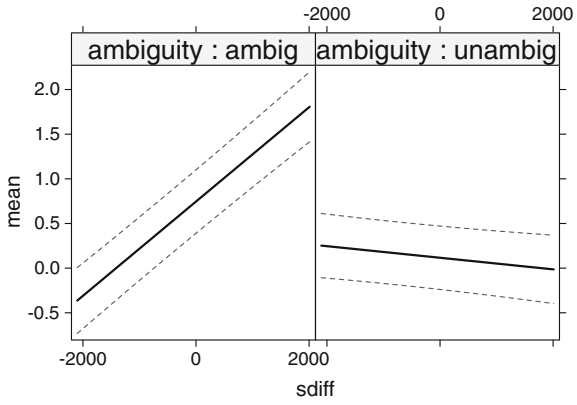


Fig. 13 Mean EEG in the P600 time window as a function of `sdiff` (*centered*) and its interaction with ambiguity. *Dashed lines* indicate the 95 % confidence interval

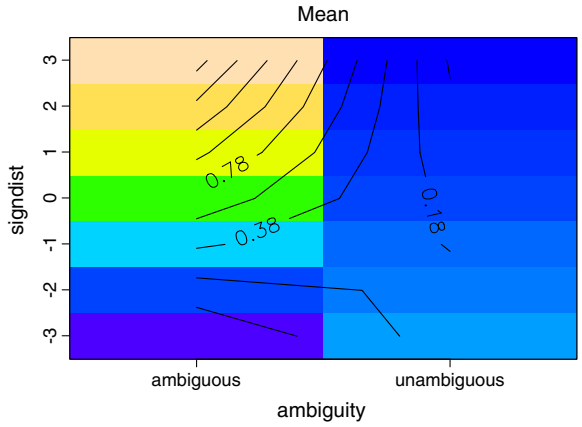


Fig. 14 Mean EEG in the P600 time window as a function of `signdist` (*centered*) and its interaction with ambiguity. The colors indicate the "height", i.e., a range of (predicted) values of the mean EEG; the value is given by the contour curves. Colors that are closer together (e.g. *light vs. dark blue*) indicate finer differences. More color indicates more variation

envisage the relation between linguistic and non-linguistic categories.

Our proposal builds on the suggestion that, in view of its cross-linguistic ubiquity, actor could be modeled as an attractor category (Bornkessel-Schlesewsky and Schlesewsky 2013a). Recent work in computational neuroscience has shown that attractor networks provide a neurobiologically plausible means of modeling decision-making processes (Deco et al. 2009; Deco et al. 2012), both for complex value-based choices as well as for perceptual classifications ("perceptual decision-making"; Heekeren et al. 2004; Basten et al. 2010). In an attractor network, decisions can be modeled via attractor states in a neural network which are associated with (stable) high firing rates. Which state "wins" during decision making is determined by the current input and the initial stochastic firing behaviour of the network. Based on the overarching (language-independent) importance of the actor category (see section "Evidence for the Actor Heuristic and for Competition for the Actor Role"), it has been proposed that an attractor network for actor categorization exists independently of language (Bornkessel-Schlesewsky and Schlesewsky 2013a). This network is universal, as it reflects the general human ability to recognize goal-directed action. The stable firing patterns inherent to this network are plausibly based on sets of input features that co-occur in domain-general actor recognition. As the linguistic actor category overlaps to a certain degree with these general features (e.g. via the features +HUMAN, +ANIMATE and +1ST. PERSON), there is a propensity for actor recognition via the general attractor network. With regard to more language-specific features (e.g. case marking), the system will learn that these correlate with the remaining (domain-general) actor features such that, in the mature system, they also push the network towards the actor recognition attractor state.[24] Crucially, an important consideration pertains to the degree of evidence for a certain decision—e.g. the classification of a certain event participant as an actor—that the current input offers. The `sdiff` metric can be viewed as a measure that captures this degree of evidence.

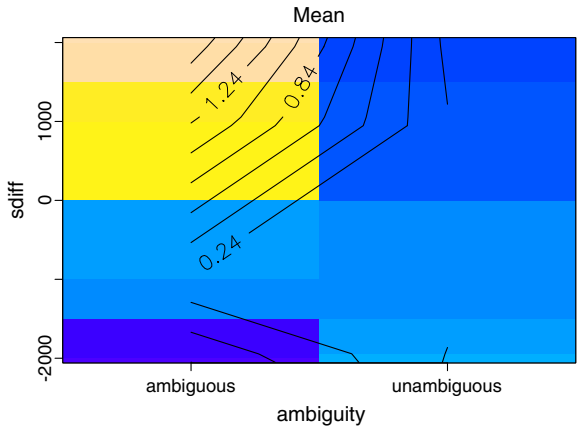


Fig. 15 Mean EEG in the P600 time window as a function of `sdiff` (*centered*) and its interaction with ambiguity. The colors indicate the "height", i.e., a range of (predicted) values of the mean EEG; the value is given by the contour curves. Colors that are closer together (e.g. *light vs. dark blue*) indicate finer differences. More color indicates more variation

Weighted, directed measures, such as `sdiff`, provide the means to quantify the effects of attractor basins. Indeed, the physical metaphor behind attractor basins also provides insight into why `sdiff` functions better than `dist` or `signdist` (Fig. 16). The proximity of an attractor is given by `dist`, but not whether it is a positive or negative attractor (true attractor vs. repulsor, or hill vs. basin; see Fig. 17). This is a decent first approximation, but quickly fails in more rolling landscapes, e.g. in languages with free word order, where it is not clear which argument will come first. The directionality of `signdist` provides a better approximation, modeling attraction and repulsion, but the best approximation comes from the strength of the attractor (the steepness of the sides of the basin / hill or equivalently, the height and depth; see Fig. 18). This is exactly what `sdiff` does—the weightedness distorts the actor space, creating stronger and weaker attractors (Fig. 19). In this sense, deterministic case marking and garden path sentences are examples of attractor basins that are too deep to escape,

[24] This proposal of a tight interrelationship between domain-general and linguistic actor features is supported by the recent observation that properties of an ideal actor may depend—at least to some degree—on the characteristics of one's native language (Fausey et al. 2010; Fausey and Boroditsky 2011).

Table 29 Statistics for the minimally adequate (stimulus-based) models in the N400 time window compared to their extension via reaction time (RT)

	Df	AIC	logLik	Chisq	Chi Df	Pr(>Chisq)
sdiff.ambiguity: mean ~ ambiguity * c.(sdiff) + ...	11	395190	−197584.42			
sdiff.ambiguity.rt: mean ~ ambiguity * c.(sdiff) + log(rt) + ...	12	395172	−197574.02	20.80	1	<5.10e−06 ***

(Random effect structure elided. See page 12)

where the language system becomes trapped at the bottom of a well, or perhaps, to use another meaning of the word "space", in a black hole.

Finally, though we have focused on the actor role here, we propose that the notion of attractor basins could be used to formalize the entire processing architecture shown in Fig. 1. Specifically, attractors could be used to model the categories assumed at every processing step within the cascade (e.g. phonemes, actor-event schemata etc.). They could further help to address an issue that is conspicuously missing from the current model implementation, namely the need to provide an estimate of the timing of the different processing steps and, accordingly, of the neurophysiological responses elicited by them. At present, the model only specifies the relative order of information processing but offers no quantifiable timing estimates. However, combining the assumption of cascaded, hierarchically organized processing steps and the attractor notion opens up a possible avenue for such a quantification. As noted in section "The Extended Argument Dependency Model (eADM) and Actor-Centered Comprehension", cascaded processing is based on the idea that, once a sufficient degree of information has accrued, processing can proceed to the next step. Drawing upon the attractor notion, we can posit that the faster the system recognizes that information is relevant for a particular attractor, the faster processing at the step relevant to that attractor will be. Accordingly, the formalization of actor space presented here could be used as the basis for estimating processing latency as well as amplitude and, in our view, should also carry over to other linguistic categories. Of course, timing estimates will not be trivial given the different levels of neuronal responses that need to be considered here: as mentioned in section "The Extended Argument Dependency Model (eADM) and Actor-Centered Comprehension", scalp ERPs as examined here are macroscopic responses with (typically) multiple underlying sources and therefore cannot be directly compared to the cascade of activity that is assumed to proceed through individual regions along the antero-ventral and postero-dorsal streams. Accordingly, latencies of language-related ERP components such as the N400 likely do not reflect the absolute timing of information processing (see Bornkessel-Schlesewsky and Schlesewsky 2013b). Nevertheless, assuming that our proposal regarding the basic relationship between evidence for an attractor and duration of the processing step in question is correct, both the direct neuronal responses and the neurophysiological responses measured by means of scalp EEG recordings should be quantifiable as some function of the degree of evidence for the respective attractor category.

Future Directions and the Role of Neuroinformatics

In the experiment presented here, `sdiff` showed the advantage of a weighted, directed distinctness measure over simple (unweighted) interference measures. Nonetheless, morphological case and ambiguities involving the same dominated the most important prominence variations. In future work, we aspire to test the metric against a wider range of stimuli, including globally ambiguous sentences and generally more naturalistic language. On account of the modular nature of the implementation, any EEG dataset could be processed and analyzed, either via adapting / constructing a suitable front-end (Stage 1) parser or via manually tagging the stimuli appropriately for Stage 2. As more work is done in this direction of quantifying linguistic differences in the brain, it becomes increasingly important to have diverse test data, especially if learning is to be

Table 30 Statistics for the minimally adequate (stimulus-based) models in the P600 time window compared to their extension via reaction time (RT)

	Df	AIC	logLik	Chisq	Chi Df	Pr(>Chisq)
sdiff.ambiguity: mean ~ ambiguity * c.(sdiff) + ...	11	403740	−201859.03			
sdiff.ambiguity.rt: mean ~ ambiguity * c.(sdiff) + log(rt) + ...	12	403505	−201740.85	236.36	1	<2.2e−16 ***

(Random effect structure elided. See page 12)

Table 31 Statistics comparing the predictors for a (unweighted) syntactic subject and (weighted) actor-prominence features P600 time window

	Df	AIC	logLik
synsigndist.ambiguity: mean ~ ambiguity * c.(synsigndist) + ...	11	404065	−202021.65
sdiff.ambiguity: mean ~ ambiguity * c.(sdiff) + ...	11	403740	−201859.03

(Random effect structure elided. See page 12)

simulated at some point. To this end, it is crucially important that databases of EEG data for diverse stimuli from typologically varied languages are available, similar to the corpora and treebanks used by researchers in natural language processing. A general model of language comprehension is the goal, not a model for particular dataset.

Call for Data (Banks)

To this end, we would like to see databanks of neuroanatomical and neurophysiological data similar to the "treebanks" common in computational linguistics and natural language processing research. Such databanks should provide a validated, state-of-the-art analysis with traditional methodologies, e.g. ANOVA (including standardized ROIs and/or single electrodes as a factor, with grand-average ERP) as well as parametric labeling of relevant linguistic information—cloze probability of each word, morphosyntactic features, thematic relations, known important semantic features (e.g. animacy and ideally other features that are expressed morphosyntactically in any of the world's languages), lexical

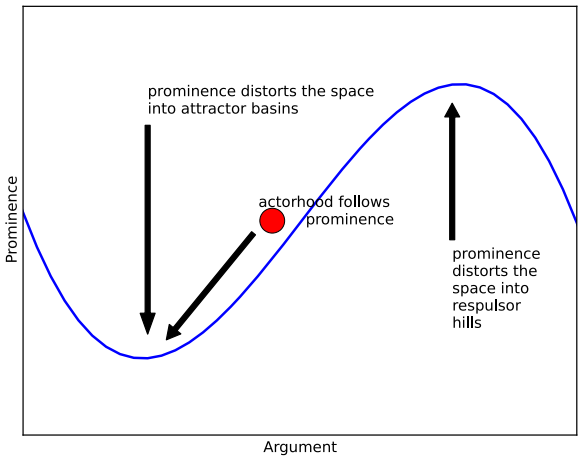


Fig. 16 Attractor basins in actor space. Prominence can be viewed as a distortion of actor space. The curvature of actor space then pulls or pushes actorhood towards a particular argument

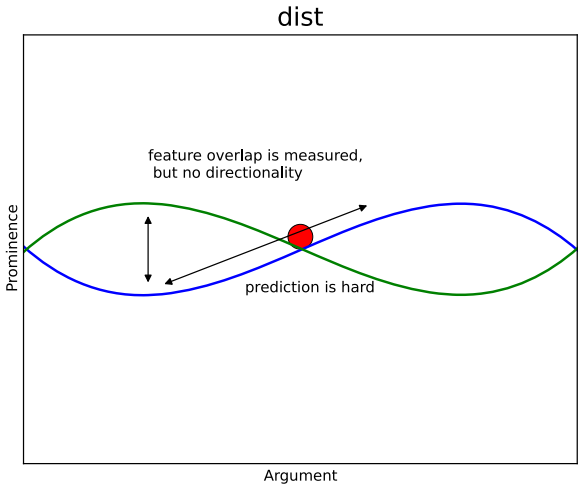


Fig. 17 Attractor basins in actor space as measured by `dist`. The directionality of distortion is lost, making prediction difficult

frequency estimates, estimates of syntactic frequency for any syntactic peculiarities (e.g. non canonical word orders), etc. The EEG data should preferably be stored in an open format, or at least in a format for which there are suitable plugins and converters—perhaps one of the formats supported by the open source EEGLAB software package. Data should not be filtered, rereferenced or otherwise manipulated offline before storage so as to not limit analysis by alternative techniques (time-frequency analysis, ICA, etc.). Instead, the measurement parameters (sampling rate, reference electrode, equipment manufacturer), experimental setup (presentation mode and aspects pertaining thereto) and anonymized subject data (age, sex, etc.) as well as

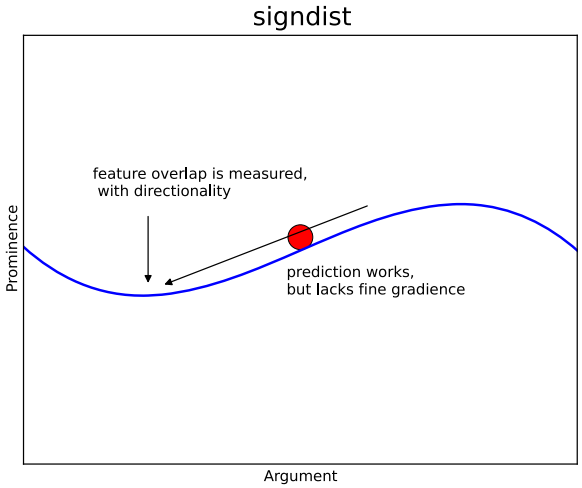


Fig. 18 Attractor basins in actor space as measured by `signdist`. The gradience of distortion is lost, leading to only qualitative predictions

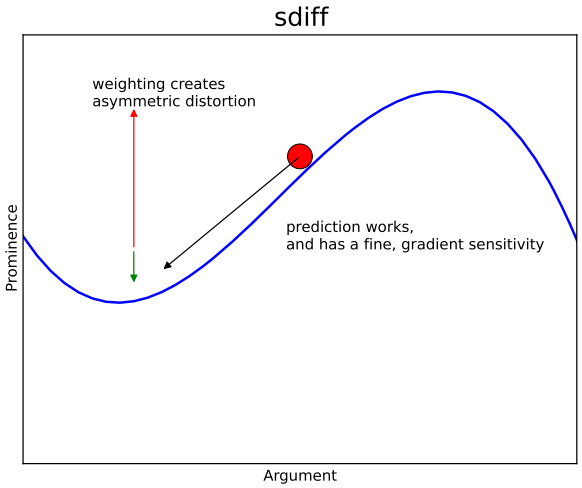


Fig. 19 Attractor basins in actor space as measured by `sdiff`. Both directionality and gradience of the distortion is preserved, thus enabling accurate prediction. The distortion here is to scale for the sentence *Die Bettlerin bedrängte den Kommissar auf der Straße*. (See Table 3)

experimental task, handedness of the test subject and task interface layout, should be stored as metadata. Optionally, the number of channels could be stored as well; however, this is not necessary. Channel names should be standardized to the 10-10/10-20 system terminology. Behavioral data should be linked not just as metadata, but also as a proper dataset unto itself. (Relational databases provide a convenient way to do this.) Only then, can we truly test our models of sentence comprehension, i.e. our "parsers", with the same rigor that has been standard in other computational disciplines for years now—with lots of large, standardized tests.

Vosse and Kempen (2008) took an important first step in this direction, analyzing data from another experiment; however, it appears important to go beyond comparisons of modeling output with the published, summarized data. As discussed by Arbib et al. (this issue), it is important to remember that summary data implies the existence of non summarized data, i.e. more complete data. As one researcher's noise is another's signal, even the most basic filtering removes important data; the usual presentation of means and ANOVA leads even more to be desired. The BrainMap database is an excellent start for fMRI data, but it only makes the want of a comparable database for EEG data more striking. Recent trends in Open Access and preregistration point to such databases as being the way of the future. We need data sharing beyond rebuttal and as common proving ground beyond the current experiments hand crafted to show off a particular model feature.

Beyond the traditional, well parameterized experimental data, we would ideally also like to see a complementary set of data acquired in a less structured, free-task environment. That is, we would like to see a similar dataset of EEG/fMRI recording of natural stories with a maximal task of a few comprehension questions taken offline after the main experiment (Whitney et al. 2009) but with fully tagged input. Computational linguists use more than sets of simple, constructed sentences to test their data and so should we—our models need to be able to handle the full complexity of human language in its actual use and not just in our idealized laboratory conditions. These more complex inputs also present us the chance to move beyond sentence-processing models towards language-processing models.

The existence of large, standardized datasets also provides for a proving ground for newer methodologies. For example, although time-frequency analysis, principal component analysis (PCA) and independent component analysis (ICA) have been used in recent years to differentiate certain subtleties not readily apparent from traditional ERP-based analyses, the world of EEG-data is still dominated by ERP. This is almost certainly related to not just the complexity of these new methods, but also their unclear relationship to ERP results. A standard dataset provides exactly the playground necessary to demonstrate and test new methodologies and their relationship to old ones.

Brief Technical Notes on Implementation

The present implementation is in Python 3.2. A previous version was written and tested in Python 2.7; however, the implementation of Stage 1 and necessity of using non ASCII encoding for German sentence data motivated the shift to the 3.x series of Python with its much more extensive Unicode support. File and directory manipulations were all tested on POSIX compatible platforms.

There are options to set the baseline weights all equal to one (no weighting) or to a priori estimated weights based on previous work done in German (Kempe and MacWhinney 1999). A further correction (from empirical data) can then be applied to the individual baseline weights via additional options.

To test the weight configuration, a set of potential constructions in German is provided in a form directly processable by Stage 2. A test mode operating purely on these preanalyzed inputs is one of three modes of operation.

The other modes are a batch mode for generating predictions about experimental stimuli and an interactive mode for demonstrations of the model, as well as a mode capable of processing Stage 1 output vectors, either as list of experimental conditions from a file or interactively. Both the batch and interactive modes use a limited version of Stage 1, featuring a small parser customized for the experiment in question (see Stage 1).

More detailed documentation can be found with the publicly available source code (see below).

Information Sharing Statement

All program source code for the implementation and generation of figures related to the mixed-models as well as the necessary input files is available to the public at https://bitbucket.org/palday/ginnungagap-code/.

Acknowledgments We would like to thank Rick Lewis and Joakim Nivre for valuable discussions and suggestions related to the development of the computational model. We would also like to thank Isabel Plauth for the data acquisition.

Appendix

Table 32 ANOVA for the N400 window

ANOVA:

	Effect	DFn	DFd	F	p	p<.05	ges
2	roi	4.00	144.00	14.47	0.00	*	0.03
3	wordOrder	1.00	36.00	19.60	0.00	*	0.01
4	ambiguity	1.00	36.00	19.11	0.00	*	0.01
5	np1type	1.00	36.00	5.61	0.02	*	0.00
6	np2type	1.00	36.00	54.44	0.00	*	0.14
7	roi:wordOrder	4.00	144.00	5.51	0.00	*	0.00
8	roi:ambiguity	4.00	144.00	7.51	0.00	*	0.00
9	wordOrder:ambiguity	1.00	36.00	21.84	0.00	*	0.01
10	roi:np1type	4.00	144.00	13.25	0.00	*	0.00
11	wordOrder:np1type	1.00	36.00	0.57	0.45		0.00
12	ambiguity:np1type	1.00	36.00	0.75	0.39		0.00
13	roi:np2type	4.00	144.00	55.78	0.00	*	0.06
14	wordOrder:np2type	1.00	36.00	0.35	0.56		0.00
15	ambiguity:np2type	1.00	36.00	0.17	0.68		0.00
16	np1type:np2type	1.00	36.00	0.17	0.68		0.00
17	roi:wordOrder:ambiguity	4.00	144.00	2.58	0.04	*	0.00
18	roi:wordOrder:np1type	4.00	144.00	1.04	0.39		0.00
19	roi:ambiguity:np1type	4.00	144.00	0.53	0.72		0.00
20	wordOrder:ambiguity:np1type	1.00	36.00	1.30	0.26		0.00
21	roi:wordOrder:np2type	4.00	144.00	12.10	0.00	*	0.00
22	roi:ambiguity:np2type	4.00	144.00	4.53	0.00	*	0.00
23	wordOrder:ambiguity:np2type	1.00	36.00	6.40	0.02	*	0.00
24	roi:np1type:np2type	4.00	144.00	1.21	0.31		0.00
25	wordOrder:np1type:np2type	1.00	36.00	1.99	0.17		0.00
26	ambiguity:np1type:np2type	1.00	36.00	0.23	0.63		0.00
27	roi:wordOrder:ambiguity:np1type	4.00	144.00	0.61	0.66		0.00
28	roi:wordOrder:ambiguity:np2type	4.00	144.00	2.68	0.03	*	0.00
29	roi:wordOrder:np1type:np2type	4.00	144.00	0.35	0.84		0.00
30	roi:ambiguity:np1type:np2type	4.00	144.00	4.03	0.00	*	0.00
31	wordOrder:ambiguity:np1type:np2type	1.00	36.00	2.88	0.10		0.00
32	roi:wordOrder:ambiguity:np1type:np2type	4.00	144.00	0.91	0.46		0.00

Table 32 (continued)

Sphericity Corrections:				
	Effect	W	p	p<.05
2	roi	0.37	0.00	*
7	roi:wordOrder	0.20	0.00	*
8	roi:ambiguity	0.28	0.00	*
10	roi:np1type	0.22	0.00	*
13	roi:np2type	0.33	0.00	*
17	roi:wordOrder:ambiguity	0.22	0.00	*
18	roi:wordOrder:np1type	0.32	0.00	*
19	roi:ambiguity:np1type	0.07	0.00	*
21	roi:wordOrder:np2type	0.19	0.00	*
22	roi:ambiguity:np2type	0.23	0.00	*
24	roi:np1type:np2type	0.12	0.00	*
27	roi:wordOrder:ambiguity:np1type	0.22	0.00	*
28	roi:wordOrder:ambiguity:np2type	0.06	0.00	*
29	roi:wordOrder:np1type:np2type	0.14	0.00	*
30	roi:ambiguity:np1type:np2type	0.32	0.00	*
32	roi:wordOrder:ambiguity:np1type:np2type	0.12	0.00	*

Table 33 ANOVA for the N400 time window resolved in the Left-Posterior Region of Interest

ANOVA:							
	Effect	DFn	DFd	F	p	p<.05	ges
2	roi	4.00	144.00	24.10	0.00	*	0.06
3	wordOrder	1.00	36.00	25.71	0.00	*	0.02
4	ambiguity	1.00	36.00	31.79	0.00	*	0.04
5	np1type	1.00	36.00	0.16	0.69		0.00
6	np2type	1.00	36.00	93.33	0.00	*	0.19
7	roi:wordOrder	4.00	144.00	7.46	0.00	*	0.00
8	roi:ambiguity	4.00	144.00	5.01	0.00	*	0.00
9	wordOrder:ambiguity	1.00	36.00	37.31	0.00	*	0.04
10	roi:np1type	4.00	144.00	0.27	0.89		0.00
11	wordOrder:np1type	1.00	36.00	0.11	0.74		0.00
12	ambiguity:np1type	1.00	36.00	2.62	0.11		0.00
13	roi:np2type	4.00	144.00	3.44	0.01	*	0.00
14	wordOrder:np2type	1.00	36.00	0.87	0.36		0.00
15	ambiguity:np2type	1.00	36.00	1.98	0.17		0.00
16	np1type:np2type	1.00	36.00	1.42	0.24		0.00
17	roi:wordOrder:ambiguity	4.00	144.00	9.99	0.00	*	0.00
18	roi:wordOrder:np1type	4.00	144.00	0.74	0.56		0.00
19	roi:ambiguity:np1type	4.00	144.00	2.76	0.03	*	0.00
20	wordOrder:ambiguity:np1type	1.00	36.00	0.23	0.63		0.00
21	roi:wordOrder:np2type	4.00	144.00	0.72	0.58		0.00
22	roi:ambiguity:np2type	4.00	144.00	1.82	0.13		0.00
23	wordOrder:ambiguity:np2type	1.00	36.00	0.50	0.48		0.00
24	roi:np1type:np2type	4.00	144.00	0.20	0.94		0.00

Table 33 (continued)

	Effect	DFn	DFd	F	p	p<.05	ges
25	wordOrder:np1type:np2type	1.00	36.00	5.24	0.03	*	0.00
26	ambiguity:np1type:np2type	1.00	36.00	0.86	0.36		0.00
27	roi:wordOrder:ambiguity:np1type	4.00	144.00	1.71	0.15		0.00
28	roi:wordOrder:ambiguity:np2type	4.00	144.00	1.85	0.12		0.00
29	roi:wordOrder:np1type:np2type	4.00	144.00	0.53	0.72		0.00
30	roi:ambiguity:np1type:np2type	4.00	144.00	0.07	0.99		0.00
31	wordOrder:ambiguity:np1type:np2type	1.00	36.00	0.89	0.35		0.00
32	roi:wordOrder:ambiguity:np1type:np2type	4.00	144.00	1.89	0.11		0.00

Sphericity Corrections:

	Effect	W	p	p<.05
2	roi	0.48	0.00	*
7	roi:wordOrder	0.20	0.00	*
8	roi:ambiguity	0.24	0.00	*
10	roi:np1type	0.15	0.00	*
13	roi:np2type	0.53	0.01	*
17	roi:wordOrder:ambiguity	0.34	0.00	*
18	roi:wordOrder:np1type	0.23	0.00	*
19	roi:ambiguity:np1type	0.13	0.00	*
21	roi:wordOrder:np2type	0.17	0.00	*
22	roi:ambiguity:np2type	0.32	0.00	*
24	roi:np1type:np2type	0.20	0.00	*
27	roi:wordOrder:ambiguity:np1type	0.21	0.00	*
28	roi:wordOrder:ambiguity:np2type	0.41	0.00	*
29	roi:wordOrder:np1type:np2type	0.12	0.00	*
30	roi:ambiguity:np1type:np2type	0.15	0.00	*
32	roi:wordOrder:ambiguity:np1type:np2type	0.14	0.00	*

Table 34 ANOVA for the P600 window

ANOVA:

	Effect	DFn	DFd	F	p	p<.05	ges
2	wordOrder	1.00	36.00	23.01	0.00	*	0.02
3	ambiguity	1.00	36.00	30.55	0.00	*	0.03
4	np1type	1.00	36.00	0.01	0.93		0.00
5	np2type	1.00	36.00	9.65	0.00	*	0.04
6	wordOrder:ambiguity	1.00	36.00	20.21	0.00	*	0.02
7	wordOrder:np1type	1.00	36.00	1.96	0.17		0.00
8	ambiguity:np1type	1.00	36.00	0.23	0.64		0.00
9	wordOrder:np2type	1.00	36.00	4.41	0.04	*	0.00
10	ambiguity:np2type	1.00	36.00	0.01	0.93		0.00
11	np1type:np2type	1.00	36.00	1.31	0.26		0.00
12	wordOrder:ambiguity:np1type	1.00	36.00	0.50	0.48		0.00
13	wordOrder:ambiguity:np2type	1.00	36.00	5.17	0.03	*	0.00
14	wordOrder:np1type:np2type	1.00	36.00	2.01	0.17		0.00
15	ambiguity:np1type:np2type	1.00	36.00	0.03	0.86		0.00
16	wordOrder:ambiguity:np1type:np2type	1.00	36.00	3.00	0.09		0.00

Table 35 ANOVA for the P600 time window resolved in the Left-Posterior Region of Interest

ANOVA:							
	Effect	DFn	DFd	F	p	p<.05	ges
2	wordOrder	1.00	36.00	28.20	0.00	*	0.02
3	ambiguity	1.00	36.00	26.00	0.00	*	0.04
4	np1type	1.00	36.00	0.00	1.00		0.00
5	np2type	1.00	36.00	85.25	0.00	*	0.23
6	wordOrder:ambiguity	1.00	36.00	43.74	0.00	*	0.05
7	wordOrder:np1type	1.00	36.00	0.56	0.46		0.00
8	ambiguity:np1type	1.00	36.00	1.33	0.26		0.00
9	wordOrder:np2type	1.00	36.00	2.09	0.16		0.00
10	ambiguity:np2type	1.00	36.00	2.88	0.10		0.00
11	np1type:np2type	1.00	36.00	1.30	0.26		0.00
12	wordOrder:ambiguity:np1type	1.00	36.00	0.50	0.48		0.00
13	wordOrder:ambiguity:np2type	1.00	36.00	0.15	0.70		0.00
14	wordOrder:np1type:np2type	1.00	36.00	7.45	0.01	*	0.00
15	ambiguity:np1type:np2type	1.00	36.00	0.85	0.36		0.00
16	wordOrder:ambiguity:np1type:np2type	1.00	36.00	0.00	0.95		0.00

Table 36 Summmary statistics for the accuracy in trials

Summary statistics:							
wordOrder	ambiguity	np1type	np2type	N	Mean	SD	FLSD
O	A	N	N	37.00	0.86	0.13	0.01
O	A	N	P	37.00	0.93	0.08	0.01
O	A	P	N	37.00	0.89	0.11	0.01
O	A	P	P	37.00	0.94	0.06	0.01
O	U	N	N	37.00	0.95	0.06	0.01
O	U	N	P	37.00	0.94	0.04	0.01
O	U	P	N	37.00	0.98	0.02	0.01
O	U	P	P	37.00	0.96	0.02	0.01
S	A	N	N	37.00	0.97	0.04	0.01
S	A	N	P	37.00	0.97	0.04	0.01
S	A	P	N	37.00	0.98	0.02	0.01
S	A	P	P	37.00	0.98	0.04	0.01
S	U	N	N	37.00	0.97	0.04	0.01
S	U	N	P	37.00	0.97	0.04	0.01
S	U	P	N	37.00	0.98	0.03	0.01
S	U	P	P	37.00	0.99	0.02	0.01

Table 37 Summmary statistics for reaction time in trials

Summary statistics:							
wordOrder	ambiguity	np1type	np2type	N	Mean	SD	FLSD
O	A	N	N	37.00	484.45	197.41	16.07
O	A	N	P	37.00	435.32	164.68	16.07
O	A	P	N	37.00	482.62	184.68	16.07
O	A	P	P	37.00	433.66	157.27	16.07
O	U	N	N	37.00	465.44	167.91	16.07
O	U	N	P	37.00	437.66	155.49	16.07
O	U	P	N	37.00	442.40	161.16	16.07
O	U	P	P	37.00	424.94	142.27	16.07
S	A	N	N	37.00	436.50	147.15	16.07
S	A	N	P	37.00	427.92	146.34	16.07
S	A	P	N	37.00	442.69	149.04	16.07
S	A	P	P	37.00	427.23	151.57	16.07
S	U	N	N	37.00	438.67	151.13	16.07
S	U	N	P	37.00	433.64	149.12	16.07
S	U	P	N	37.00	437.05	151.31	16.07
S	U	P	P	37.00	420.69	127.54	16.07

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi:10.1109/TAC.1974.1100705.

Arbib, M., Bonaiuto, J., Bornkessel-Schlesewsky, I., Kemmerer, D., MacWhinney, B., Nielsen, F., Oztop, E. (this issue). Action and language mechanisms in the brain: data, models and neuroinformatics. *Neuroinformatics*.

Baayen, R., Davidson, D., Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Basten, U., Biele, G., Heekeren, H., Fiebach, C. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*, *107*(50), 21,767–21,772.

Bates, D., Maechler, M., Bolker, B. (2013). lme4: linear mixed-effects models using S4 classes. http://CRAN.R-project.org/package=lme4rpackageversion0.999999-2.

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., Smith, S. (1982). Functional constraints on sentence processing: a cross-linguistic study. *Cognition*, *11*, 245–299.

Bates, E., Devescovi, A., Wulfeck, B. (2001). Psycholinguistics: a cross-language perspective. *Annual Review of Psychology*, *52*, 369–396.

Bornkessel, I. (2002). *The argument dependency model: a neurocognitive approach to incremental interpretation* (Vol. 28). Leipzig: MPI Series in Cognitive Neuroscience.

Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. *Psychological Review*, *113*(4), 787–821.

Bornkessel, I., Fiebach, C.J., Friederici, A.D. (2004a). On the cost of syntactic ambiguity in human language comprehension: an individual differences approach. *Cognitive Brain Research*, *21*, 11–21.

Bornkessel, I., McElree, B., Schlesewsky, M., Friederici A.D. (2004b). Multi-dimensional contributions to garden path strength: dissociating phrase structure from case marking. *Journal of Memory and Language*, *51*, 495–522.

Bornkessel, I., Zysset, S., Friederici, A.D., von Cramon, D.Y., Schlesewsky, M. (2005). Who did what to whom? The neural basis of argument hierarchies during language comprehension. *Neuroimage*, *26*(1), 221—233.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*, 55–73.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). The role of prominence information in the real-time comprehension of transitive constructions: a cross-linguistic approach. *Language and Linguistics Compass*, *3*(1), 19–58.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013a). Neurotypology: Modelling cross-linguistic similarities and differences in the neurocognition of language comprehension. In M. Sanz, I. Laka, M.K. Tanenhaus (Eds.), *The cognitive and biological basis for linguistic structure: new approaches and enduring themes*, (pp. 241–252). Oxford: Oxford University Press.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013b). Reconciling time, space and function: a new dorsal-ventral stream model of sentence comprehension. *Brain and Language*, *125*(1), 60–76. doi:10.1016/j.bandl.2013.01.010.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (to appear). Competition in argument interpretation: evidence from the neurobiology of language. In B. MacWhinney, A. Malchukov, E. Moravcsik (Eds.), *Competing motivations in grammar and usage*. Oxford: Oxford University Press.

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., Roehm, D., Schlesewsky, M. (2011). Think

globally: cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and Language*, *117*(3), 133–152.

Bourguignon, N., Drury, J.E., Valois, D., Steinhauer, K. (2012). Decomposing animacy reversals between agents and experiencers: an ERP study. *Brain and Language*, *122*(3), 179–189.

Brouwer, H., Fitz, H., Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143.

Burnham, K.P., & Anderson, D.R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.

Choudhary, K.K., Schlesewsky, Bickel, B., Bornkessel-Schlesewsky, I. (2010). An actor-preference in a split-ergative language: electrophysiological evidence from Hindi. In *Proceedings from 23rd annual meeting of the cuny conference on human sentence processing*. New York City.

Coulson, S., King, J.W., Kutas M (1998). ERPs and domain specificity: beating a straw horse. *Language and Cognitive Processes*, *13*, 653–672.

Crocker, M.W. (1994). On the nature of the principle-based sentence processor. In J.C. Clifton, L. Frazier, K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 245–266). Hillsdale: Erlbaum.

Croft, W.A. (2001). *Radical construction grammar: syntactic theory in typological perspective*. Oxford: Oxford University Press.

Dahl, Ö. (2008). Animacy and egophoricity: grammar, ontology and phylogeny. *Lingua*, *118*, 141–150.

Deco, G., Rolls, E.T., Romo, R. (2009). Stochastic dynamics as a principle of brain function. *Progress in Neurobiology*, *88*(1), 1–16.

Deco, G., Rolls, E.T., Albantakis, L., Romo, R. (2012). Brain mechanisms for perceptual and reward-related decision-making. *Progress in Neurobiology*.

DeLong, K.A., Urbach, T.P., Kutas, M. (2005). Probabilistic word preactivation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121.

Demiral, Ş., Schlesewsky, M., Bornkessel-Schlesewsky, I. (2008). On the universality of language comprehension strategies: evidence from Turkish. *Cognition*, *106*(1), 484–500.

DeWitt, I., & Rauschecker, J. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, E505–E514.

Fausey, C., & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin and Review*, *18*, 150–157.

Fausey, C.M., Long, B.L., Inamori, A., Boroditsky, L. (2010). Constructing ageny: the role of language. *Frontiers in Psychology*, *1*(162).

Federmeier, K.D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology*, *44*, 491–505.

Felleman, D., & Van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, *1*(1), 1–47.

Fodor, J.A. (1983). *Modularity of mind. An essay on faculty psychology*. Cambridge: MIT Press.

Frazier, L. (1987). Syntactic processing: evidence from dutch. *Natural Language and Linguistic Theory*, *5*, 519–559. doi:10.1007/BF00138988.

Frenzel, S., Schlesewsky, M., Bornkessel-Schlesewsky, I. (2011). Conflicts in language processing: a new perspective on the N400–P600 distinction. *Neuropsychologia*, *49*(3), 574–579.

Friederici, A.D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84.

Frisch, S., & Schlesewsky, M. (2001). The N400 reflects problems of thematic hierarchizing. *NeuroReport*, *12*(15), 3391–3394.

Frith, U., & Frith, C.D. (2010). The social brain: allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society B*, *365*, 165–176.

Grewe, T., Bornkessel, I., Zysset, S., Wiese, R., von Cramon, D.Y., Schlesewsky, M. (2006). Linguistic prominence and broca's area: The influence of animacy as a linearization principle. *Neuroimage*, *32*, 1395–1402.

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience*, *9*, 934–946.

Hagoort, P. (2003). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *Neuroimage*, *20*, 18–29.

Hagoort, P. (2005). On broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, *9*(9), 416–422.

Hahne, A., & Friederici, A.D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research*, *13*, 339–356.

Haupt, F.S., Schlesewsky, M., Roehm, D., Friederici, A.D., Bornkessel-Schlesewsky, I. (2008). The status of subject–object reanalyses in the language comprehension architecture. *Journal of Memory and Language*, *59*, 54–96.

Heekeren, H., Marrett, S., Bandettini, P., Ungerleider, L. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, *431*(7010), 859–862.

Huynh, H., & Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact f-distributions. *Journal of the American Statistical Association*, *65*(332), 1582–1589.

Jonides, J., Lewis, R., Nee, D., Lustig, C. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, *59*, 193–224.

Kempe, V., & MacWhinney, B. (1999). Processing of morphological and semantic cues in Russian and German. *Language and Cognitive Processes*, *14*(2), 129–171.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225.

Kolk, H.H., Chwilla, D.J., van Herten, M., Oor, P. (2003). Structure and limited capacity in verbal working memory: a study with event-related potentials. *Brain and Language*, *85*, 1–36.

Kretzschmar, F. (2010). *The electrophysiological reality of parafoveal processing: on the validity of language-related ERPs in natural reading*. PhD thesis, University of Marburg.

Kuperberg, G.R. (2007). Neural mechanisms of language comprehension: challenges to syntax. *Brain Research*, *1146*, 23–49.

Kutas, M., & Federmeier, K.D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470.

Kutas, M., & Hillyard, S.A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Lau, E.F., Phillips, C., Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933.

Leslie, A.M. (1995). A theory of agency. In D. Sperber, D. Premack, A.J. Premack (Eds.), *Causal cognition. A multidisciplinary debate* (pp. 121–141). Oxford: Clarendon Press.

Lewis, R., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science: A Multidisciplinary Journal*.

Lewis, R., Vasishth, S., Dyke, J.V. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.

Lewis, R.L. (2000). Specifying architectures for language processing: process, control, and memory in parsing and interpretation. In *Mechanisms for language processing*.

Lotze, N., Tune, S., Schlesewsky, M., Bornkessel-Schlesewsky, I. (2011). Meaningful physical changes mediate lexical-semantic integration: top-down and form-based bottom-up information sources interact in the N400. *Neuropsychologia*, *49*, 3573–3582.

MacWhinney, B., & Bates, E. (1989). *The cross-linguistic study of sentence processing*. New York: Cambridge University Press.

MacWhinney, B., Bates, E., Kliegl, R. (1984). Cue validity and sentence intERPretation in english, German and Italian. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 127–50.

Magnusdottir, S., Fillmore, P., den Ouden, D., Hjaltason, H., Rorden, C., Kjartansson, O., Bonilha, L., Fridriksson, J. (2012). Damage to left anterior temporal cortex predicts impairment of complex syntactic processing: a lesion-symptom mapping study. *Human Brain Mapping*.

Manning, C.D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Maris, E. (2004). Randomization tests for ERP topographies and whole spatiotemporal data matrices. *Psychophysiology*, *41*, 142–151.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*, 522–533.

Martin, A.E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, *58*, 879–906.

McElree, B. (2006). Accessing recent events. In B.H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 46). San Diego: Academic Press.

McElree, B., Foraker, S., Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91.

Muralikrishnan, R., Schlesewsky, M., Bornkessel-Schlesewsky, I. (2008). Universal and cross-linguistic influences on the processing of word order and animacy: neurophysiological evidence from Tamil. In *Proceedings from 21st annual CUNY conference on human sentence processing.* Chapel Hill.

New, J., Cosmides, L., Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, *104*(42), 16,598–16,603.

Newell, A. (1990). *Unified theories of cognition*. Cambridge: Harvard University Press.

Nieuwenhuis, S., Aston-Jones, G., Cohen, J.D. (2005). Decision making, the p3, and the locus coerulus-norepinephrine system. *Psychological Bulletin*, *131*, 510–532.

Oldfield, R.C. (1971). The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, *9*, 97–113.

Penolazzi, B., Vincenzi, M.D., Angrilli, A., Job, R. (2005). Processing of temporary syntactic ambiguity in Italian who-questions: a study with event-related potentials. *Neuroscience Letters*, *377*(2), 91–96.

Philipp, M., Bornkessel-Schlesewsky, I., Bisang, W., Schlesewsky, M. (2008). The role of animacy in the real time comprehension of Mandarin Chinese: evidence from auditory event-related brain potentials. *Brain and Language*, *105*(2), 112–133.

Rauschecker, J. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology*, *8*(4), 516–521.

Rauschecker, J., & Scott, S. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat Neurosci*, *12*(6), 718–724.

Roehm, D., Schlesewsky, M., Bornkessel, I., Frisch, S., Haider, H. (2004). Fractionating language comprehension via frequency characteristics of the human EEG. *NeuroReport*, *15*(3), 409–412.

Roehm, D., Bornkessel-Schlesewsky, I., Rösler, F., Schlesewsky, M. (2007). To predict or not to predict: influences of task and strategy on the processing of semantic relations. *Journal of Cognitive Neuroscience*, *19*, 1259–1274.

Sassenhagen, J., Kretzschmar, F., Mueller, E., Schlesewsky, M., Bornkessel-Schlesewsky, I. (2013). *Independent components dominating ERP responses to linguistic stimuli also respond to domain-general events*. Manuscript submitted for publication.

Schlesewsky, M., & Bornkessel, I. (2004). On incremental interpretation: degrees of meaning accessed during sentence comprehension. *Lingua*, *114*(9–10), 1213–1234.

Schlesewsky, M., Fanselow, G., Kliegl, R., Krems, J. (2000). The subject preference in the processing of locally ambiguous wh-questions in German. In B. Hemforth, & L. Konieczny (Eds.), *German sentence processing* (pp. 65–93). Dordrecht: Kluwer.

Schriefers, H., Friederici, A.D., Kuhn, K. (1995). The processing of locally ambiguous relative clauses in German. *Journal of Memory and Language*, *34*(4), 499–520.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Silverstein, M. (1976). Hierarchy of features and ergativity. In R.M. Dixon (Ed.), *Grammatical categories in Australian languages* (pp. 112–171). New Jersey: Humanities Press.

Simon, H.A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, *106*, 467–482.

Stevens, S.S. (1951). Mathematics, measurement and psychophysics. In S.S. Stevens (Ed.) *Handbook of experimental psychology* (pp. 1–49). Wiley.

Stroud, C., & Phillips, C. (2012). Examining the evidence for an independent semantic analyzer: an ERP study in spanish. *Brain and Language*, *120*, 108–126.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge: Harvard University Press.

Tomlin, R. (1986). *Basic word order: Functional principles*. London: Croom Helm.

van de Meerendonk, N., Kolk, H.H., Chwilla, D.J., Vissers, C.T.W.M. (2009). Monitoring in language perception. *Language and Linguistics Compass*, *3*, 1211–1224.

van de Meerendonk, N., Kolk, H.H., Vissers, C.T.W.M., Chwilla, D.J. (2010). Monitoring in language perception: mild and strong conflicts elicit different ERP patterns. *Journal of Cognitive Neuroscience*, *22*, 67–82.

Vincenzi, M.D. (1991). Filler-gap dependencies in a null-subject language: referential and non-referential whs. *Journal of Psycholinguistic Research*, *20*(3), 197–213.

Vissers, C.T.W.M., Chwilla, D., Kolk, H. (2006). Monitoring in language perception: the effect of misspellings of words in highly constrained sentences. *Brain Research*, *1106*, 150–163.

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition*, *75*, 105–143.

Vosse, T.G., & Kempen, G.A.M. (2008). Parsing verb-final clauses in German: garden-path and ERP effects modeled by a parallel dynamic parser. In B. Love, K. McRae, V. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society (Washington, DC, July 2008)*. Austin: Cognitive Science Society.

Vosse, T., & Kempen, G. (2009). The unification space implemented as a localist neural net: predictions and error-tolerance in a constraint-based parser. *Cognitive Neurodynamics*, *3*, 331–346.

Wang, L., Schlesewsky, M., Bickel, B., Bornkessel-Schlesewsky, I. (2009). Exploring the nature of the 'subject'-preference: evidence from the online comprehension of simple sentences in Mandarin Chinese. *Language and Cognitive Processes*, *24*(7/8), 1180–1226.

Weckerly, J., & Kutas, M. (1999). An electrophysiological analysis of animacy effects in the processing of object relative sentences. *Psychophysiology*, *36*(05), 559–570.

Whitney, C., Huber, W., Klann, J., Weis, S., Krach, S., Kircher, T. (2009). Neural correlates of narrative shifts during auditory story comprehension. *NeuroImage*, *47*, 360–366.

Wolff, S., Schlesewsky, M., Hirotani, M., Bornkessel-Schlesewsky, I. (2008). The neural mechanisms of word order processing revisited: Electrophysiological evidence from Japanese. *Brain and Language*, *107*, 133–157.

Zipf, G.K. (1935). *The psycho-biology of language*. Boston: Houghton Mifflin Company.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.

4. Decisions, Decisions: Quantifying Cue Contributions

> There are two kinds of statistics, the kind you look up and the kind you make up.
>
> Archie Goodwin

Following the success in quantifying prominence (Alday, Schlesewsky, and Bornkessel-Schlesewsky 2014, see Chapter 3), it became clear that the weighting of the individual prominence features presented a problematic set of free parameters. Free parameters represent not just potential underspecification in a model but also a source of possible interindividual variation, as the space of possible parameters also creates a solution space, where multiple solutions may exist for a given computational problem (cf. Howes, Lewis, and Vera 2009). We addressed the issue of both free parameters and interindividual variation in Alday, Schlesewsky, and Bornkessel-Schlesewsky (in press), where we presented a technique for estimating parameter weights at an individual level.

4.1. Brief Summary of Methods and Results

In a short experiment (30-40 minutes), subjects were shown verb-final sentences with two nominal arguments with rapid serial visual presentation. The nominal arguments appeared fully crossed for animacy (animate, inanimate), case (nominative, accusative) and number (singular, plural), with the verb agreeing in number with at least one of the nominal arguments. In addition to unambiguous sentences, the full-crossing also leads to globally ambiguous sentences, both well-formed (e.g. two plural nouns, because plural nouns in German are case-ambiguous) and ill-formed (e.g. double singular accusative or double singular nominative). After the end of the sentence, subjects had to answer a comprehension question within four seconds, whose answer indicated their assignment of the agent/actor role.

Due to the large factorial design, subjects were not exposed to every item in every condition; rather, sentences were chosen at random from a large pool of possible stimuli. Although all conditions were equally represented in the stimuli pool, the random sampling and randomized presentation means that not all conditions were necessarily equally represented for each subject.

Analyses were performed on a per-subject level using binomial regression for the assignment decision (actor-initial vs. actor-second) and linear regression for the reaction times. The reaction time analysis yielded large intra-subject variation, possibly due to the complexity of the task, and hence unreliable parameter estimates. The actor-assignment analyses, however, yielded comparatively low intra-subject variation and stable estimates. Data from four subjects did show an interesting amount of interindividual variance in terms of exact numerical estimates, but a remarkable consistency in qualitative estimates. In other words, subjects largely developed the same ranking for the prominence features but distinct weightings.

Subsequently, data from all four test subjects were pooled and analyzed using mixed-effects models. We then reanalyzed the EEG data from Alday, Schlesewsky, and Bornkessel-Schlesewsky (2014) with both the pooled weights and the weights from a single subject as well as the original *a priori* weights and found a comparable fit across models.

4.2. Relevance

In this paper, we demonstrated the feasibility of estimating the free parameters found in our computational implementation of the actor strategy quickly and easily at the single-subject level. Moreover, the framework we implemented is extensible, open source and based on free software and will run on any modern laptop. Together, this allows for the reduction of free parameters using rapidly acquired data.

4.3. Publication

Peer-Reviewed Article P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (in press). "Discovering Prominence and its Role in Language Processing: An Individual (Differences) Approach". In: *Linguistic Vanguard*. DOI: 10.1515/lingvan-2014-1013

My Contribution For this paper, I conceived, designed and carried out the experiment, including writing the necessary software. Additionally, I performed the analysis and wrote the entire paper with the exception of the section describing the eADM.

Discovering prominence and its role in language processing: An individual (differences) approach

Phillip M. Alday [a*], Matthias Schlesewsky [b], and Ina Bornkessel-Schlesewsky [a,c]

[a]University of Marburg
[b]Johannes-Gutenberg University Mainz
[c]University of South Australia

16 September 2014

Abstract

It has been suggested that, during real time language comprehension, the human language processing system attempts to identify the argument primarily responsible for the state of affairs (the "actor") as quickly and unambiguously as possible. However, previous work on a prominence (e.g. animacy, definiteness, case marking) based heuristic for actor identification has suffered from underspecification of the relationship between different cue hierarchies. Qualitative work has yielded a partial ordering of many features (e.g. MacWhinney, Bates, and Kliegl 1984), but a precise quantification has remained elusive due to difficulties in exploring the full feature space in a particular language. Feature pairs tend to correlate strongly in individual languages for semantic-pragmatic reasons (e.g., animate arguments tend to be actors and actors tend to be morphosyntactically privileged), and it is thus difficult to create acceptable stimuli for a fully factorial design even for binary features. Moreover, the exponential function grows extremely rapidly and a fully crossed factorial design covering the entire feature space would be prohibitively long for a purely within-subjects design.

Here, we demonstrate the feasibility of parameter estimation in a short experiment. We are able to estimate parameters at a single subject level for the parameters animacy, case and number. This opens the door for research into individual differences and population variation. Moreover, the framework we introduce here can be used in the field to measure more "exotic" languages and populations, even with small sample sizes. Finally, pooled single-subject results are used to reduce the number of

*corresponding author: phillip.alday@staff.uni-marburg.de

free parameters in previous work based on the extended Argument Dependency Model (Bornkessel and Schlesewsky 2006; Bornkessel-Schlesewsky and Schlesewsky 2009; Bornkessel-Schlesewsky and Schlesewsky 2013; Bornkessel-Schlesewsky and Schlesewsky in press; Alday, Schlesewsky, and Bornkessel-Schlesewsky 2014).

Multimedia: OpenSesame experiment and Python support scripts; sample stimuli; R scripts for analysis
Keywords: computational model; language processing; emergence; ambiguity resolution; actor identification; prominence; individual differences

Introduction

Parameter underspecification is a critical issue in modern linguistic models, with too many parameters typically dismissed to the periphery of qualitative description and "performance". The return on investment for working out the precise mechanistic and quantitative "details" of a model often seems too poor, especially in light of the many levels of linguistic variation: language > dialect > idiolect (inter-speaker) > intra-speaker. Yet, it is exactly these parameters and how they can vary that is interesting when discussing *language* instead of *a language*.

Even well-formulated psycholinguistic and neurolinguistic models often suffer from underspecification with many parameters omitted and many more never empirically estimated. Implemented computational models suffer less from the underspecification problem, but still have many issues with free parameters (Howes, Lewis, and Vera 2009) and researcher degrees of freedom (Simmons, Nelson, and Simonsohn 2011). Previously, we presented a computational model of language processing based on the interaction of weighted *prominence features* (Alday, Schlesewsky, and Bornkessel-Schlesewsky 2014). While our models provided a good fit for event-related potential data (i.e. electrical brain activity time-locked to a critical word within a sentence) which has a very poor signal-to-noise ratio, we nonetheless relied on a somewhat problematic conversion of ordinally scaled data to ratio-scaled data using simple logarithmic scaling. In the following we present a framework for empirically quantifying the parameters of well-defined computational models based on competition and constraint-satisfaction, focusing on the class of prominence-based models.

Using a small experiment and a basic statistical technique, we demonstrate that it is possible to estimate parameters at the single subject level in less than half an hour and perhaps a good cup of coffee. The ease of this approach opens the door to quantitative study of interindividual variation and linguistic settings in which only small samples of speakers are accessible (e.g. less-researched languages, clinical populations).

Prominence, the Extended Argument Dependency Model (eADM) and Actor-Centered Comprehension

Before turning to the parameter estimation approach that is the focus of the present paper, we will briefly describe the empirical neurocognitive model on which it is based. This framework will provide two critical concepts for the parameter estimation: *prominence* (the independent variable) and the *actor role* (the dependent variable).

The extended Argument Dependency Model (eADM) is a neurocognitive, and more recently neurobiologically grounded model of cross-linguistic language comprehension which places particular emphasis on the role of the "actor" participant (Bornkessel and Schlesewsky 2006; Bornkessel-Schlesewsky and Schlesewsky 2009; Bornkessel-Schlesewsky and Schlesewsky 2013; Bornkessel-Schlesewsky and Schlesewsky in press) The actor, a term taken from Role and Reference Grammar (Van Valin 2005) and termed Proto-Agent in other approaches (Dowty 1991; Primus 1999), refers to the event instigator / participant primarily responsible for the state of affairs being described. Based on the results of electrophysiological studies across a range of typologically diverse languages, the eADM posits that comprehension is actor-centered in the sense that the human language comprehension system endeavours to identify the actor participant as quickly and unambiguously as possible while comprehending a sentence. Accordingly, if several candidates are available, they compete for the actor role and actor competition has measurable neurophysiological repercussions (Bornkessel-Schlesewsky and Schlesewsky 2009; Alday, Schlesewsky, and Bornkessel-Schlesewsky 2014).

Actor identification in language processing is based both on domain-general features (e.g. animacy, certain movement parameters such as autonomous and/or biological motion, similarity to the first person etc.) and on language-specific features such as case marking or word order. In accordance with language-external observations regarding the importance of actor entities for mechanisms such as attentional orienting (New, Cosmides, and Tooby 2007) or social cognition (U. Frith and Frith 2010), the eADM assumes that the actor can be viewed as a cognitive and neurobiological attractor category, with domain-general actor features allowing for the *bootstrapping* of language-specific actor characteristics during language development (Bornkessel-Schlesewsky and Schlesewsky in press). Clearly, individual actor-related features will be more important for actor identification in certain languages as opposed to others (e.g. case marking in German, Japanese or Hindi versus English) and, within a particular language, some actor-related features will be weighted more strongly than others (Bates et al. 1982; Bates and MacWhinney 1989; Bates, Devescovi, and Wulfeck 2001; MacWhinney, Bates, and Kliegl 1984).

In this regard, parameter estimation – i.e. estimating the weighting of individual actor-related prominence features in a given language – becomes a central

modelling problem. In the following, we introduce an initial, empirically-based framework for parameter estimation that is flexible, based on open source software components and thus freely distributable and requires only a minimal time commitment from test subjects (i.e. native speakers of a given language). We thereby intend to establish a basis for examining (a) inter-individual differences in parameter weightings, and (b) lesser-studied languages for which only a small number of speakers is available to participate in linguistic experiments.

Previous Computational Work

Alday, Schlesewsky, and Bornkessel-Schlesewsky (2014) presented the first computational implementation of actor competition, with a strong focus on distinctness (similarity / distance in the space of prominence features) as a predictor of mean EEG signal in time windows previously associated with actor competition. Due to high variance in EEG data – both inter- and intra subject – mixed effect models with crossed random factors for subjects and items were used. Moreover, the dependent variable was not a single offline behavioral measurement but rather an online measure of brain activity. The independent variables were different notions of distance, i.e. different mathematical ways of combining prominence features and weights into a single distinctness score. While models involving neurophysiological data are arguable much closer to the actual biological reality of language processing, they measure processes at a level where the correspondence between conscious intuition and subconscious computation is far from clear. As such, while the parameter estimation used here is of utmost importance for continued work on such models, the results of the two approaches are not directly comparable but rather complementary.

Individual Experimentation

Robust parameter estimation must apply at the single subject (i.e. individual native speaker) level for several reasons. Language comprehension in a given language arguably involves a "strategy space" rather than hard-and-fast, deterministic processing strategy (see Howes, Lewis, and Vera (2009), for a more general cognitive perspective). Thus, by estimating inter-individual variability, we can establish an estimate of the breadth of the strategy space. Secondly, under certain circumstances (e.g. languages with few remaining or available speakers, clinical populations, children) it may not be possible to obtain data from a large pool of participants. Hence, the framework described here aims to provide a first step towards parameter estimation for individual participants, using the actor competition / prominence feature approach of the eADM as a test case. Of course, the approach is in principle applicable to any type of linguistic feature / model parameter.

Experiment

In order to maximize the portability and availability of individual parameter estimation, the experiment is restricted in equipment and duration. The experiment is programmed in OpenSesame (Mathôt, Schreij, and Theeuwes 2012), a freely available, Open Source software package written in Python for cognitive science experiments that runs on Windows, Mac OS X and Linux. No further equipment is required for the experiment itself. Similarly, the other parts of the proposed toolchain (R, Python and various packages for them) are all free software and available on all three platforms.

The much harder restriction is the duration of the experiment. While many psycholinguistic and neurolinguistic experiments last several hours per test subject, we restricted ourselves to a run time of between 30 and 40 minutes. This clearly restricts the number of trials available, which forces a tradeoff between a fully factorial exploration of differing conditions and the number of trials per condition. In the provided example experiment, the stimulus preparation script `load_data.py` generates the fullest factorial design allowed by its inputs (for our current sample stimuli, [ANIMACY x CASE x NUMBER] x [NP1, NP2], a total of 16 conditions, including violations) across all items and takes a random sample to generate 200 trials (see Table 1).

Table 1: Sample stimuli. All sentences began with *Gestern wurde erzählt* ('Yesterday, it was told'). Due to case syncretism in the German plural, all plural nouns were ambiguous and thus encoded in the subsequent models as being the average of nominative and accusative. The active task for the first sentence is _ *hat/haben angeremeplt* ('_s has/have bumped into'), with the two nouns placed on either side (left-right placement was random.) Because the article in German carries most case information and some number information, it was omitted in the task.

dass	die Pfarrer	die Magier	angerempelt	haben.
that	the pastors	the magicians	bumped-into	have
dass	den Wirt	den Einbrecher	eingeladen	hat.
that	the host.ACC	the thief.ACC	invited	has
dass	die Bürostühle	der Kellner	gespendet	hat.
that	the office chairs	the waiter.NOM	donated	has
dass	die Zäune	die Tischler	bedauert	haben.
that	the fences	the carpenters	regretted	have
dass	die Räume	der Veranstalter	eingeschaltet	hat.
that	the rooms	the organizer.NOM	turned-on.	has.
dass	den Bauer	der Obdachlose	gerettet	hat.
that	the farmer.ACC	the homeless person.NOM	saved	has

This leads to an extremely sparse sample which will differ from run to run. The variation is in and of itself interesting, as it gives some indication of minimal learnability requirements.

The task for the experiment is a comprehension question, asking either for the actor or the undergoer ("Who did X?" / "Someone did X to whom?" or passive variants of the same).[1] For the syntactically ambiguous or ungrammatical sentences, this task attempts to force the subject to arrive at some interpretation of the sentence (as would happen in normal conversation). This serves as an explicit task somewhat similar to traditional acceptability judgements. The task is also timed with a moderately hard timeout of 4 seconds, which should push the subject to answer more intuitively and less metalinguistically. The answer is encoded as having assigned actorhood to the first or the second NP (cf. Bates et al. 1982; MacWhinney, Bates, and Kliegl 1984; Li, Bates, and MacWhinney 1993; Kempe and MacWhinney 1999). "Correctness" is not a valid measure across conditions because the ambiguous and ungrammatical conditions lack a canonical answer. Moreover, the interesting question is how the prominence heuristic allows for decision under uncertainty. The response time is also recorded, under the assumption that prominence features misaligned with their weightings and the actor prototype will lead to higher reaction times.

Parameter Estimation

Ultimately, the computational problem presented by the actor strategy is classification. The language system must assign actorhood to a single argument and, in order to do that, depends on classifying an individual argument's probability of being an actor. Probabilitistic classification into two groups is a well-researched problem with many methods available. The simplest method, based on the general linear model, is probit regression (Bliss 1934).

The dependent variable in probit regression is a probabilisitic binary classification, while the independent variables are the feature encodings.[2] The model weights correspond to the coefficient estimates, allowing direct extraction of the weights and easy interpretation. Although the better known logistic regression yields similar results and has coefficients that are slightly easier to interpret (as an odds ratio), probit regression has several advantages for modelling the role of prominence features.[3] Logistic regression is more difficult to implement in a

[1] In initial tests, it seems that test-subjects felt more comfortable when only active or passive questions were presented. The results for one volunteer who completed both the mixed and pure passive variants were similar, resulting in the same rankings. However, the two trial runs are not directly comparable because each run used a different subset of the possible stimuli.

[2] Currently, the features are encoded as binary pairs with 1 for marked/more prominent and 0 for less prominent. For many things, a discrete scale seems unnatural, and the model can accommodate continuous scales on [0,1] without modification. This is currently used for ambiguous case marking, encoded as 0.5.

[3] The model coefficients in logit regression correspond to changes in the odds ratio for the dependent variable per unit change in the independent variable. In probit regression, the

Bayesian framework, which is a disadvantage for planned future work utilizing estimates from several sources (i.e. work involving non-flat priors and pooling). More importantly, probit regression is the better model for the dichotomization of a continuous variable. Prominence is a continuous variable, but the actor strategy is a dichotomization strategy,[4] and probit regression is thus better suited.[5]

For a transitive relationship, there are two arguments each carrying their own set of prominence features. While it is possible that the weights for the features are position dependent (i.e. that there is an interaction term for argument position by prominence), we make the simplifying assumption that this is not the case. Accordingly, we can collapse the two sets of features into a single set of pairwise differences, thus reducing the number of parameters to be estimated. This also makes the work more compatible with the types of models used in Alday, Schlesewsky, and Bornkessel-Schlesewsky (2014), where the weights actually applied to the pairwise differences. Here we use NP1 - NP2 to model expectations: NP1 - NP2 < 0 means that a more prominent argument comes in a later position, which is known to be dispreferred (preference for initial actors).

Due to the sparseness of the data, we also exclude interaction terms between features.

Sample Analysis

Actor Identification Sample analysis scripts and data sets collected from graduate students are included in the supplementary materials. In the following, we present the results from `sample01a`.[6]

For our regression, a 1 encodes an initial actor, while a 0 encodes an initial undergoer. Because of the mutual exclusivity of the actor-undergoer relation, we do not need to encode the other argument. We chose one to correspond to initial actor so that more prominence would correlate positively with more actorhood. Table 2 provides the results of the regression with `sample01a`.

While the exact meaning of the estimates in probit regression is difficult, the relationship in the size of the estimates is straight forward. Case clearly has

coefficients represent change in the z-score of the dependent variable per unit change in the independent variable. The errors in logistic regression follow the logistic function, while the errors in probit regression follow the normal distribution. This leads to the "tipping" behavior of emergent binary categories from a continuous scale. Both methods depend on a logarithmic link function.

[4] Indeed, the currently postulated processing model assumes that the threshold for this dichotomization is dynamic, adapting to contextual demands.

[5] To our knowledge, no psycholinguistic study has utilized probit regression. However, logistic regression, which is generally better known, and its mixed-effect extension have been proposed as a more appropriate way to analyze categorical responses, cf. (Jaeger 2008) and others.

[6] The number refers to the test subject, while the `a` refers to the task (active question, i.e. name the actor). Other possible codes are `p` (passive question, i.e. name the undergoer) and `b` (both types of questions randomly mixed).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.2878	0.2064	1.4	0.163
animacy	0.6269	0.167	3.8	0.0001743
case	1.3480	0.2318	5.8	6.014e-09
number	-0.4678	0.1596	-2.9	0.003381
index	0.0005	0.001767	0.3	0.7614

Table 2: **AIC:** 215.6 **Deviance:** 205.6 **Null Deviance:** 263.7 Results of probit regression for actor-initial order. The large residual deviance reflects the great variety of items, including syntactic und semantic violation conditions.

the strongest estimate, which fits well given that unambiguous case marking is known to work deterministically in German. Case also has the least amount of variance relative to its influence – this is reflected in the large z-score. Animacy has the next highest estimate and z-score, with both about half as large as for case. Number has the estimate and z-score with the smallest magnitude. Interestingly, the sign is also reversed for number. This could reflect the late disambiguating nature of number agreement in the verb final sentences. Index (i.e. trial number within the experiment) has a very small estimate and z-score, which indicates that the test subject was unable to develop and apply a strategy during the course of the experiment.[7]

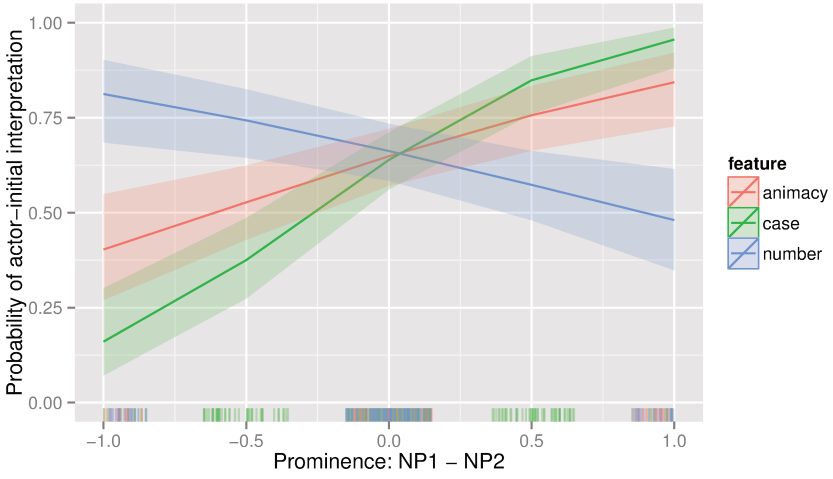


Figure 1: Actor Identification by Feature

This is also clear graphically. Figure 1 presents the likelihood of choosing an initial actor based on the difference in prominence. High initial prominence

[7] This of course says nothing about whether the test subject developed one beforehand while looking at an example run.

followed by low initial prominence – the high end of the scale – increases the odds of assigning actorhood to the first argument. (Shaded regions indicate the 95% confidence interval; the number of samples of each condition is shown as a rug plot.) The strength of a cue is reflected in the slope of the individual lines. The preference for an initial actor can also be clearly seen here. At 0, i.e. at a tie in prominence between the two arguments, all features show a preference for an actor-initial interpretation. Despite the low power from a short, unbalanced design, a clear ranking is visible.

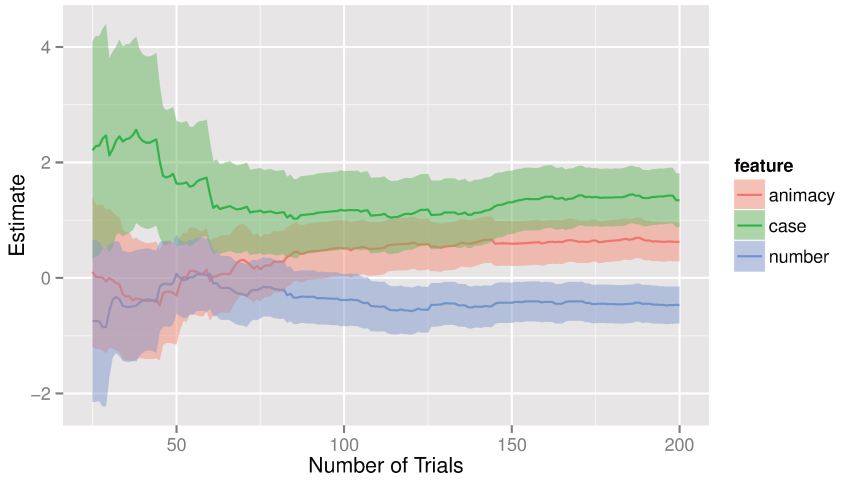


Figure 2: Convergence of Estimates

The emergence of such a clear ranking is also interesting for exploring the learnability of the actor strategy. Figure 2 shows the convergence of the parameter estimates by recomputing the model for the first n trials, starting with trial 25. Despite the low statistical power in such an experiment, the parameter estimates quickly sort themselves into a clear ranking. Figure 3 shows how this learning would look in terms of language processing, displaying the identification curve as in Figure 1, but after 50, 100, 150 and 200 trials. Again, after 50 trials, the strength of case is established, but the strength of the next strongest cue, animacy, is established after twice as many (100), and the third strongest cue settles down between 150 and 200 trials (1.5-2x as many as animacy), suggesting perhaps a rank-power law in cue strength (such as in Zipf's Laws).

Although we did not model separate interaction terms here, it is nonetheless interesting to consider how the model handles interaction. For this, we used the model to predict outputs for the grid of animacy x case, treating both as semi-continuous measures. Figure 4 shows the contour for this simulation holding constant index = 1 and number = 0 (plural). Darker tones indicate a

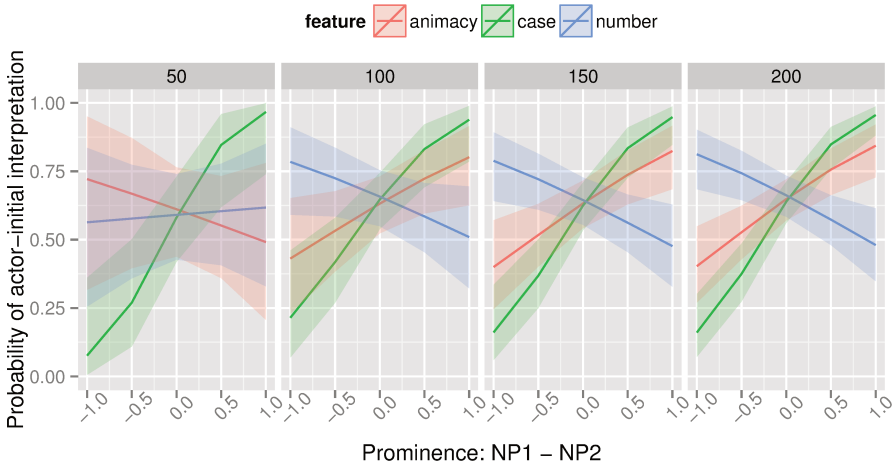


Figure 3: Effect of Estimate Precision on Actor Identification by Feature

higher probability of initial actor interpretation.[8] Using a physical metaphor, we can view the darker tones as being valleys and the lighter tones as being hills. Actorhood works as an attractor, with the ideal attractor basin being an initial, animate, nominative argument. However, even an inanimate initial nominative is more likely to be interpreted as an actor than an animate initial accusative – the basin slopes more sharply along case than along animacy.

Figures and models for four test subjects can be found in the appendix.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	863.2570	57.57	15	2.618e-34
animacy	-48.1050	44.68	-1.1	0.283
case	-66.2133	56.59	-1.2	0.2434
number	-16.8376	42.74	-0.39	0.694
index	-1.8386	0.4956	-3.7	0.0002706

Table 3: **AIC:** 2973.8 **Adjusted R^2:** 0.1 **Residual standard error:** 402.7 on 195 degrees of freedom, $F(4, 195) = 3.96$, $p = 0.0041$. Results of linear regression model for reaction time.

[8] Because the calculated model includes index and number as independent variables, they have to be nominally set. Index has little effect and so the choice is completely arbitrary: 1 is reasonable because it reflects a neutral state (not influenced by the odd experimental context) and is the smallest valuable possible. Similarly, the effect for number was quite small and the choice is again arbitrary. Number = 0 (plural) reflects the base state.

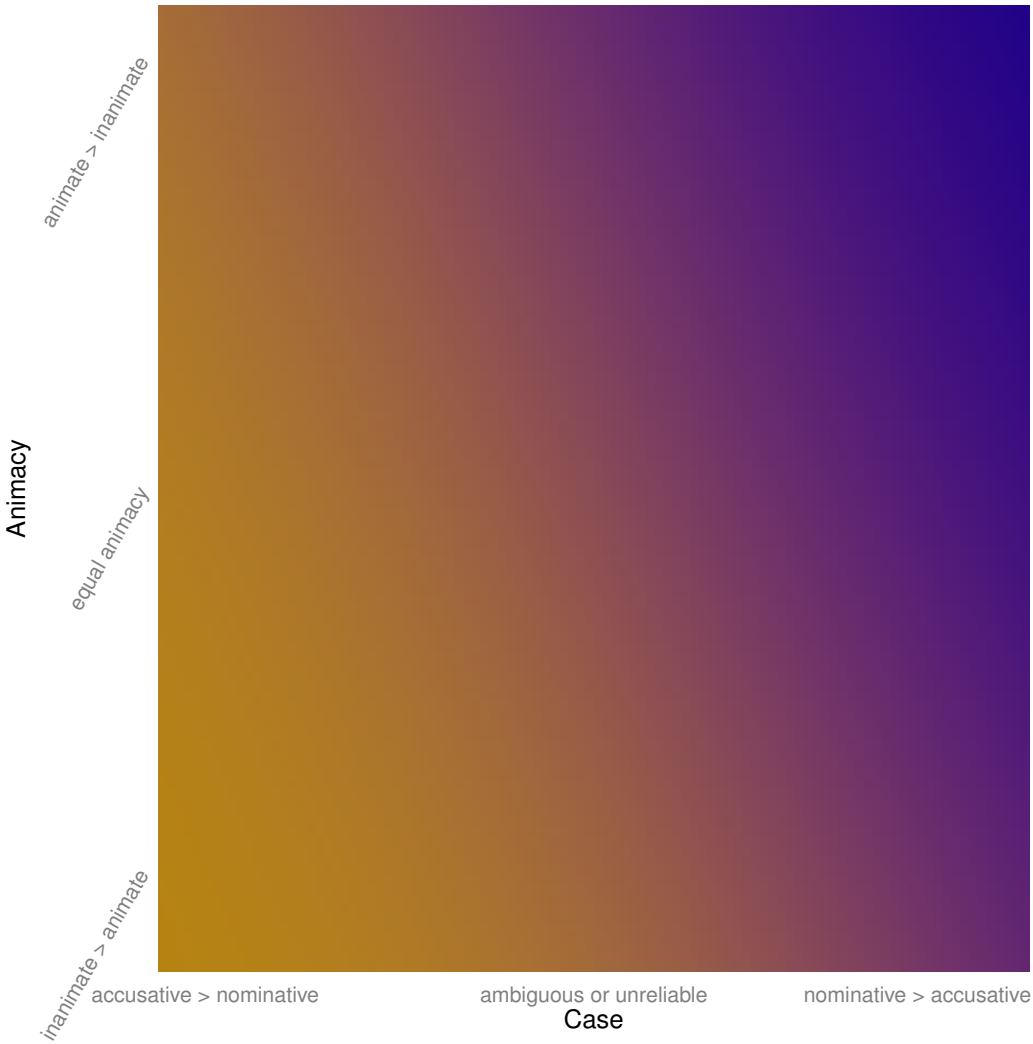


Figure 4: Individual Actor Space

Reaction Time The reaction time model shows the same general trend in the magnitude of the various estimates[9] but much higher variance. This variance leads to poor t-values. Index has a very small effect here – roughly 2ms reduction in reaction time for each successive item. This reflects a "training effect", where the test subject adapts to the experimental conditions and task. Although this effect is *statistically* highly significant, the magnitude is quite small.

The high degree of variance in the reaction time (related to the complexity of the task) and the limited power of such a small experiment leads to promising yet not reliable results. Training has a significant effect on test subject ability: a roughly 400ms reduction over the course of the experiment. Over a larger experiment, we expect that this effect would reach some natural asymptote as the test subjects become comfortable with the task and that the accompanying reduction in the variance would lead to larger t-values.

More research is required in order to investigate how best to integrate reaction time into the parameter estimation.

Comparison to Previous Work

	Df	AIC	BIC	logLik
model.apriori	11	395190	395291	-197584
model.emp	11	395223	395324	-197600
model.pooled	11	395252	395352	-197615

Table 4: Comparison of sdiff performance for a priori weights, weights from `sample01a`, and all samples. Models fitted to the EEG data in the N400 window (Alday et al, 2014).

Using the model estimates collected here, it is possible to compare empirical weights with the a priori estimates presented in Alday, Schlesewsky, and Bornkessel-Schlesewsky (2014). More precisely, the weight of a given feature is given by e^{β}, where β is the coefficient in the probit model. The exponentiation is necessary because the probit link function is logarithmic. We can also compute a mixed effects model over the four subjects tested thus far and extract the fixed-effect relevant coefficient for the pooled weightings. A comparison of the **apriori**, single-subject **emp**irical and **pooled** models of the best distinctness measure (sdiff, Alday, Schlesewsky, and Bornkessel-Schlesewsky 2014) is presented in Table 4. Critically, although two different sets of test subjects were used, the models are all extremely similar in their fit. The t-score for the distinctness metric was also similar.

[9] The reversed sign reflects that increased prominence aligns better with general expectations (actor preference) and thus *reduced* reaction times.

Better Data through Openness

An exciting aspect of estimating model parameters based on individual performance in a quick experiment is the possibility of making science accessible, available and touchable to everyone, which should open the door for exploration of areas where data acquisition has been difficult, such as the study of individual differences and less-researched languages. This depends on the software and underlying methodology being freely accessible, free to modify and free to distribute. All software used here is licensed under the GNU Public License (GPL). For the portions we wrote, we encourage you to fork us on Bitbucket: bug fixes and improvements are of course welcome, but example stimuli for different languages, sample data and alternative analyses would contribute far more towards our and the broader community's understanding of language.

Future Plans

Our own future plans for the software include a more integrated tool chain. Currently, the user has to install several programs (Python + several extensions, OpenSesame, R), but it should be possible to move core features into the OpenSesame experiment. The user would perhaps no longer have access to more advanced features (for which she would need some programming know-how anyway), but a core set of features for spontaneously testing a single-subject would fit into a single OpenSesame experiment file pool. As part of this, we are currently implementing a framework for combining the estimated parameters with an existing computational framework and providing an animation of how sentence processing in an individual works. All models are wrong, including ours, but some are useful (Box and Draper 1987) and the most useful are the ones everybody can see and tinker with.

More Data, Less Uncertainty

The framework presented here shows that parameter estimation is possible even with few trials from a single subject. With minimal equipment and quick parameter estimation, it is now possible to gather data from more languages, and we have another tool to remove our Indo-European blinders. At the other end of the spectrum, model fit may be poorer with a few test subjects than with a single subject, if the variance between subjects is large. This would be an interesting result within itself, indicating the size of the strategy space for a single language (population). More data from more languages and more individuals will help us to better understand both the cognitive mechanisms underlying language in general and the speaker-level adaption to a particular language.

Acknowledgements

Parts of the research reported here were supported by the LOEWE programme funded by the German state of Hesse. The authors would like to thank Alexander Dröge for contributing the stimuli for the German experiment as well as Jona Sassenhagen and Elisabeth Rabs for their extensive help in testing the experiment. Jona Sassenhagen also gave extensive feedback on the data visualization. Miriam Burk, Christina Lubinus, Pia Schoknecht and Fiona Weiß kindly consented to us using their performance as sample data.

Appendix: Figures for Four Subjects

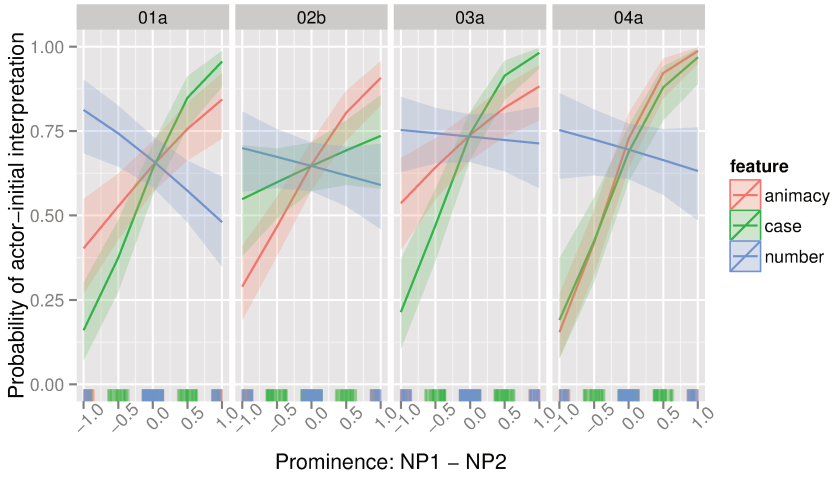


Figure 5: Actor Identification by Feature

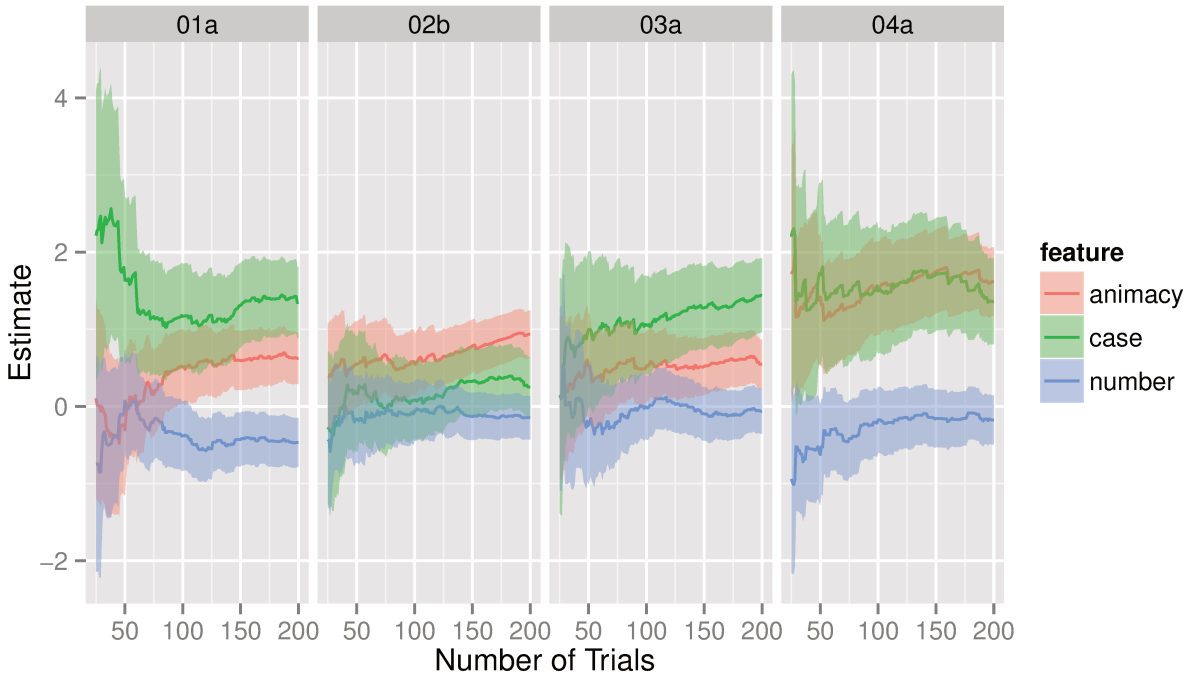


Figure 6: Convergence of Estimates

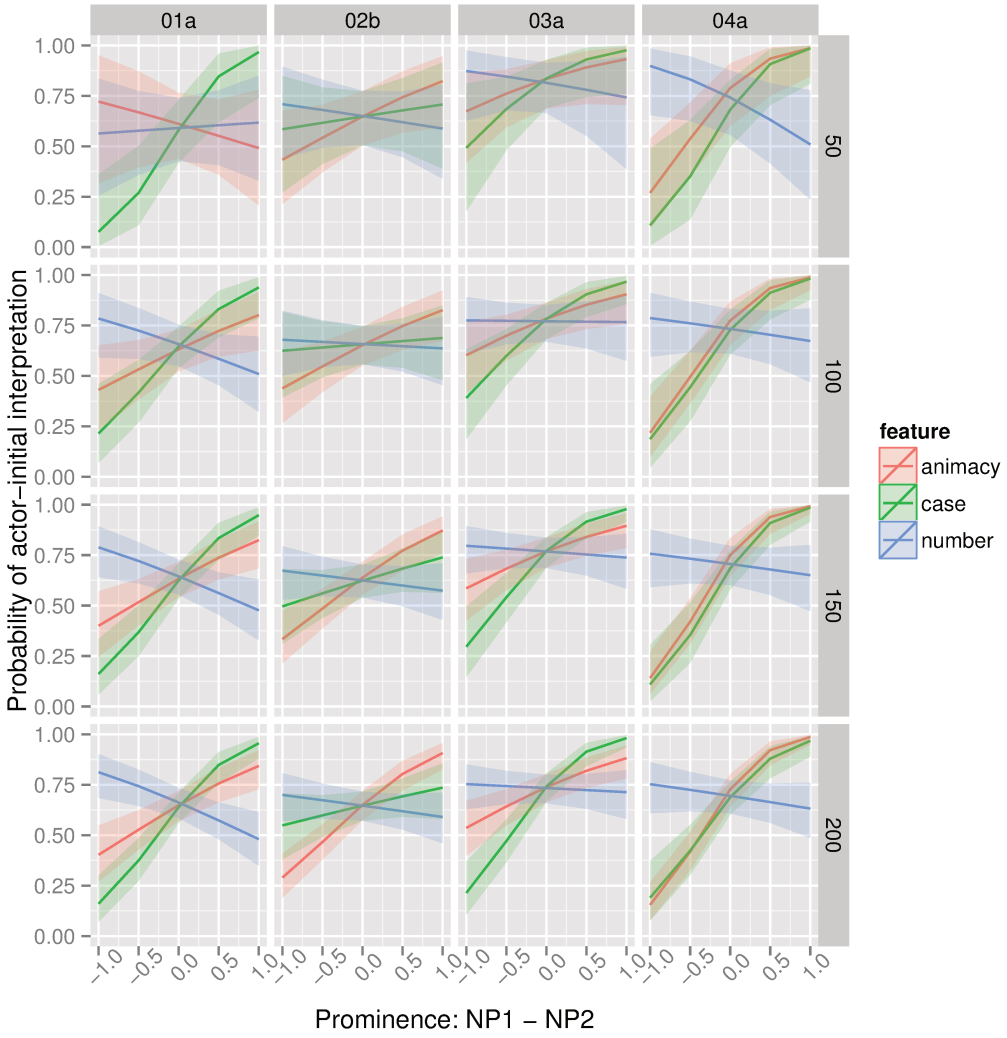


Figure 7: Effect of Estimate Precision on Actor Identification by Feature

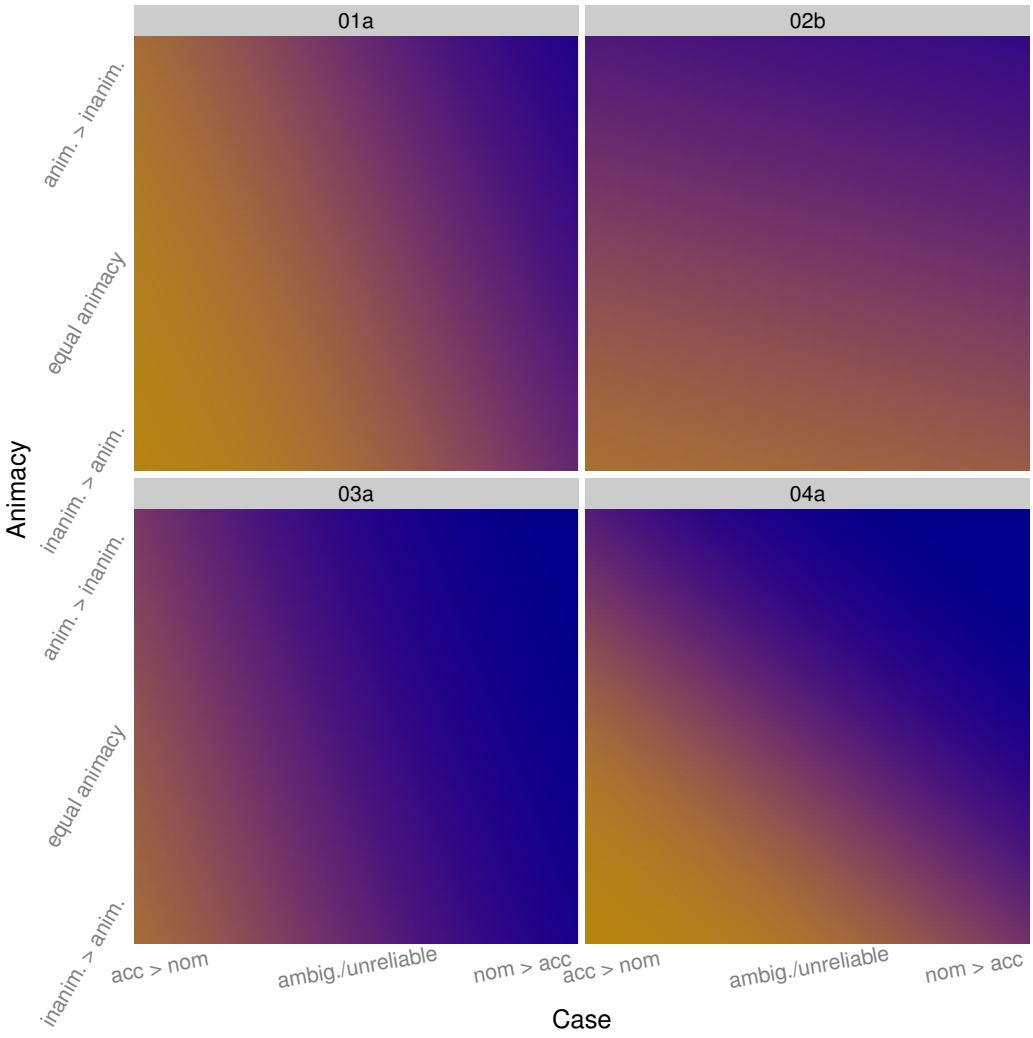


Figure 8: Individual Actor Space

References

Alday, Phillip M., Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. 2014. "Towards a Computational Model of Actor-Based Language Comprehension." *Neuroinformatics*.

Bates, Douglas, Martin Maechler, and Ben Bolker. 2013. *Lme4.0: Linear Mixed-Effects Models Using S4 Classes*.

Bates, Elizabeth, and Brian MacWhinney. 1989. "Cross-Linguistic Research in Language Acquisition and Language Processing." In *Proceedings of the World Conference on Basque Language and Culture*. San Sebastian: Basque Regional Government.

Bates, Elizabeth, Antonella Devescovi, and Beverly Wulfeck. 2001. "Psycholinguistics: A Cross-Language Perspective." *Annual Review of Psychology* 52: 369–96.

Bates, Elizabeth, Sandra McNew, Brian MacWhinney, Antonella Devescovi, and Stan Smith. 1982. "Functional Constraints on Sentence Processing: A Cross-Linguistic Study." *Cognition* 11: 245–99.

Bliss, C. I. 1934. "The Methods of Probits." *Science* 79 (2037): 38–39. doi:10.1126/science.79.2037.38.

Bornkessel, Ina, and Matthias Schlesewsky. 2006. "The Extended Argument Dependency Model: A Neurocognitive Approach to Sentence Comprehension Across Languages." *Psychological Review* 113 (4): 787–821.

Bornkessel-Schlesewsky, Ina, and Matthias Schlesewsky. 2009. "The Role of Prominence Information in the Real-Time Comprehension of Transitive Constructions: A Cross-Linguistic Approach." *Language and Linguistics Compass* 3 (1): 19–58.

———. 2013. "Reconciling Time, Space and Function: A New Dorsal-Ventral Stream Model of Sentence Comprehension." *Brain and Language*.

———. in press. "Competition in Argument Interpretation: Evidence from the Neurobiology of Language." In *Competing Motivations in Grammar and Usage*, edited by Brian MacWhinney, Andrej Malchukov, and Edith Moravcsik. Oxford: Oxford University Press.

Box, George E. P., and Norman R. Draper. 1987. *Empirical Model-Building and Response Surfaces*. Probability and Mathematical Statistics. Oxford, England: John Wiley; Sons.

Dahl, David B. 2014. *Xtable: Export Tables to LaTeX or HTML*.

Dowty, David. 1991. "Thematic Proto-Roles and Argument Selection." *Language* 67 (3). Linguistic Society of America: pp. 547–619.

Fox, John. 2003. "Effect Displays in R for Generalised Linear Models." *Journal of Statistical Software* 8 (15): 1–27.

Fox, John, and Jangman Hong. 2009. "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the Effects Package." *Journal of Statistical Software* 32 (1): 1–24.

Frith, U, and C D Frith. 2010. "The Social Brain: Allowing Humans to Boldly Go Where No Other Species Has Been." *Philosophical Transactions of the Royal Society B* 365: 165–76.

Howes, Andrew, Richard L Lewis, and Alonso Vera. 2009. "Rational Adaptation Under Task and Processing Constraints: Implications for Testing Theories of Cognition and Action." *Psychological Review* 116 (4): 717–51. doi:10.1037/a0017187.

Jaeger, T. Florian. 2008. "Categorical Data Analysis: Away from ANOVAs (Transformation or Not) and Towards Logit Mixed Models." *Journal of Memory and Language* 59 (4): 434–46. doi:10.1016/j.jml.2007.11.007.

Kempe, Vera, and Brian MacWhinney. 1999. "Processing of Morphological and Semantic Cues in Russian and German." *Language and Cognitive Processes* 14 (2): 129–71.

Li, Ping, Elizabeth Bates, and Brian MacWhinney. 1993. "Processing a Language Without Inflections: A Reaction Time Study of Sentence Interpretation in Chinese." *Journal of Memory and Language(Print)* 32 (2): 169–92.

MacWhinney, Brian, Elizabeth Bates, and Reinhold Kliegl. 1984. "Cue Validity and Sentence Interpretation in English, German and Italian." *Journal of Verbal Learning and Verbal Behavior* 23 (2): 127–50.

Mathôt, Sebastiaan, Daniel Schreij, and Jan Theeuwes. 2012. "OpenSesame: An Open-Source, Graphical Experiment Builder for the Social Sciences." *Behavior Research Methods* 44 (2). Springer New York: 1–11.

New, Joshua, Leda Cosmides, and John Tooby. 2007. "Category-Specific Attention for Animals Reflects Ancestral Priorities, Not Expertise." *Proceedings of the National Academy of Sciences* 104 (42): 16598–603. doi:10.1073/pnas.0703913104.

Primus, Beatrice. 1999. *Cases and Thematic Roles*. Tübingen: Niemeyer.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science*.

Team, R Core. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Van Valin, Robert D. 2005. *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20.

———. 2009. *Ggplot2*. Use R! Springer.

5. Natural Stories: New Perspectives on ERPs and the Role of Frequency

> Science is the belief in the ignorance of experts.
>
> Richard Feynman

One key advantage of prominence-based processing compared to grammatical-relation-based processing is the parsimonious account of the role of contextual (including extralinguistic) information. The domain of syntactic subjects is inherently the sentence level,[1] thus requiring a separate mechanism for integrating contextual-pragmatic information into language processing, while the domain of the actor role is at the level of the "plot" or event. We therefore expect that the actor heuristic will particularly shine in a larger and less constrained context. However, the high temporal resolution of EEG, or more precisely, the resulting extreme temporal sensitivity, becomes problematic in such naturalistic settings, and conventional wisdom states that it is not possible to extract meaningful results in such a setting. In Alday, Schlesewsky, and Bornkessel-Schlesewsky (submitted), we demonstrate the feasibility of analyzing EEG data from an experiment with naturalistic stimuli using the N400 as a test case and reported results compatible with previous findings from more controlled experimental conditions.

5.1. Brief Summary of Methods and Results

5.1.1. Methods

EEG data were collected from 57 test subjects, five of which were excluded from the final analysis, yielding 52 test subjects. Participants answered a short questionnaire after the experiment to control for attentiveness; no additional task was performed. The experimental

[1] This is a slight oversimplification, but it fits in well with traditional notions of "subject" used in parsing, which are based on formal syntax and grammatical relations (i.e. the strategy we have been comparing the actor strategy to). In particular, it is not coincidence that the central theme of a discussion is called the "subject of conversation" and that the focal point of a research area is called its "subject (matter)". Syntactic subjects do function as a cataphoric discourse device, indicating topicality, even in non topic-marking languages (cf. Bornkessel-Schlesewsky and Schlesewsky 2014). A clear example of this is the shift in focus from *John hit Mary* to *Mary was hit by John* — although the underlying physical-event semantics are arguably identical, the pragmatic usage is not.

stimulus was auditorily presented using a recording from Whitney et al. (2009) and consisted of a 23 minute story adapted from an out of copyright short story.

The EEG data were subsequently cleaned of electrical noise and artifacts via sine-wave removal and ICA (Winkler, Haufe, and Tangermann 2011) and high-pass filtered at 0.3 Hz. The single-trial mean amplitude in the time window was extracted from 1682 epochs per subject, time locked to the start of content words. For this exploratory study, analyses were restricted to the time window 300-500ms post onset, a typical window for the N400, because the N400 is one of the most extensively studied language related components and has proven to be a very robust effect (cf. Kutas and Federmeier 2011). Nonetheless, the techniques used here are quite general and should apply equally well to other ERP effects.

5.1.2. Results

Analysis was performed using linear mixed-effects models and restricted to the centro-parietal midline electrodes (Cz, CPz and Pz). As an initial proof of concept, models were calculated for both corpus frequency and relative frequency within the story; only the relative-frequency model was improved by including ordinal position in the story as a covariate. Relative frequency with ordinal position yielded a similar fit to corpus frequency. Both frequency models showed the well-established effect for decreasing N400 amplitude with increasing frequency.

The Role of Frequency

Based on the similarity between the best models for corpus frequency and relative frequency, frequency seems to be a dynamic entity. In line with findings that the N400 indexes the fulfillment of expectations (or lack thereof), we suggested that frequency be viewed as a prior in a Bayesian sense. In particular, corpus frequency reflects an always available, weakly informative prior, while relative frequency reflects a locally constrained prior. As more information becomes available, the informativeness of the local prior increases, which is reflected in the interaction between relative frequency and ordinal position.

Prominence Features

Additionally, in line with previous findings on the N400, effects were found for animacy (Weckerly and Kutas 1999; Philipp et al. 2008) as well as the interaction of morphology and position (cf. Frisch and Schlesewsky 2001), which also matches predictions from the eADM for processing low-prominence referents. Finally, combination models were computed using the different frequency measures, ordinal position and orthographic length (as a proxy for stimulus duration) with overall improved model fits but similar results. Interestingly, the different prominence features interacted differently with the two frequency measures.

The Nature of the ERP

The replication of results from the literature using a fixed time window in a heterogeneous context presents a challenge for the usual perspective on the nature of ERP components. While most theories of the ERP do not explicitly assume that ERPs operate on discrete entities, traditional strictly controlled experimental environments have yielded homogeneous data with fairly well-defined waveforms that give the impression of discrete processing phases. We argue that ERPs instead reflect continuous information processing, whose peak-like appearance in traditional experiments results from the uneven information flow due to rapid serial visual presentation or fixed-length auditory stimuli. From this perspective, we then argue that component latencies reflect not only position within the processing pipeline but also the time scale upon which they operate. As a specific example, the N200/MMN and N400 index mismatches, but their particular sensitivity is a reflection of their time scale. The N200/MMN is sensitive to mismatches perceptible on a very short time scale, such as physical properties of the stimulus, while the N400 operates on a time scale roughly corresponding to single worlds and is thus sensitive to semantic mismatches at the level of individual words. A similar argument applies to P300 and P600 effects.

Although counterintuitive, this perspective is compatible with existing theories of electrophysiology. It also fits in well with the recent suggestion within the broader eADM framework that human language comprehension arises from a difference in quantity and not quality between humans and other primates (Bornkessel-Schlesewsky, Schlesewsky, et al. in press), as increasing quantity could lead to additional increasingly large temporal scales and hence more hierarchical complexity.

5.2. Relevance

In this paper, we demonstrated the feasibility of analyzing electrophysiological data elicited from a more ecologically valid setting than previously thought possible. The compatibility with existing results has two implications: (1) the proposed method yields valid results and (2) the rigors of controlled experimental manipulations do not induce a special processing mode in the brain distinct from "natural", normal language processing. This opens the doors to more comprehensive studies of language "at scale", including the effects of rich, extended contexts.

Using richer contexts showed a dynamicism for the effect of frequency, which previously had not been apparent. Additionally the role of prominence features was shown to be measurable in a naturalistic environment, which suggests that previous results were not merely the result of experimental manipulation. This in turn opens the door for studying prominence features and their interactions with one another, which are postulated to be every bit as important as their individual effects, in a way not possible using traditional factorial designs.

Finally, this manuscript proposes a new perspective on the nature of ERP components, which, while compatible with existing neurobiological theories of the ERP, contradicts the usual intuition about the nature of individual components.

5.3. Publication

Peer-Reviewed Article P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (submitted). "Electrophysiology Reveals the Neural Dynamics of Naturalistic Auditory Language Processing: Event-Related Potentials Reflect Continuous Model Updates". In: *Journal of Neuroscience*

Conferences P. M. Alday, A. Nagels, et al. (2011). *Actor Identification in Natural Stories: Qualitative Distinctions in the Neural Bases of Actor-related Features*. Talk presented at the Neurobiology of Language Conference. Annapolis

P. M. Alday, J. Sassenhagen, and I. Bornkessel-Schlesewsky (2014b). *Tracking the Emergence of Meaning in the Brain during Natural Story Comprehension*. Poster presented at the International Conference on Cognitive Neuroscience (ICON). Brisbane

P. M. Alday, J. Sassenhagen, and I. Bornkessel-Schlesewsky (2014a). *Neural Signatures of Incremental Text Processing Correlate with Word Entropy in a Natural Story Context*. Poster presented at the Society for Neurobiology of Language Conference. Amsterdam

My Contribution For this paper, I extended an existing experiment by collecting data from additional test subjects, developed the necessary analysis techniques by combining modern statistical and signal processing methods, proposed an interpretation and wrote the majority of the final manuscript.

5. *Natural Stories: New Perspectives on ERPs and the Role of Frequency*

Full title: Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: event-related potentials reflect continuous model updates
Running title: Electrophysiology of natural language processing
Date: February 2015
Authors: Phillip M. Alday (a) Matthias Schlesewsky (b) Ina Bornkessel-Schlesewsky (a,c)
Corresponding and communicating author: Phillip M. Alday, phillip.alday@staff.uni-marburg.de
Affiliations and Addresses:
a:
Department of Germanic Linguistics
University of Marburg
Deutschhhausstr. 3
35032 Marburg
Germany
b:
Department of English and Linguistics
Johannes-Gutenberg University Mainz
Jakob-Welder Weg 18
55128 Mainz
Germany
c:
Cognitive Neuroscience Laboratory; School of Psychology, Social Work & Social Policy
University of South Australia
GPO Box 2471, Adelaide, SA 5001
Australia

Number of Figures: 5
Number of Tables: 11
Number of Pages: 19
Abstract length: 113 words
Introduction length: 454 words
Discussion length: 1305 words
Conflict of Interest: The authors declare no conflicts of interest.
Acknowledgements: We would like to thank Fritzi Milde for her help in annotating the stimulus, Jon Brennan for helpful discussions related to naturalistic stimulus presentation and EEG/MEG measures, Jona Sassenhagen and Franziska Kretzschmar for engaging discussions, and Jona Sassenhagen again for his help with EEGLAB.

5. *Natural Stories: New Perspectives on ERPs and the Role of Frequency*

Abstract

Recent advances in statistical computing have made it possible to use experimental designs beyond traditional factorial manipulations, thus allowing investigations into the neurobiology of cognition to employ more naturalistic and ecologically valid designs. Using mixed effects models for epoch-based regression, we demonstrate the feasibility of examining event-related potentials (ERPs) to study the neural dynamics of auditory language processing in a naturalistic setting. We replicated previous findings from the literature as a proof of concept, despite the large variability between trials during naturalistic stimulation. This suggests a new perspective on ERPs, namely as a continuous modulation reflecting continuous model updates (cf. Friston, 2005) instead of a series of discrete and essentially sequential processes.

Introduction

In real-life situations, the human brain is routinely confronted with complex, continuous and multimodal sensory input. Such natural stimulation differs strikingly from traditional laboratory settings, in which test subjects are presented with controlled, impoverished and often isolated stimuli (e.g. individual pictures or words) and often perform artificial tasks. Accordingly, cognitive neuroscience has seen an increasing trend towards more naturalistic experimental paradigms (Hasson and Honey, 2012), in which complex, dynamic stimuli (e.g. movies, natural stories) are presented without an explicit task (e.g. Hasson et al., 2004, 2008; Skipper et al., 2009; Whitney et al., 2009; Lerner et al., 2011; Brennan et al., 2012; Conroy et al., 2013; Hanke et al., 2014).

In spite of being uncontrolled, naturalistic stimuli have been shown to engender distinctive and reliable patterns of brain activity (Hasson et al., 2010). However, they also pose unique challenges with respect to data analysis (e.g. Hasson and Honey, 2012 cf. also the 2014 Real-life neural processing contest, in which researchers were invited to develop novel analysis techniques for brain imaging data obtained using complex, naturalistic stimulation). To date, the discussion of these challenges has focused primarily on neuroimaging data and, in the majority of cases, on visual stimulation. Naturalistic stimuli in the auditory modality, by contrast, give rise to an additional and unique set of problems, particularly when examined using techniques with a high temporal resolution such as Electroencephalography (EEG) or Magnetoencephalography (MEG). Consider the case of language processing: in contrast to typical, controlled laboratory stimuli, a natural story or dialogue contains words that vary vastly in length, a stimulus property to which EEG and MEG are particularly sensitive because of their superb temporal resolution. The characteristic unfolding over time of auditory stimuli is already evident when evoked electrophysiological responses are compared in more traditional, controlled studies – the endogenous components show increased latency and a broader temporal distribution (see for example Wolff et al., 2008, where the same study was carried out in the auditory and visual modalities). EEG and MEG studies with naturalistic stimuli consequently tend to use the less naturalistic visual modality (segmented, rapid-serial visual presentation, e.g. Frank et al. (2015); or natural reading combined with eye-tracking, e.g. Kretzschmar et al. (2013); Hutzler et al. (2007)).

Given current data analysis techniques, these distinctive properties of the auditory modality impose severe limitations on our ability to conduct and interpret naturalistic auditory experiments, particularly when seeking to address questions related to time course information in the range of tens – or even hundreds – of milliseconds. Here, we present a new analysis technique that addresses this problem using linear mixed effects modeling. We further provide an initial demonstration of the feasibility of this approach for studying

auditorily presented naturalistic stimuli using electrophysiology.

Materials and methods

Participants

Fifty-seven right-handed, monolingually raised, German native speakers with normal hearing, mostly students at the Universities of Marburg and of Mainz participated in the present study after giving written informed consent. Three subjects were eliminated due to technical issues, one for psychotropic medication, and one for excessive yawning, leaving a total of 52 subjects (mean age 24.2, std.dev 2.55; 32 women) for the final analysis.

Experimental stimulus and procedure

Participants listened passively to a story roughly 23 minutes in length while looking at a fixation star. The story recording, a slightly modified version of the German novella "Der Kuli Klimgun" by Max Dauthendey read by a trained male native speaker of German, was previously used in an fMRI study by Whitney et al. (2009). Subjects were instructed to blink as little as possible, but that it was better to blink than to tense up from discomfort. After the auditory presentation, test subjects filled out a short comprehension questionnaire to control for attentiveness.

EEG recording and preprocessing

EEG data was recorded from 27 Ag/AgCl electrodes fixed in an elastic cap (Easycap GmbH, Herrsching, Germany) using a BrainAmp amplifier (Brain Products GmbH, Gilching, Germany). Recordings were sampled at 500 Hz, referenced to the left mastoid and re-referenced to linked mastoids offline. Using sine-wave fitting, the EEG data were first cleaned of line noise, and then automatically cleaned of artifacts using ICA (Winkler et al., 2011). The ICA decomposition was performed via Adaptive-Mixture ICA on data high-pass filtered at 1 Hz and downsampled to 100Hz (Palmer et al., 2007) and backprojected onto the original data high-pass filtered at 0.1 Hz. Subsequently, the original data were high-pass filtered at 0.3 Hz and 1682 segments extracted per test subject, time locked to the onset of content words. This filter was chosen to remove slow signal drifts as traditional baselining makes little sense in the heterogeneous environment of

naturalistic stimuli (cf. Frank et al., 2015, who also found that a heavier filter helped to remove correlation between the pre-stimulus and component time windows).

Data analysis

For this initial exploratory study, we focus on the N400 event-related potential (ERP), a negative potential deflection with a centro-parietal maximum and a peak latency of approximately 400 ms, but the methodology should apply to other ERP components as well.

The N400 is well suited to the purposes of the present study, since it is highly robust and possibly the most researched ERP component in the neurobiology of language (see Kutas and Federmeier, 2011, for a recent review). Although the exact neurocognitive mechanism(s) that the N400 indexes are still under debate, it can be broadly described as being sensitive to manipulations of expectation and its fulfillment (cf. Kutas and Federmeier, 2000, 2011; Hagoort, 2007; Lau et al., 2008; Lotze et al., 2011). This can be seen most clearly in the sensitivity of the N400 to word frequency, cloze probability and contextual constraint, but also to manipulations of more complex linguistic cues such as animacy, word order and morphological case and the interaction of these factors (Bornkessel and Schlesewsky, 2006; Bornkessel-Schlesewsky and Schlesewsky, 2009). Importantly for the examination of naturalistic stimuli, N400 amplitude is known to vary parametrically with modulations of these cues, thus making it well suited to modeling neural activity based on continuous predictors and activity fluctuations on a trial-by-trial basis (cf. Cummings et al., 2006; Roehm et al., 2013; Sassenhagen et al., 2014).

More recently, researchers have attempted to quantify expectation using measures derived from information theory, such as surprisal. These have enjoyed some success as a parsing oracle in computational psycholinguistics (Hale, 2001; Levy, 2008) and have been shown to correlate with N400 amplitude for naturalistic stimuli (real sentences taken from an eye-tracking corpus) presented with RSVP (Frank et al., 2015).

We examined single trial mean amplitude in the time window 300-500ms, a typical time window for the N400 effect (Kutas and Federmeier, 2011; cf. Frank et al., 2015). To simplify the analysis, both computationally and in terms of comprehensibility, only data from the electrodes Cz, CPz, and Pz were used, following the centro-parietal distribution of the N400. Data from these electrodes were analyzed using linear mixed effects models (Pinheiro and Bates, 2000)

Statistical Methods

Mixed effects models present several advantage over traditional repeated-measures ANOVA for the exploration presented here. First, they yield quantitative results, estimating the actual difference between conditions instead of merely the significance of the difference. Second, they can easily accommodate both quantitative and qualitative independent variables, allowing us to integrate measures such as frequency without relying on dichotomization and the associated loss of power (cf. MacCallum et al., 2002). Finally, they are better able to accommodate unbalanced designs than traditional ANOVA methods.

A major topic of debate in the application of mixed models to psycho- and neurolinguistic data is the structure of the random effects. While Baayen et al. (2008) recommend forward selection of the random-effect structure, starting from the minimal intercepts-only structure, Barr et al. (2013) recommend backwards selection from the maximal random-effect structure, and Barr (2013) takes this suggestion one step further and suggests including *all* interactions. In practice, Barr et al. (2013)'s suggestion is somewhat problematic as complex random effect structures are costly to compute and often fail to converge on real data sets. Moreover, the backward selection procedure suggested by Barr et al potentially leads to issues with overparameterization (see for example this thread and this comment by Doug Bates, author of the popular `lme4` and `nlme` packages, on the mailing list for the R special interest group for mixed models). Another suggestion common to the mixed model literature is to follow the random-effect structure that best models the experimental design (see for example the GLMM wiki).

Here, we use a minimal model with a single random-effect term for the intercept of the individual subjects. This is equivalent to assuming that all subjects react the same way to each experimental manipulation but may have different "baseline" activity. This is a plausible assumption for an initial exploration, where we focus less on interindividual variation and instead focus on the feasibility of measuring population-level effects across subjects. Furthermore, this is not in violation of Barr et al. (2013)'s advice, which is explicitly directed at *confirmative* studies. The reduced random-effect structure reduces the number of parameters to estimate, which (1) greatly increases the computational tractability of the exploration at hand and (2) allows us to focus the relatively low power of this experimental setup on the parameters of interest.

We omit a random effect term for "item" as there are no "items" in the traditional psycholinguistic sense here. A random effect for "lexeme" is also not appropriate because while some lexemes appear multiple times (e.g., "Ali", the name of the title character), many lexemes appear only once and this would lead to overparameterization.

No parameter for electrode was introduced into the model as this would have reduced overall power and

increased computational complexity. The three electrodes used are close enough together that they should all have correlated values, which means more data and thus more precise estimates.

In order to make the models and their fits more readily comparable with each other, all models were estimated with Maximum Likelihood Estimation (i.e. with `REML=FALSE` in `lme4`, cf. Pinheiro and Bates, 2000; Baayen et al., 2008). For simpler models, we present the full model summary, including an estimation of the inter-subject variance and all estimated coefficients for the fixed effects, but for more complicated models, we present Type-II Wald χ^2 tests for readability. Type-II Wald tests have a number of problems (see for example the discussion here), but even assuming that their results yield an anti-conservative estimate, we can use them to get a rough impression of the overall effect structure (cf. Bolker et al., 2009). For groups of several similar models, e.g. adding or removing a single term, we generally present a likelihood ratio test.

For the model summaries, we view $|t| > 2$ (i.e., the estimate of the coefficient is more than twice as large as the error in the estimate) as being indicative of a reliable estimate. We view $|t| < 2$ as being unreliable estimates, which may be an indicator of low power or of a generally trivial effect. (We note that Baayen et al. (2008) use $|t| > 2$ as approximating the 5%-significance level.) For the Type-II Wald tests, we use the p-values as a rough indication of reliability of the estimate across factor levels, which each receive their own coefficient in the model (e.g. a single "morphology" factor in the Wald tests, but two coefficients for the three levels: "unambiguous nominative" and "unambiguous accusative", with the third level "ambiguous" being encoded as part of the intercept term.) This will become clearer with an example, and so we begin with a well-known modulator of the N400: frequency of a word in the language as a whole.

Results

Proof of Concept: Frequency

In a natural story context, traditional ERP methodology with averaging and grand averaging yields waveforms that appear uninterpretable or even full of artifacts. From the perspective of continuous processing, this is not surprising at all. Some information is present before word onset via context (e.g. modifiers before a noun), which leads to ERPs that seem to show an effect very close to or even before zero. Some words are longer than others, which leads to a smearing of the traditional component structure, both at a single-trial and at the level of averages. These problems are clearly visible in Figure 1, which shows an ERP image (Jung et al., 2001) for a single participant for initial accusatives, which are known to be dispreferred to initial nominatives (Frisch and Schlesewsky, 2001) and thus should engender an N400 effect. However, a modula-

tion of the ERP signal may nonetheless be detectable in the N400 time window, indexing the processing of the new information available at the trigger point. As a proof of concept for our method, we first examine the well-established effect of frequency on N400 amplitude (see Kutas and Federmeier, 2011 for a review), the results of which are in present in Table 1.

Corpus Frequency

The frequency of a word in the language as whole, *corpus frequency*, is known to correlate with N400 amplitude and interact with cloze probability (see Kutas and Federmeier, 2011 for a review). Using the logarithmic frequency classes from the Leipzig Wortschatz, we can see in Table 1 that corpus frequency has a small, but reliable effect (only -0.6 µV per frequency class, but $t < -13$ in the N400 time window). This is exactly what the literature predicts – frequency is not dominant in context-rich environments, but plays a distinct role (Dambacher et al., 2006; cf. Kutas and Federmeier, 2011).

Moreover, corpus frequency is insensitive to context as it represents global and not local information. Adding index, i.e. the ordinal position in the story, to the corpus frequency model does not improve it, as shown in Table 2. This lack of improvement reflects the context insensitivity of corpus frequency, which is a global measure not dependent on the story context. (At the sentence level, there is some evidence that ordinal position plays modulates the role of frequenc, e.g. Van Petten and Kutas (1990), but the ordinal position in the story averages out this modulation across the entire story. Short stimuli are dominated by boundary effects but longer naturalistic stimuli are not.) This is also visible in Figure 2, in which the regression lines have roughly the same slope regardless of index.

Relative Frequency

The relative frequency of a word in a story is also known to correlate with N400 amplitude (cf. Van Petten et al., 1991, which found a repetition priming effect for words repeated in natural reading). This is seen indirectly in repetition priming (which is essentially a minimal, binary context) and information-theoretic surprisal, which can be seen as a refinement of relative frequency. In contrast to corpus frequency, incorporating index does improve the relative frequency model (Table 3). The improved model is presented in Table 4; relative frequency was divided into classes using the same algorithm as for corpus frequency, but applied exclusively to the smaller "corpus" of the story. Interestingly, the interaction of index with relative frequency has a smaller estimated value than the main effect for index, but a larger t-value, indicating a more reliable estimate and a clearer effect. This interaction is visible in the clearly differing slopes in Figure

3. The main effect for relative frequency has both a larger estimate and t-value than the terms with index.

Frequency is Dynamic

Somewhat surprisingly, the model for relative frequency with index provides nearly as good a fit as the model for corpus frequency (Table 5). Adopting a Bayesian perspective on the role of prior information (here: frequency), this result is less puzzling. From a Bayesian perspective, corpus frequency is a nearly universally applicable but weakly informative prior on the word, while the relative frequency is (part of) a local prior on the word. This is clearly seen in the interaction with position in the story – corpus frequency's informativeness does not improve over the course of the story, but relative frequency's does as the probability model it represents is asymptotically approached. Thus, (corpus) frequency makes a small but measurable contribution in a rich context, while it tends to dominate in more restricted contexts. Relative frequency becomes a more accurate model of the world, i.e. a more informative prior, as the length of the context increases. Corpus frequency is thus in some sense an approximation of the relative frequency calculated over the context of an average speaker's lifetime of language input.

In this sense, we can say that frequency is dynamic and not a static, inherent property of a word. In the absence of local context, frequency is calculated according to the most general context available – the sum total of language input. With increasing local context, a narrower context for calculating frequency is determined, increasingly cut down from the global language input (which now of course includes the new local context). From this perspective, it is less surprising that a model incorporating the development of relative frequency over time yields results that are nearly as good as a model based on the well-established effect of corpus frequency. Frequency is an approximation for expectation, and a larger context leads to expectation that is better predicted from that context than from general trends.

The present approach: examining complex influences within a fixed epoch

The results for frequency in both its forms are not surprising in the sense that they match previous results. Nonetheless, it is perhaps somewhat surprising that it is possible to extract the effects in such a heterogeneous and noisy environment. Part of the problem with the type of presentation in Figure 1 is that the influences on N400 (and ERP in general) amplitude are many, including frequency, and this three dimensional representation (time on the x-axis, trial number sorted by orthographic length on the y-axis, and amplitude as color, or equivalently, on the z-axis) shows only some of them. Some hint of this complexity is visible in the trends between trials – the limited coherence of vertical stripes across trials reflects the sorting accord-

ing to orthographic length. Unsorted, the stripes are greatly diminished. Similarly, other patterns emerge when we (simultaneously) sort by other variables, but our ability to represent more dimensions graphically is restricted.

A further complication is the inclusion of continuous predictors. Traditional graphical displays – and statistical techniques – are best suited for categorical predictors, which we can encode with different colors, line types or even subplots. With continuous predictors, this is more difficult (and indeed the reason why we did not include an ERP image of frequency to accompany the model sanity checks). This distorts our perspective as to the true "shape" of ERPs. The sharply defined ERP curves that are familiar from traditional experiments are simply level curves in a multidimensional space, much like lines of equal height on a topographic map. However, even a mountain that appears as a series of coherent rings on a topographic map will tend to be jagged and craggy when viewed in its full multidimensional splendor instead of a series of averages on a two-dimensional piece of paper.

In the graphical presentation of the ERP, we have held only two (morphology and position) of many influences constant and sorted along another dimension (stimulus length), but ran out of visual dimensions to present other influences graphically. However, the mixed-effects models are capable of incorporating many dimensions simultaneously, including continuous dimensions like frequency, which have been traditionally difficult to present as an ERP without resorting to methods like dichotomization (see Smith and Kutas, 2014a, 2014b for a similar but complementary approach using continuous-time regression). In other words, traditional graphical representations of ERPs have difficulty displaying more complex effects and interactions.

One approach is to pick a fixed time-window, freeing up the horizontal axis for something other than time, which fits well with the epoch-based regression approach used here. Displays of the regression at a particular time point are also level curves at a particular time and provide clarity about the shape effect at a particular time, but are less useful for exploring the time course of the ERP. Nonetheless, this perspective allows us to study the modulation of the ERP in a given epoch via more complex influences, such as those that arise in a natural story context. The implications of this perspective – complex influences in a fixed epoch – are discussed more fully below.

Animacy, Case Marking and Word Order

In addition to frequency as a relatively basic, word-level property, we examined the effects of several higher-level cues to sentence interpretation – animacy, case marking and word order – in order to determine whether our methodology is also suited to examining neural activity related to the interpretation of linguistically

expressed events. Psycholinguistic studies using behavioral methods have demonstrated that such cues play an important role in determining real-time sentence interpretation (e.g. with respect to the role of a participant in the event being described; a human is a more likely event instigator, as is an entity that is mentioned early rather than late in a sentence etc.) – and, hence, expectations about upcoming parts of the stimulus (e.g. Bates et al., 1982; MacWhinney et al., 1984). Electrophysiological evidence has added support to this claim, with an increased N400 amplitude for dispreferred yet grammatically correct constructions (for animacy effects in English, Chinese and Tamil Weckerly and Kutas, 1999; Bornkessel et al., 2003; e.g. for accusative-initial sentences in several languages including German, Swedish and Japanese, Schlesewsky et al., 2003; Philipp et al., 2008; Wolff et al., 2008; Bourguignon et al., 2012; Hörberg et al., 2013; Muralikrishnan et al., in press). As a further exploration, we examine the feasibility of measuring these effects in the natural story context.

For the following analyses, we further restricted the trials to full noun phrases occurring as main arguments of verbs that were in the nominative or accusative case (roughly "subjects" and "objects", not including indirect objects). This matches previous work most closely and avoids more difficult cases where the theory is not quite as developed (i.e., what is the role of animacy in prepositional phrases?). In the following, '+' indicates preferred (i.e., animate, initial position, or unambiguous nominative) and '-' indicates dispreferred (i.e., inanimate, non-initial position, unambiguous accusative). For morphology, there is an additional neutral classification for ambiguous case marking.

We begin with a model for these linguistic cues and their interactions with each other, shown in Table 6. For comparison, we include the Wald tests for this simple model in Table 7. From the model summary, we can see a main effect for animacy: animate/preferred is more positive, or in other words, there is a negativity for inanimate/dispreferred. Similarly, we see main effects for both types both types of unambiguous case marking, with a negativity for unambiguous nominative / preferred and a positivity for unambiguous accusative / dispreferred, which at first seems to contradict previous evidence that dispreferred cue forms elicit a negativity. This somewhat surprising result is quickly explained by the interaction between morphology in position, which shows a negativity for the dispreferred initial-accusative word order. The "missing" main effect for (ordinal) position is not surprising for German data, where case and animacy drive the interpretation (cf. MacWhinney et al., 1984) – the role of position is driven more by its interactions than its main effect.

The Wald tests show similar results with the curious exception that position is significant. This is likely a result of the strength of position's interaction with morphology; position is important for the model, the interactions "absorb" some of the effect. However, the Wald tests are *marginal tests*, they test the effect of completely removing a given term – and thus all of its interactions – from the model. With this in mind,

it becomes clear that position achieves significance via its interactions. Since it is problematic to interpret main effects in the presence of interactions anyway, this is not a large problem.

Index and Corpus Frequency: Covariates, not confounds

We also considered more extensive models with the covariates index and corpus frequency. Table 8 shows the results for the model comparison: including index and corpus frequency improves the model fit. The Wald tests for this more extensive model are shown in Table 9.

In the full model, we find main effects for index, corpus frequency, morphology and position. There is no longer an effect for animacy. This can be explained by the reliable correlation between animacy and frequency (in this story, Kendall's $\tau = -0.24$, $p =< 0.001$), and so the variance explained by animacy is absorbed into the frequency term. The interaction between morphology and position is again present. Both morphology and position interact with position individually and in a three-way interaction. There is also a three-way interaction between the linguistic cues (Figure 4). Moreover, morphology and position have a three-way interaction with corpus frequency (Figure 5). Additionally, there are number of higher level interactions between morphology or position, but we avoid interpreting these further than to note that they are compatible with results in the literature.

Word Length

Due to convergence issues, it was not possible to create a maximum model including orthographic length, index, corpus frequency, and all the linguistic cues, but the model with corpus frequency and orthographic length as covariates for the prominence features shows a similar set of effects (Table 10). This again serves as a validity check that the effects for the linguistic cues are not merely the result of confounds with other properties of the stimulus.

Frequency is Dynamic, Redux

We can also examine the interplay between linguistic cues and the two types of frequency in a single model, shown in Table 11. Due to convergence issues, it was not possible to include index or orthographic length in this model, but nonetheless several interesting patterns emerge.

There are main effects for both types of frequency as well as morphology; additionally corpus and relative frequency interact with each other. The interaction between morphology and position is again present as well

as an interaction between animacy and morphology and a three-way interaction between all three features. Interestingly, there appears to be a division in the interactions between linguistic cues and frequency type. Corpus frequency interacts with position, morphology, and with both in a three-way interaction, while relative frequency interacts with animacy and with animacy and morphology and with morphology and position in three-way interactions. There are also higher-order interactions including both frequency types and the prominence features.

Discussion

We have presented a new approach to analyzing electrophysiological data collected in response to a naturalistic auditory stimulus (a natural story). Strikingly, the current results mirror a number of well-established findings from traditional, highly controlled studies. This is somewhat surprising given the large amount of jitter in naturalistic stimuli. The words themselves have different lengths and different phonological and acoustic features; moreover, the phrases have different lengths, which are often longer than in traditional experiments. This leads to the information carried by the acoustic-phonological signal being more broadly and unevenly distributed in time. Yet, we still see clear effects at a fixed latency, which seems to be at odds with traditional notions of ERPs as successive, if occasionally overlapping events, reflecting various (perhaps somewhat parallel) processing stages. In the following, we discuss the implications of our results for the interpretation of ERP responses in cognitive neuroscience research – both in a naturalistic auditory environment and beyond.

Implications for Electrophysiological Research in Cognitive Neuroscience: ERP Components as Ongoing Processes

In cognitive neuroscience research, ERPs are often treated as discrete events. From this perspective, individual components within the electrophysiological signal (e.g. the N200, N400, P300 and P600 to name just a small selection of examples) are interpreted as indexing particular cognitive processes which occur at certain, clearly defined times within the overall time course of processing (see e.g. Friederici, 2011, for a recent review in the language domain). However, ERP data recorded in response to naturalistic, auditory language challenge this traditional view: in contrast to ERPs in studies employing segmented visual presentation (RSVP), components no longer appear as well defined peaks during ongoing auditory stimulation and this applies equally to the early exogenous components and to endogenous components.

Let us first consider the exogenous components. The fact that these no longer appear during continuous auditory stimulation other than at stimulus onset does not mean that the neurocognitive processes indexed by these early components do not take place later in the stimulus, but rather that their form is no longer abrupt enough to be visually distinct from other signals in the EEG. The abruptness of stimulus presentation in RSVP leads to the abruptness of the components, but continuous stimulation, as in a naturalistic paradigm, leads to a continuous modulation of the ERP waveform without the typical peaks of RSVP.

More precisely, the relevant continuity is not that of the stimulus itself, but rather of the information it carries. In RSVP, *all* external information for a given presentation unit is immediately available, although there may be certain latencies involved in processing this information and connecting to other sources of information (e.g. binding together multimodal aspects of conceptual knowledge). Thus, as the information passes through the processing system, it is available in its entirety and there are sharp increases in neural activity corresponding to this flood of new information resulting in sharp peaks. In auditory presentation, the amount of external information is transmitted over time (instead of over space), and thus the clear peaks fall away as the incoming information percolates continuously through the processing system, yielding smaller and temporally less well defined modulations of the ERP. In summary, we propose that the appearance of ERP components as small modulations or large peaks is a result of the relative change in the degree of information processed. In studies employing visual presentation, time-locking to recognition point (e.g. Brink and Hagoort, 2004; Wolff et al., 2008) or employing other similar jitter-controlling measures in auditory presentation, ERPs thus reflect the state of processing *at the climax of (local) information input* and fail to provide information about incrementality below the level of units such as words.

This proposal accords well with a predictive coding-based approach to electrophysiological responses, in which ERP responses such as the mismatch negativity (MMN) reflect both bottom-up adaptation to the stimulus and modulation of top-down predictions / adjustment of an internal model (Friston, 2005; Garrido et al., 2009). Predictive coding posits that the brain constantly attempts to match sensory input sampled from the external world to predictions about the state of the world derived from an internal model, accomplished by means of hierarchically organised forward and inverse models and thought to be implemented by hierarchically organised cortical networks. At the lowest level, predictions are matched against sensory input and any resulting mismatch (prediction error) is propagated back up the hierarchy via feedforward connections (bottom-up adaptation), thereby initiating model updates to minimise prediction errors both at the current level and the level below (top-down modulation). From the predictive coding perspective, the MMN for deviant stimuli within a series of standards reflects an attenuation of the response to the standards rather than the generation of an additional mismatch response to the deviants: stimulus repetition leads

to model adjustment and the minimization of prediction error for subsequent standard presentations and, accordingly, a disappearance of the MMN. An approach along these lines straightforwardly accounts for the apparent discrepancy between ERP responses in traditional and naturalistic paradigms. In naturalistic settings, continuous stimulation in conjunction with rich contextual information leads to increased model update and adaptation, particularly for early sensory aspects of processing, thereby resulting in an attenuation of ERP components. In other words, the prediction errors and resulting model updates are necessarily more pronounced in isolated stimuli than in stimuli encountered in a naturalistic context.

Continuous Components, Continuous Processing, and Growing Representations

We propose that this continuous, subsymbolic incrementality can be extended to also account for a broader range of stimulus-locked components such as the N200 and N400. Specifically, we suggest that the account of the MMN outlined above can be straightforwardly extended to these components in the sense that they reflect similar stimulus-related processing mechanisms as the MMN (bottom-up adaptation and top-down modulation), but at different levels of the processing hierarchy (for a somewhat similar view, see Pulvermüller et al., 2009). This view is not entirely new: early research concerning the N400 examined the possibility that it was a member of the N200 family (Kutas and Federmeier, 2011), much like the long-standing debate about whether the P600 belongs to the P300 family (Osterhout et al., 1996; e.g. Gunter et al., 1997; Coulson et al., 1998; Sassenhagen et al., 2014). The notion of continuous processing presented here hints at a coherent account for such component families, related to their temporal resolution. Following Giraud and Poeppel (2012)'s suggestion that the frequency bands in cortical oscillations track the time resolution of hierarchical structure in speech processing, we can consider similar ERP components with different time-scales as tracking the time resolution of different stimulus features (Dogil et al., 2004; see also Bornkessel and Schlesewsky, 2006; Roehm et al., 2007). In this view, the MMN and N200 are similar to the N400 but react to more basic features of the stimulus at a lower latency because they reflect a similar neural process earlier in the processing hierarchy. This leads to a higher temporal resolution but a smaller analysis time window, in accordance with the frequency of the oscillation under consideration. This perspective accounts for the apparent paradox of MMN effects for manipulations more typical to N400 experiments (cf. "ultrafast processing" in recent studies such as Pulvermüller et al., 2001; MacGregor et al., 2012; Shtyrov et al., 2014); or other fast recognitions of large-scale stimulus change (e.g. category error in Dikker et al., 2009) as reflecting predictions that are exceedingly precise and can thus be falsified quickly. Moreover, similar mechanisms operating at different scales is compatible with the recent proposal that the mechanisms for human language processing arise from a difference from nonhuman primates in quantity rather than quality (Bornkessel-Schlesewsky et al., in press)

and is compatible with the account that the neural aspects of early language acquisition follow increasing time scales (Friederici, 2005). More complex processing arises as fundamental processing mechanisms are repeated and expanded across multiple time scales.

Conclusion

We have demonstrated the feasibility of studying the electrophysiology of speech processing with a naturalistic stimulus. The replication of well-known effects served as a proof of concept, while initial exploration of the more complex interactions possible in a rich context suggested new courses of study. Surprisingly, we found robust manipulations at a fixed latency from stimulus onset in spite of the extreme jitter from differences in word and phrase length. This suggests that ERP responses should be viewed as continuous modulations and not discrete, yet overlapping waveforms.

References

Baayen R, Davidson D, Bates D (2008) Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59:390–412.

Barr DJ (2013) Random effects structure for testing interactions in linear mixed-effects models. Frontiers in Psychology 4.

Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language 68:255–278.

Bates E, McNew S, MacWhinney B, Devescovi A, Smith S (1982) Functional constraints on sentence processing: A cross-linguistic study. Cognition 11:245–299.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS (2009) Generalized linear mixed models: A practical guide for ecology and evolution. Trends Ecol Evol 24:127–135.

Bornkessel I, Schlesewsky M (2006) The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. Psychological Review 113:787–821.

Bornkessel I, Schlesewsky M, Friederici AD (2003) Contextual information modulates initial processes of syntactic integration: The role of inter- vs. intra-sentential predictions. Journal of Experimental Psychology: Learning, Memory and Cognition 29:269–298.

Bornkessel-Schlesewsky I, Schlesewsky M (2009) The role of prominence information in the real-time comprehension of transitive constructions: A cross-linguistic approach. Language and Linguistics Compass 3:19–58.

Bornkessel-Schlesewsky I, Schlesewsky M, Small SL, Rauschecker JP (in press) Neurobiological roots of language in primate audition: Common computational properties. Trends in Cognitive Sciences.

Bourguignon N, Drury JE, Valois D, Steinhauer K (2012) Decomposing animacy reversals between agents and experiencers: An ERP study. Neurobiology of language 2010 neurobiology of language conference 122:179–189.

Brennan J, Nir Y, Hasson U, Malach R, Heeger DJ, Pylkkänen L (2012) Syntactic structure building in the anterior temporal lobe during natural story listening. Brain and Language 120:163–173.

Brink D van der, Hagoort P (2004) The influence of semantic and syntactic context constraints on lexical selection and integration in spoken-word comprehension as revealed by ERPs. Journal of Cognitive Neuroscience 16:1068–1084.

Conroy BR, Singer BD, Guntupalli JS, Ramadge PJ, Haxby JV (2013) Inter-subject alignment of human cortical anatomy using functional connectivity. NeuroImage 81:400–411.

Coulson S, King JW, Kutas M (1998) Expect the unexpected: Event-related brain response to morphosyntactic violations. Language and Cognitive Processes 13:21–58.

Cummings A, Čeponienė R, Koyama A, Saygin A, Townsend J, Dick F (2006) Auditory semantic networks for words and natural sounds. Brain Research 1115:92–107.

Dambacher M, Kliegl R, Hofmann M, Jacobs AM (2006) Frequency and predictability effects on event-related potentials during reading. Brain Res 1084:89–103.

Dikker S, Rabagliati H, Pylkkänen L (2009) Sensitivity to syntax in visual cortex. Cognition 110:293–321.

Dogil G, Frese I, Haider H, Röhm D, Wokurek W (2004) Where and how does grammatically geared processing take place – and why is broca's area often involved. a coordinated fMRI/ERBP study of language processing. Language and motorIntegration 89:337–345.

Frank SL, Otten LJ, Galli G, Vigliocco G (2015) The ERP response to the amount of information conveyed by words in sentences. Brain and Language 140:1–111.

Friederici AD (2005) Neurophysiological markers of early language acquisition: From syllables to sentences. Trends in Cognitive Sciences 9:481–488.

Friederici AD (2011) The brain basis of language processing: From structure to function. Physiol Rev 91:1357–1392.

Frisch S, Schlesewsky M (2001) The N400 reflects problems of thematic hierarchizing. NeuroReport 12.

Friston K (2005) A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences 360:815–836.

Garrido MI, Kilner JM, Stephan KE, Friston KJ (2009) The mismatch negativity: A review of underlying mechanisms. Clinical Neurophysiology 120:453–463.

Giraud A-L, Poeppel D (2012) Cortical oscillations and speech processing: Emerging computational principles and operations. Nature Neuroscience 15:511–517.

Gunter TC, Stowe LA, Mulder G (1997) When syntax meets semantics. Psychophysiology 34:660–676.

Hagoort P (2007) The memory, unification and control (MUC) model of language. In: Automaticity and control in language processing. Psychology Press.

Hale J (2001) A probabilistic earley parser as a psycholinguistic model. In: Proceedings of the second meeting of the north american chapter of the association for computational linguistics on language technologies, pp 1–8 NAACL '01. Stroudsburg, PA, USA: Association for Computational Linguistics.

Hanke M, Baumgartner FJ, Ibe P, Kaule FR, Pollmann S, Speck O, Zinke W, Stadler J (2014) A high-resolution 7-tesla fMRI dataset from complex natural stimulation with an audio movie. Scientific Data 1.

Hasson U, Honey CJ (2012) Future trends in neuroimaging: Neural processes as expressed within real-life contexts. Neuroimage 62:1272–1278.

Hasson U, Malach R, Heeger DJ (2010) Reliability of cortical activity during natural stimulation. Trends in Cognitive Sciences 14:40–48.

Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. Science 303:1634–1640.

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N (2008) A hierarchy of temporal receptive windows in human cortex. The Journal of Neuroscience 28:2539–2550.

Hörberg T, Koptjevskaja-Tamm M, Kallioinen P (2013) The neurophysiological correlate to grammatical function reanalysis in swedish. Language and Cognitive Processes 28:388–416.

Hutzler F, Braun M, Võ ML-H, Engl V, Hofmann M, Dambacher M, Leder H, Jacobs AM (2007) Welcome to the real world: Validating fixation-related brain potentials for ecologically valid settings. Brain Res 1172:124–129.

Jung T-P, Makeig S, Westerfield M, Townsend J, Courchesne E, Sejnowski TJ (2001) Analysis and vIsualization of single-trial event-related potentials. Human Brain Mapping 14:166–185.

Kretzschmar F, Pleimling D, Hosemann J, Füssel S, Bornkessel-Schlesewsky I, Schlesewsky M (2013) Subjective impressions do not mirror online reading effort: Concurrent EEG-eyetracking evidence from the reading of books and digital media. PLoS One 8:e56178.

Kutas M, Federmeier KD (2000) Electrophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences 4:463–470.

Kutas M, Federmeier KD (2011) Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual review of psychology 62:621–647.

Lau EF, Phillips C, Poeppel D (2008) A cortical network for semantics: (De)constructing the N400. Nat Rev Neurosci 9:920–933.

Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. The Journal of Neuroscience 31:2906–2915.

Levy R (2008) Expectation-based syntactic comprehension. Cognition 106:1126–1177.

Lotze N, Tune S, Schlesewsky M, Bornkessel-Schlesewsky I (2011) Meaningful physical changes mediate lexical-semantic integration: Top-down and form-based bottom-up information sources interact in the n400. Neuropsychologia 49:3573–3582.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. Psychological Methods 7:19–40.

MacGregor LJ, Pulvermüller F, Casteren M van, Shtyrov Y (2012) Ultra-rapid access to words in the brain. Nature Communications 3:711.

MacWhinney B, Bates E, Kliegl R (1984) Cue validity and sentence interpretation in English, German and Italian. Journal of Verbal Learning and Verbal Behavior 23:127–150.

Muralikrishnan R, Schlesewsky M, Bornkessel-Schlesewsky I (in press) Animacy-based predictions in language comprehension are robust: Contextual cues modulate but do not nullify them. Brain Research.

Osterhout L, Mckinnon R, Bersick M, Corey V (1996) On the language specificity of the brain response

to syntactic anomalies: Is the syntactic positive shift a member of the p300 family. Journal of Cognitive Neuroscience 8:507–526.

Palmer JA, Kreutz-Delgado K, Rao BD, Makeig S (2007) Modeling and estimation of dependent subspaces with non-radially symmetric and skewed densities. In: Proceedings of the 7th international symposium on independent component analysis (Davies ME, James CJ, Abdallah SA, Plumbley MD, eds), pp 97–104 Lecture notes in computer science. Springer Berlin Heidelberg.

Philipp M, Bornkessel-Schlesewsky I, Bisang W, Schlesewsky M (2008) The role of animacy in the real time comprehension of mandarin chinese: Evidence from auditory event-related brain potentials. Brain and Language 105:112–133.

Pinheiro J, Bates D (2000) Mixed-effects models in S and S-PLUS. Springer New York.

Pulvermüller F, Kujala T, Shtyrov Y, Simola J, Tiitinen H, Alku P, Alho K, Martinkauppi S, Ilmoniemi RJ, Näätänen R (2001) Memory traces for words as revealed by the mismatch negativity. NeuroImage 14:607–616.

Pulvermüller F, Shtyrov Y, Hauk O (2009) Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. Brain and Language 110:81–94.

Roehm D, Bornkessel-Schlesewsky I, Schlesewsky M (2007) The internal structure of the N400: Frequency characteristics of a language related ERP component. Chaos and Complexity Letters 2:365–395.

Roehm D, Sorace A, Bornkessel-Schlesewsky I (2013) Processing flexible form-to-meaning mappings: Evidence for enriched composition as opposed to indeterminacy. Language and Cognitive Processes 28:1244–1274.

Sassenhagen J, Schlesewsky M, Bornkessel-Schlesewsky I (2014) The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. Brain and Language 137:29–39.

Schlesewsky M, Bornkessel I, Frisch S (2003) The neurophysiological basis of word order variations in German. Understanding language 86:116–128.

Shtyrov Y, Butorina A, Nikolaeva A, Stroganova T (2014) Automatic ultrarapid activation and inhibition of cortical motor systems in spoken word comprehension. Proc Natl Acad Sci U S A 111:E1918–E1923.

Skipper JI, Goldin-Meadow S, Nusbaum HC, Small SL (2009) Gestures orchestrate brain networks for language understanding. Current Biology 19:661–667.

532 Smith NJ, Kutas M (2014a) Regression-based estimation of ERP waveforms: I. the rERP framework. Psy-
533 chophysiology.

534 Smith NJ, Kutas M (2014b) Regression-based estimation of ERP waveforms: II. nonlinear effects, overlap
535 correction, and practical considerations. Psychophysiology.

536 Van Petten C, Kutas M (1990) Interactions between sentence context and word frequency in event-related
537 brain potentials. Memory and Cognition 4:380–393.

538 Van Petten C, Kutas M, Kluender R, Mitchiner M, McIsaac H (1991) Fractionating the word repetition
539 effect with event-related potentials. Journal of Cognitive Neuroscience 3.

540 Weckerly J, Kutas M (1999) An electrophysiological analysis of animacy effects in the processing of object
541 relative sentences. Psychophysiology 36:559–570.

542 Whitney C, Huber W, Klann J, Weis S, Krach S, Kircher T (2009) Neural correlates of narrative shifts
543 during auditory story comprehension. NeuroImage 47:360–366.

544 Winkler I, Haufe S, Tangermann M (2011) Automatic classification of artifactual ICA-components for artifact
545 removal in EEG signals. Behavioral and Brain Functions 7:30.

546 Wolff S, Schlesewsky M, Hirotani M, Bornkessel-Schlesewsky I (2008) The neural mechanisms of word order
547 processing revisited: Electrophysiological evidence from Japanese. Brain and Language 107:133–157.

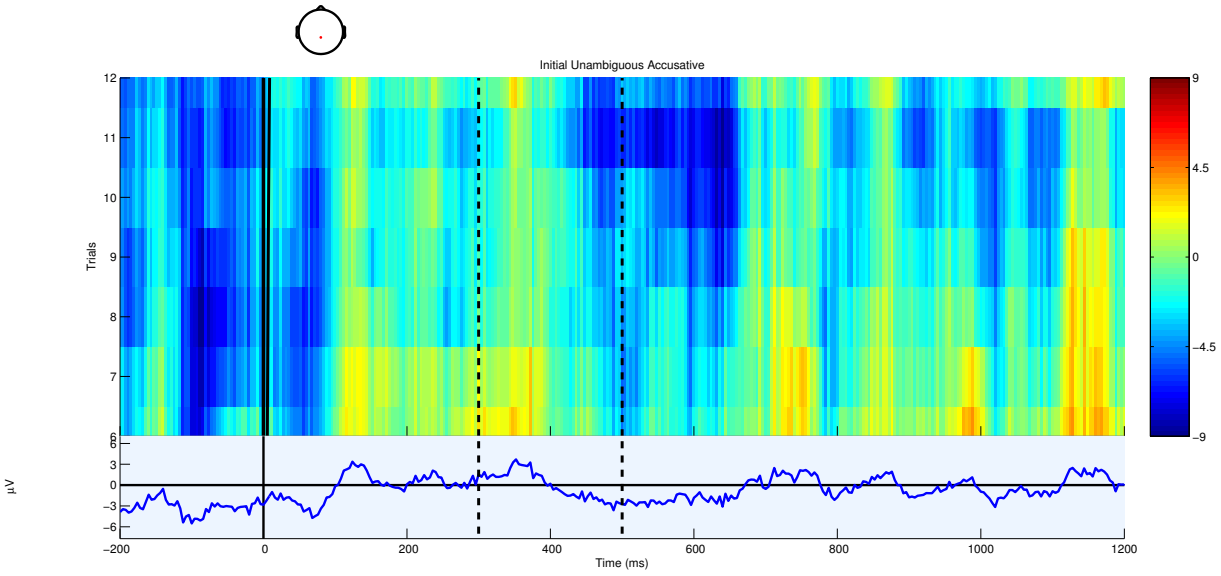


Figure 1: Single trial and average ERPs from electrode CPz from a single subject for unambiguous accusatives placed before a nominative. In the upper part, single trials are displayed stacked and sorted from top to bottom in decreasing orthographic length as a weak proxy for acoustic length, while the lower part displays the average ERP. Amplitude is given by color in the upper part and by the y-axis in the lower part. The dashed vertical lines indicate the boundaries of the N400 time window, 300 and 500ms post stimulus onset.

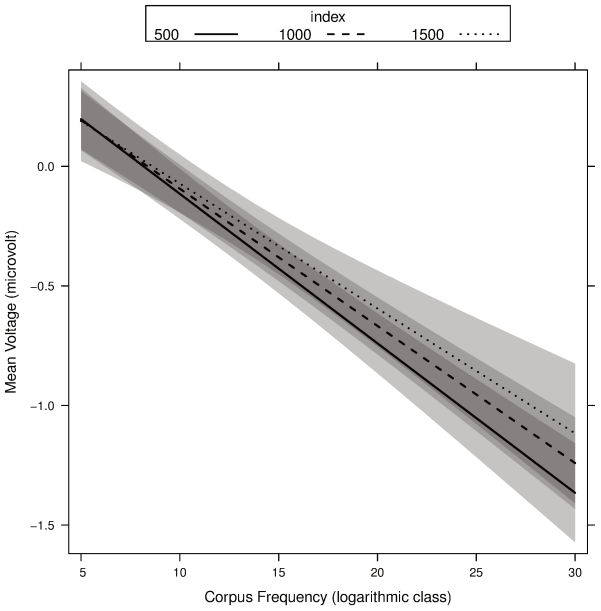


Figure 2: Plot of effects for corpus frequency interacting with index (ordinal position in the story). Shaded areas indicate 95% confidence intervals. Index is divided into tertiles and plotted in an overlap to make the lack of interaction more prominent. There is an increasing negativity with decreasing frequency (higher logarithmic class), which is unaffected by position in the story.

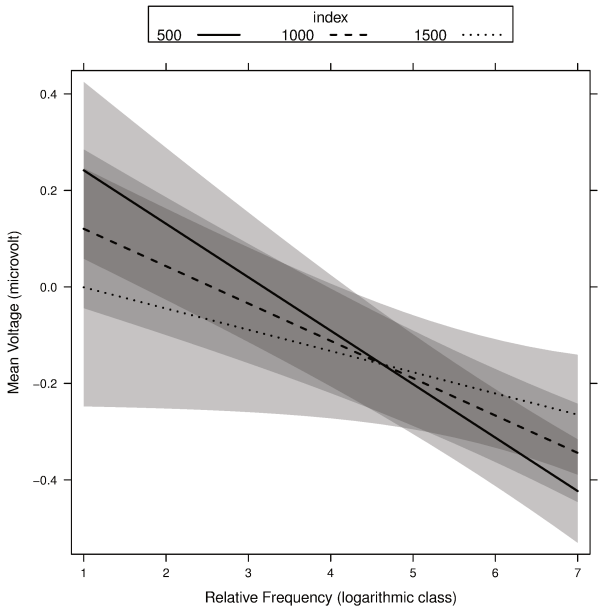


Figure 3: Plot of effects for relative frequency interacting with index. Index is divided into tertiles and plotted in an overlap to make the interaction more prominent.

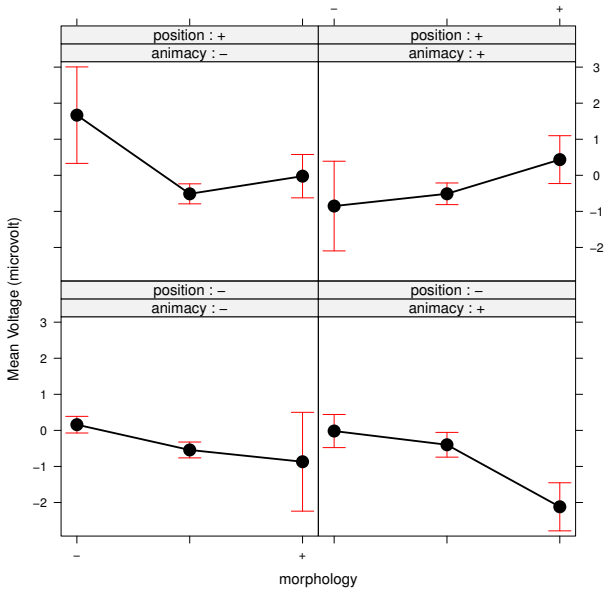


Figure 4: Interaction of animacy, morphology and position from the full prominence model with index and frequency class.

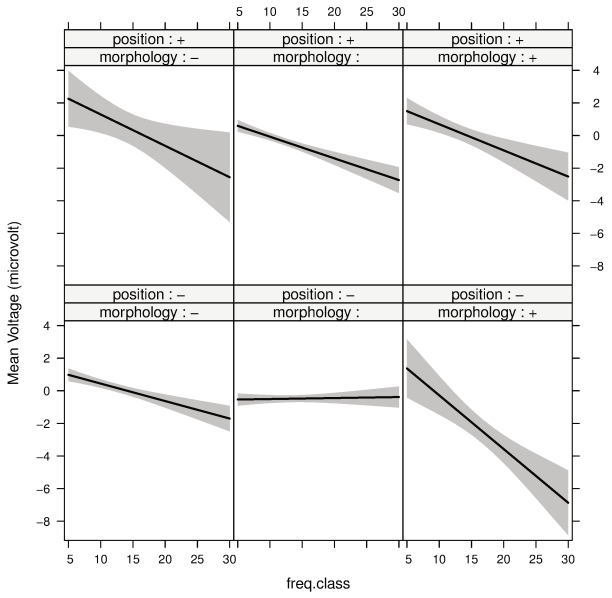


Figure 5: Interaction of position, morphology and corpus frequency from the full prominence model with index and frequency class.

Table 1: Summary of model fit for (corpus) frequency class in the time window 300–500ms from stimulus onset using all content words.

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance	
2021954	2021996	-1010973	2021946	
Scaled residuals:				
Min	1Q	Median	3Q	Max
-33.06	-0.53	0	0.53	38.43
Random effects:				
Groups	Name	Variance	Std.Dev	
subj	(Intercept)	0.10	0.31	
Residual		130.01	11.40	
Number of obs: 262392, groups: subj, 52.				
Fixed effects:				
	Estimate	Std. Error	t value	
(Intercept)	0.5	0.075	6.6	
freq.class	−0.059	0.0045	−13	

Table 2: Comparison of models for (corpus) frequency class with and without index (ordinal position). Including index does not significantly improve model fit.

	Df	AIC	BIC	logLik	deviance	χ^2	χ^2 Df	Pr($>\chi^2$)
m.freq	4	2021954	2021995	-1010973	2021946			
m.freq.index	6	2021954	2022017	-1010971	2021942	3.74	2	0.154

Table 3: Comparison of models for relative frequency class with and without index (ordinal position). Including index significantly improves model fit.

	Df	AIC	BIC	logLik	deviance	χ^2	χ^2 Df	Pr($>\chi^2$)
m.rel	4	2022083	2022125	-1011037	2022075			
m.rel.index	6	2022078	2022141	-1011033	2022066	8.61	2	0.0135

Table 4: Summary of model fit for relative frequency class and index (ordinal position) in the time window 300–500ms from stimulus onset using all content words. The interaction term yields a reliable estime, while the main effect for index is not quite reliable.

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance	
2022079	2022141	-1011033	2022067	
Scaled residuals:				
Min	1Q	Median	3Q	Max
-33.07	-0.53	0	0.53	38.4
Random effects:				
Groups	Name	Variance	Std.Dev	
subj	(Intercept)	0.10	0.31	
Residual		130.07	11.40	
Number of obs: 262392, groups: subj, 52.				
Fixed effects:				
	Estimate	Std. Error	t value	
(Intercept)	0.51	0.17	3	
index	−0.00031	0.00017	−1.8	
rel.freq.class	−0.14	0.027	−5.3	
index:rel.freq.class	6.7e−05	2.8e−05	2.4	

Table 5: Comparison of best models for corpus and relative frequency. Both models yield similar fits.

	Df	AIC	BIC	logLik
m.freq	4	2021954	2021995	-1010973
m.rel.index	6	2022078	2022141	-1011033

Table 6: Summary of model fit for linguistic cues (animacy, morphology, linear position) known to elicit N400-like effects. Dependent variable are single-trial means in the time window 300–500ms from stimulus onset using only subjects and (direct) objects. '+' indicates preferred (i.e., animate, initial position, or unambiguous nominative) and '-' indicates dispreferred (i.e., inanimate, non-initial position, unambiguous accusative). Morphology also has an additional 'neutral' level for ambiguous case marking.

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance	
530425	530553	-265199	530397	
Scaled residuals:				
Min	1Q	Median	3Q	Max
-11.87	-0.54	0	0.54	12.94
Random effects:				
Groups	Name	Variance	Std.Dev	
subj	(Intercept)	0.20	0.44	
Residual		125.99	11.22	
Number of obs: 69108, groups: subj, 52.				
Fixed effects:				
	Estimate	Std. Error	t value	
(Intercept)	−0.59	0.11	−5.4	
animacy+	0.34	0.17	2	
morphology-	0.6	0.13	4.6	
morphology+	−0.93	0.33	−2.8	
position+	0.17	0.15	1.1	
animacy+:morphology-	−0.026	0.27	−0.096	
animacy+:morphology+	−0.8	0.48	−1.7	
animacy+:position+	−0.11	0.23	−0.48	
morphology-:position+	−0.2	0.48	−0.41	
morphology+:position+	1.4	0.43	3.4	
animacy+:morphology-:position+	−0.65	0.65	−1	
animacy+:morphology+:position+	0.9	0.61	1.5	

Table 7: Type-II Wald tests for the model presented in Table 6

	χ^2	Df	$\Pr(>\chi^2)$	
animacy	6.46	1	0.011	*
morphology	33.69	2	< 0.001	***
position	9.10	1	0.00255	**
animacy:morphology	0.66	2	0.721	
animacy:position	0.11	1	0.74	
morphology:position	47.22	2	< 0.001	***
animacy:morphology:position	3.71	2	0.157	

Table 8: Model comparison for linguistic-cue based models with index and (corpus) frequency

	Df	AIC	BIC	logLik	deviance	χ^2	χ^2 Df	Pr($>\chi^2$)
prom	14	530425	530553	-265198	530397			
prom.index	26	530392	530630	-265170	530340	56.43	12	< 0.001
prom.freq.index	50	530281	530738	-265090	530181	159.18	24	< 0.001

Table 9: Type-II Wald tests for the model combining index, (corpus) frequency and linguistic cues.

	χ^2	Df	Pr($>\chi^2$)	
index	6.84	1	0.00892	**
freq.class	55.06	1	< 0.001	***
animacy	0.01	1	0.919	
morphology	39.32	2	< 0.001	***
position	4.07	1	0.0438	*
index:freq.class	0.80	1	0.371	
index:animacy	1.96	1	0.161	
freq.class:animacy	1.19	1	0.276	
index:morphology	2.75	2	0.253	
freq.class:morphology	11.03	2	0.00404	**
animacy:morphology	0.76	2	0.685	
index:position	8.85	1	0.00293	**
freq.class:position	19.92	1	< 0.001	***
animacy:position	0.13	1	0.722	
morphology:position	23.41	2	< 0.001	***
index:freq.class:animacy	0.50	1	0.481	
index:freq.class:morphology	5.08	2	0.0787	.
index:animacy:morphology	7.62	2	0.0221	*
freq.class:animacy:morphology	5.48	2	0.0645	.
index:freq.class:position	4.47	1	0.0345	*
index:animacy:position	6.01	1	0.0142	*
freq.class:animacy:position	1.29	1	0.256	
index:morphology:position	1.94	2	0.378	
freq.class:morphology:position	11.79	2	0.00275	**
animacy:morphology:position	6.93	2	0.0313	*
index:freq.class:animacy:morphology	16.73	2	< 0.001	***
index:freq.class:animacy:position	2.47	1	0.116	
index:freq.class:morphology:position	0.64	2	0.725	
index:animacy:morphology:position	0.76	2	0.685	
freq.class:animacy:morphology:position	9.50	2	0.00863	**
index:freq.class:animacy:morphology:position	1.15	2	0.563	

Table 10: Type-II Wald tests for the model combining orthographic length, (corpus) frequency and linguistic cues.

	χ^2	Df	$\Pr(>\chi^2)$	
ortho.len	3.14	1	0.0765	.
freq.class	9.86	1	0.00169	**
animacy	0.17	1	0.68	
morphology	44.45	2	< 0.001	***
position	8.39	1	0.00377	**
ortho.len:freq.class	0.06	1	0.81	
ortho.len:animacy	1.49	1	0.222	
freq.class:animacy	0.30	1	0.584	
ortho.len:morphology	6.04	2	0.0489	*
freq.class:morphology	10.70	2	0.00474	**
animacy:morphology	2.86	2	0.239	
ortho.len:position	2.07	1	0.15	
freq.class:position	6.03	1	0.0141	*
animacy:position	2.36	1	0.125	
morphology:position	39.94	2	< 0.001	***
ortho.len:freq.class:animacy	15.97	1	< 0.001	***
ortho.len:freq.class:morphology	6.94	2	0.0311	*
ortho.len:animacy:morphology	4.70	2	0.0951	.
freq.class:animacy:morphology	3.69	2	0.158	
ortho.len:freq.class:position	10.39	1	0.00127	**
ortho.len:animacy:position	6.52	1	0.0107	*
freq.class:animacy:position	1.32	1	0.251	
ortho.len:morphology:position	5.31	2	0.0702	.
freq.class:morphology:position	2.96	2	0.228	
animacy:morphology:position	2.60	2	0.272	
ortho.len:freq.class:animacy:morphology	49.04	2	< 0.001	***
ortho.len:freq.class:animacy:position	34.32	1	< 0.001	***
ortho.len:freq.class:morphology:position	0.54	2	0.764	
ortho.len:animacy:morphology:position	5.94	2	0.0514	.
freq.class:animacy:morphology:position	8.56	2	0.0138	*
ortho.len:freq.class:animacy:morphology:position	8.47	2	0.0145	*

Table 11: Type-II Wald tests for the model combining linguistic cues with both corpus and relative frequency.

	χ^2	Df	Pr(>χ^2)	
rel.freq.class	3.86	1	0.0495	*
freq.class	40.65	1	< 0.001	***
animacy	0.15	1	0.702	
morphology	37.05	2	< 0.001	***
position	1.92	1	0.166	
rel.freq.class:freq.class	9.31	1	0.00228	**
rel.freq.class:animacy	10.52	1	0.00118	**
freq.class:animacy	0.00	1	0.998	
rel.freq.class:morphology	2.43	2	0.296	
freq.class:morphology	19.54	2	< 0.001	***
animacy:morphology	7.67	2	0.0217	*
rel.freq.class:position	0.26	1	0.607	
freq.class:position	10.12	1	0.00146	**
animacy:position	0.37	1	0.541	
morphology:position	31.17	2	< 0.001	***
rel.freq.class:freq.class:animacy	13.27	1	< 0.001	***
rel.freq.class:freq.class:morphology	13.48	2	0.00118	**
rel.freq.class:animacy:morphology	24.83	2	< 0.001	***
freq.class:animacy:morphology	4.68	2	0.0965	.
rel.freq.class:freq.class:position	0.16	1	0.688	
rel.freq.class:animacy:position	0.03	1	0.864	
freq.class:animacy:position	0.39	1	0.534	
rel.freq.class:morphology:position	17.97	2	< 0.001	***
freq.class:morphology:position	10.82	2	0.00447	**
animacy:morphology:position	8.38	2	0.0151	*
rel.freq.class:freq.class:animacy:morphology	13.64	2	0.00109	**
rel.freq.class:freq.class:animacy:position	25.67	1	< 0.001	***
rel.freq.class:freq.class:morphology:position	2.95	2	0.229	
rel.freq.class:animacy:morphology:position	3.85	2	0.146	
freq.class:animacy:morphology:position	13.76	2	0.00103	**

6. The New Old Thing: Memory, Models, and Prediction

> Whereas Newton could say, "If I have seen a little farther than others, it is because I have stood on the shoulders of giants," I am forced to say, "Today we stand on each other's feet."
>
> Richard Hamming

The results from Chapter 5 suggest a new perspective on the nature of ERP components as indices of continuous processes modulated by information flow, which goes against traditional notions of ERPs as successive, if overlapping events (cf. Friederici 2002, 2011). As we suggested in Alday, Schlesewsky, and Bornkessel-Schlesewsky (submitted), our account is largely compatible with the notion of predictive coding (Friston 2005, 2009). In the following, we use this perspective to reformulate a recent suggestion by Bornkessel-Schlesewsky and Schlesewsky (in press) and thus provide a suggestion of the types of computational problems (cf. Marr's levels of description in Chapter 1) indexed by language-related ERP components.

6.1. A Neurocomputational Proposal

Although these suggestions are to a limited extent informed by the neuroanatomic principles suggested in recent formulations of the eADM (Bornkessel-Schlesewsky and Schlesewsky 2013, in press) and a mathematical formalism believed to be neurobiologically plausible (Friston 2005, 2009), they are formulated in computational and cognitive terms and should not be taken as having a one-to-one mapping to brain structures. Rather, the main goal is provide another perspective and thus insight on the types of computations which we have examined throughout this dissertation.

In the sense of a Lakatosian research programme (see Chapter 1), we suggest as a first proposal the following (semi-)hard core:

6.1.1. Supporting Assumptions and Hypotheses

Hypothesis 1 (Continuity of processing in a cascading architecture)**.** *Neurocomputation is a continuous process. Information is processed as soon as it is available; moreover, processed informa-*

tion is immediately passed along the processing pipeline in a cascade. (cf. Alday, Schlesewsky, and Bornkessel-Schlesewsky submitted; Bornkessel-Schlesewsky and Schlesewsky 2009)

In other words, neurocomputation is incremental in the most extreme sense. This proposal is similar to the cascading architecture proposed by Bornkessel-Schlesewsky and Schlesewsky (2009) and has previously been described as non-strict seriality.

Corollary (Hierarchy-time-scale correspondence)**.** *The canonical latency of an ERP component reflects not only its place in the processing stream but also its temporal resolution and thus the time-scale of information and level of hierarchical organization it operates upon. (cf. Alday, Schlesewsky, and Bornkessel-Schlesewsky submitted; Bornkessel-Schlesewsky and Schlesewsky 2013, in press)*

For example, the N200 and N400 reflect similar processes but at the time scales associated with increasingly complex information, e.g. individual phonemes and individual words (cf. Bornkessel-Schlesewsky and Schlesewsky 2013; Alday, Schlesewsky, and Bornkessel-Schlesewsky submitted), and a similar proposal has been suggested for the P3b and P600 (cf. Coulson, King, and Kutas 1998b,a; Sassenhagen, Schlesewsky, and Bornkessel-Schlesewsky 2014).[1] This correspondence is not strict, due to the two-way flow of information in the processing streams (cf. Bornkessel-Schlesewsky and Schlesewsky 2013; see also Friston 2005, and Section 6.1.3, below). Moreover, a single information unit at a particular time scale may have implications that are first problematic for information processing at larger time scales.

In Alday, Schlesewsky, and Bornkessel-Schlesewsky (submitted), we suggested that this correspondence is compatible with recent proposals that the discrete time scales seen in language processing (e.g. phoneme, prosodic word, prosodic sentence) are deeply tied to the division of oscillatory rhythms in the brain into discrete bands (cf. Giraud and Poeppel 2012; Bornkessel-Schlesewsky and Schlesewsky in press; see also Bornkessel and Schlesewsky 2006, for related considerations on the scale of language processing). As such, it may be the case that the apparent division of processing into discrete stages as evidenced by rank-ordering especially for early components, may be an epiphenomenon emerging from the interaction of these more fundamental aspects of the processing architecture.

[1] There is one critical difference between the proposal here and that of Coulson, Sassenhagen and their respective colleagues, namely that we are not suggesting that the P600 is a special instance of the P3 (the so-called identity hypothesis). Rather, we are claiming that positivities reflect a "family" of computationally similar yet temporally and hierarchically distinct processes. As such, the P600 is not a delayed P3b, but rather a distinct component with distinct neural generators yet similar properties. Indeed, the whole notion of *components* is a term of convenience reflecting clusters of emergent phenomena with similar computational and, as suggested by Coulson, Sassenhagen and their respectives colleagues, underlying neurobiological principles. Similarly, it is equally problematic to speak of "the" N400, both generally and in light of several studies demonstrating distinct groups of N400-like components (Roehm et al. 2004; Kretzschmar, Bornkessel-Schlesewsky, and Schlesewsky 2009; Dröge, Schlesewsky, and Bornkessel-Schlesewsky 2012; Van Petten and Luka 2012; Knoeferle, Urbach, and Kutas 2014). The difference in nomenclature emphasizes our perspective that scalp ERP patterns likely reflect a dynamic mixture of neural generators (cf. Bornkessel-Schlesewsky and Schlesewsky in press).

6.1.2. Central Proposal

Combining these assumptions with characterizations of the major components from the literature, we arrive at the following parsimonious account:

Hypothesis 2 (Neurocomputational basis of cognitive ERP components)**.** *The endogeneous ERP components related to language can be broadly divided into two computational categories along the line of their polarity.* Negativities *are indicative of neurocomputational processes related to phenomena which are best termed "model adaptation" or "representation activation".* Positivities *are indicative of neurocomputational processes related to phenomena which in traditional cognitive-psychological models would be termed "evaluation" or "decision making". (cf. Bornkessel-Schlesewsky and Schlesewsky in press; Brouwer, Fitz, and Hoeks 2012)*

Predictive Coding

Especially the negativity-related portion of this proposal is perhaps best understood in light of predictive coding (Friston 2005), and as such, we begin with a very brief summary.

Predictive coding posits that basic neural computation implemented by the brain is an expectation-maximization approach to an empirical and hierarchical Bayesian model (Friston 2005). In particular, "the" model is actually a collection of models such that constituent submodels (in Friston's account, cortical areas) have directed, i.e. hierarchical, connections to other models, roughly corresponding to the notion of a *partial order*.[2] This model is generative and projects its predictions backwards, i.e. from top to bottom, while errors in predictions are projected forwards, i.e. from bottom to top. At the lowest level, predictions are compared against sensory input and the resulting error is propagated upwards, which has a two-fold effect: (1) it generates additional predictions, or equivalently, adapts the current set of predictions, which are then propagated back to the source of the error, i.e. the source of the mismatch, and (2) it subserves an additional comparison, which then generate additional errors, resulting in a cyclic, almost co-recursive series of updates across the entire model hierarchy.

Following this perspective, Friston and colleagues (Friston 2005; Garrido et al. 2009) explain the mismatch negativity (MMN; an early ERP component which, amongst other conditions, arises for "mismatched" tones during unattended listening to a series of matched tones). In

[2] An ordering in mathematical terms is a relation such as "less than" or "contains". Formally, we call a relation $\subseteq$ on a set S a *partial order* if for all $x, y \in S$, the following hold:

1. (Reflexivity) $a \subseteq a$
2. (Antisymmetry) If $a \subseteq b$ and $b \subseteq a$, then $a = b$.
3. (Transitivity) If $a \subseteq b$ and $b \subseteq c$, then $a \subseteq c$.

This corresponds to the usual intuition of operations like "less than [or equal to]", "contains", "greater than [or equal to]", but does not require that an ordering exist between any two elements. If we add the restriction that the order is defined for every pair of elements, i.e. $a \subseteq b$ or $b \subseteq a$ for all $x, y \in S$, then we have a *total order*.

particular, the MMN reflects the large error for a deviant tone following a period of continuous model refinement on the standard tones. Thus, the MMN reflects less the absolute deviance of the mismatched tone and more the contrast to the extremely restricted and hence extremely predictable standard tones, and as such is perhaps better characterized as an attenuation of the error signal for the other tones.

In Alday, Schlesewsky, and Bornkessel-Schlesewsky (submitted), we applied these principles to explain the absence of early perceptual components in auditory studies as well as the continuous modulation of the ERP in a naturalistic context as a series of "drifts" instead of a series of "peaks".

Negativities

Following and extending this predictive-coding perspective, we claim that negativities reflect "model adaptation" in the sense of indexing the amount of error in previous predictions, or equivalently, indexing the extent of the mismatch between top-down and bottom-up influences (cf. Lau, Phillips, and Poeppel 2008; Lotze et al. 2011). Although our account shares some similarities to the model proposed by Lau, Phillips, and Poeppel (2008), we do not assume static *a priori* or symbolic levels, but rather a dynamic, interactive interplay of subsymbolic influences.

In this sense, we can also say that negativities reflect "representation activation" as we understand that representations *are* models. Default representations, e.g. less dynamic entities such as the semantic field of a word, reflect in the Bayesian account a default, possibly minimally informative prior (cf. "Frequency is Dynamic" in Alday, Schlesewsky, and Bornkessel-Schlesewsky submitted), which can be dynamically adapted and modified.[3] In particular, activated representations are continuously updated and adjusted to fit the current (stimulatory) context.

In this sense, this account is similar to activation-based accounts of working memory (cf. Jonides et al. 2008),[4] and indeed negativities have been shown to correlate with visual working memory (Vogel and Machizawa 2004; Luck and Vogel 2013). Activation-based accounts of working memory have also been applied to language processing as skilled retrieval modulated by cue interference (Lewis and Vasishth 2005; Lewis, Vasishth, and Van Dyke 2006). However, as Alday, Schlesewsky, and Bornkessel-Schlesewsky (2014) (see Chapter 3) showed, the lack of feature weighting in such working memory accounts is somewhat problematic. In the Bayesian predictive coding framework, this can be viewed as not yet properly formulated marginal (conditional) distributions for the priors on individual features.

[3] This is superficially similar to prototype-based models of semantics, but we note that traditional linguistic models of such phenomena are not necessarily easily mapped onto a neurocomputational framework.

[4] Friston (2005) discusses a possible neural mechanism for short and long term memory based on different neurotransmitters and physiological changes, which is largely compatible with Jonides et al. (2008)'s theoretical synthesis of existing research, but examining this claim and its implications in detail exceeds the bounds of this dissertation.

In other words, we can also describe model adaptation as *memory access*, in that it refers to the dynamics of making information available for further processing. In language research, this has been best described as a pattern of spreading activation in the neuropsychological architecture and is broadly compatible with the dynamical accounts given by Kutas and Federmeier (2011) and Brouwer, Fitz, and Hoeks (2012). The amplitude of a negativity thus correlates (subject to all the usual issues concerning component overlap, conductive cancellation, etc.) broadly with the cost of memory access, i.e. the extent to which information must be activated, or equivalently, to which a model must be updated.

At this point, a metaphor will help present a comparison of these different, yet deeply related perspectives. We can understand "spreading activation" as being like a stone dropped in water — the ripples are largest closest to the point of impact but extend potentially indefinitely with decreasing strength. Moreover, like a stone in water, certain configurations can result in oscillatory behavior, which can either enhance the original effect (constructive interference) or decrease it (destructive interference).[5] In unweighted feature models, the body of water is a simple, symmetric reservoir, and so the effect of a single cue (including interference) propagates in a smooth, symmetric way. In weighted feature models, the body of water is a lake with an uneven bottom and jagged coast, shaped by evolutionary and developmental demands, and so the effect of a single cue reflects in a chaotic way with some stable (attractor) and some unstable (repulsor) configurations.

In summary, the neurocomputational mechanism proposed for negativities can be seen as a generalization of several current theories related broadly to prediction and representation activation.

Positivities

As in Chapter 2, we can understand *evaluation* in a computational sense, i.e. in the sense of function evaluation, which aligns loosely with *evaluation* in the sense of 'value judgment' because executing a model reveals aspects about its fit to the data, i.e. reality. *Decision making* fits into this scheme because evaluation often leads to action in a broad sense, whether physical action coordinated by the motor system and realized by the body, or "mental" action in the sense of "thought", whether subconscious or conscious. In some sense, positivities turn models into behavior.

Bornkessel-Schlesewsky and Schlesewsky (in press), building on the work of Sassenhagen (Sassenhagen 2014; Sassenhagen, Schlesewsky, and Bornkessel-Schlesewsky 2014), provide a broadly compatible view, namely that positivities reflect a re-orientation towards "motivationally significant events". Combining our computational perspective with this neurobiological one, evaluation could potentially viewed as model re-structuring, in the sense of forced updates in the predictive-coding framework.[6]

[5] It is quite felicitous that "interference" here refers both to the physics of the metaphor and to actual phenomena in working memory research.

[6] This would also conveniently explain the early positivity found in Lotze et al. (2011): meaningful physical

This proposal is in line with previous proposals (for the P300, cf. "context updating", Donchin 1981; Donchin and Coles 1988; "stimulus evaluation", Kutas, McCarthy, and Donchin 1977; "event categorization", Kok 2001; for the P600, cf. "model representation composition", Brouwer, Fitz, and Hoeks 2012; "event-structure updating", Schumacher 2011, but we note that many of these proposals are not completely compatible in all their details). Although we assume that decision making exhibits threshold-based behavior (i.e. an all-or-nothing response upon reaching a certain "tipping point"), positivities are, like negativities, continuous modulations of the ERP signal. Thresholds can be reached either by accumulated signal change (i.e. "drift") or by a forced, sudden update based on the immediate processing window (cf. O'Connell, Dockree, and Kelly 2012). In general, only the latter are readily apparent as ERP effects, though O'Connell, Dockree, and Kelly (2012) have demonstrated a "continuous oddball" design for non-linguistic visual stimuli. In other words, slow positive drifts may reflect continuous refinements, while peak-like behavior reflects abrupt re-evaluation associated with e.g. decision making.

At this point, an additional metaphor may help highlight the difference between the evaluations reflected by positivities compared to the updates reflected by negativities. *Evaluation* means in some sense to attempt to answer the question, *Is this an appropriate model?*,[7] while *update* means changing the current model.

Hierarchical regression modeling provides a convenient metaphor. Negativities correspond to the (group) variance or error terms (random effects), while positivities correspond to the "big picture" of a given model, i.e. whether the choice of ecological predictors (fixed effects) and dependent variable (target) is appropriate.

6.1.3. Implications

In the following, we present in brief a few possible implications of this initial proposal. In particular, the proposed architecture strongly suggests the following corollaries:

Corollary (Prediction is pre-activation)**.** *Prediction-based processes in processing are the result of spreading activation from a prior stimulus and the resulting processing. Equivalently, spreading activation reflects model updates in a predictive coding sense. (cf. Friston 2005)*

Corollary (Neurocomputational implementation of "top-down" vs. "bottom-up" processing)**.** *Top-down influences reflect back propagation, while bottom-up influences reflect forward propagation of activation along the processing stream. (cf. Friston 2005)*

changes are informative enough to force model restructuring, which is reflected in an early positivity. This restructuring subsequently decreases prediction error, and thus modulates the N400 amplitude.

[7] This also explains the absence of positivity *effects* in non-binary contexts (Bornkessel-Schlesewsky, Kretzschmar, et al. 2011). Because there is no immediate, clean answer, the positivity only appears as a small modulation the ERP signal and not as a classical, peaky components (see above and Chapter 5).

Although specific predictions are often seen as distinct from spreading activation in working-memory accounts, this is not necessary in our view of spreading activation. If we consider the metaphor of a stone dropped into water (see Section 6.1.2, above), then we can arrive at strong, focal activation by an appropriately constrained initial configuration, where the ripples of activation are shaped and guided by environmental factors. In particular, predictions that are ecologically salient, such as those required by the task, are stronger and longer lasting because the task creates a positive feedback loop.

Corollary (Mismatches are an epiphenomenon)**.** *Mismatch-related phenomena are simply the result of insufficient pre-activation, i.e. a failed prediction. (cf. Friston 2005)*

These corollaries are largely direct implications of Friston (2005)'s account and our application of it here, where predictive coding permeates the entire processing pipeline. These implications are also compatible with recent surprisal-based accounts of the N400 (Frank et al. 2015), suggesting that mathematical information theory may provide the tools needed to better quantify this proposal. Indeed, Friston (Friston 2005, 2009; Friston and Kiebel 2009) uses *free energy* as a neurobiologically plausible mathematical-physical formalism,[8] and free energy is a function of entropy.

6.2. Divergence from Existing Theories

The ideas presented here are not new, as we have emphasized above. In particular, much of the work conducted here was done in parallel to work by Sassenhagen, and so it is worth briefly emphasizing one key difference. In particular, Sassenhagen (2014, p. 196) suggested that there is a basic biphasic pattern consisting of a negativity followed by a positivity, and, in particular, that negativities represent "incongruences between multiple unattended streams" and positivities the "transitioning of the cortex to an appropriate state following the evaluation of the incongruent event".

Our theory differs from this latter proposal in its gradedness. Neurocomputation at the level revealed by EEG is not all or nothing. In the case of negativities, there is *always* an incongruency,[9] albeit often very small, because every model has some level of error (cf. Friston 2005). Positivities indeed reflect the ongoing evaluation of the stimulus, congruent or not. The sharp difference and peaky nature of ERP components observed in experiments is a result of experimental manipulation and discrete nature of stimulus presentation in traditional designs (cf. Alday, Schlesewsky, and Bornkessel-Schlesewsky submitted). Compared to a congruent stimulus, an incongruent stimulus does impose a greater computational demand,

[8] Free energy in thermodynamics is the energy available in physical system to do work. There are several formulations of this principle in the physical science with different restrictions (Gibbs for uniform temperature and pressure, Helhomltz for constant temperature and volume, etc.), but much like entropy, free energy also has a statistical or information-theoretic interpretation, namely as a lower bound on "the surprise from sampling some data, given a generative model" (Friston 2009).

[9] Even Kutas and Hillyard (1980) found an N400 for *all* words, but it was only an *effect*, in the sense of differing between conditions, for certain contrasts.

which is clearly seen in experimental manipulations, while the ongoing low-level positivities for continuous evaluation are absorbed into the background. In this sense, Sassenhagen's statements are true if we restrict them to experimental *effects* rather than *components*.

6.3. Relationship to Previous Computational Work on the Actor Heuristic

In light of these considerations, we can consider what this theory means for the central topic of this dissertation, the actor heuristic.

In particular, ideal actors tend to elicit reduced N400s compared to poor actors (increased N400 amplitude for object-initial constructions, Frisch and Schlesewsky 2001; increased N400 for inanimates, Weckerly and Kutas 1999, etc.). One possibility is a purely frequency-based account (cf. frequency as a prior, Alday, Schlesewsky, and Bornkessel-Schlesewsky in press) — animates are more frequent than inanimates, initial nominatives are more common than initial accusatives — but it is possible to dissociate these frequency effects (Bornkessel, Schlesewsky, and Friederici 2002). Moreover, this account does not answer the *why*, but only pushes it down a level — why are these configurations more common? The actor heuristic grounds itself in environmental demands and thus provides a more satisfying answer to the "why" part, but how does that influence the amplitude of the N400? The actor, as an essential category in language processing, is a pre-activated or potentially even default state (see also "default representation" under *Negativities*, above). Forcing an argument away from actorhood takes energy (cf. attractor basin in Chapter 3, and above). Thus prominence tends to correlate inversely with markedness in a linguistic sense.

Positivites are somewhat more complicated. For less frequent or dispreferred constructions that are nonetheless well-formed, they could simply reflect the necessary model changes, i.e. re-orientation, necessary to accomodate unusual input. Indeed, this matches well with the idea that positivities also correlate with attention as part of behavioral re-orientation and e.g. non-canonical word orders are often used to draw attention to unusual circumstances. "Erroneous" input (both syntactic and semantic, see below) is also equally demanding of attention and thus also elicits a late positivity.

In terms of the parsing work from Chapter 2, we expect N400 amplitude to correlate with configurations where the parser shows reduced accuracy across trials, particularly in arc-direction, because the increased competition reflects a less clear ordering of the arguments. Late positivities should be reflected in "double-attachment" errors, i.e. errors where the parser generates two actor or two undergoer attachments and fails to generate the other attachment. This follows from two major findings concerning late positivities. For "syntactic" errors, which will obviously mislead the parser and from which even humans may have trouble escaping, we can expect attachment errors as a direct result of the ill-formedness, i.e. lack of a canonical tree representation. In severe cases, this may be reflected by dangling elements being coerced into tree form by an attachment to the ROOT node. For "semantic" errors, such as semantic-reversal anomalies (e.g., *The ham ate Steve*, for a review, see Brouwer,

Fitz, and Hoeks 2012; but N.B. Tune et al. 2014), the parser lacks the necessary "evaluation" or "re-orientation" mechanisms to correct a previous mistake and is forced to cut the Gordian knot by a double attachment. The proposed correspondence, however, is very preliminary and has not been explicitly tested, and assumes more direct mapping between parser behavior and electrophysiology than we assumed in Chapter 2.

6.4. Review and Outlook

In this chapter, we outlined a parsimonious neurocomputational account of many ERP components and showed that it subsumes many previous models from the literature. In particular, we claim that there are two large component groups whose differing polarity is indicative of distinct computational operations. This proposal answers many long standing debates related to the electrophysiology of language, often by showing them to be moot or ill-formed (e.g., Does the P600 belong to the P300 family? Does the N400 belong to the N200 family? Which neurocognitive process does the N400 index?), but does not address questions regarding the underlying neurobiological implementation in neuroanatomy and -physiology. Further research should focus on supplying the missing neuroanatomical and even neurophysiological details.

7. Conclusion

> Da stehe ich nun, ich armer Thor!
> Und bin so klug als wie zuvor.
>
> Johann Wolfgang von Goethe, *Faust*

The work presented here unabashedly borrowed tools, methods and perspectives from a range of fields.

We started with **computational linguistics** and examined the possibility of using a successful parsing technique to examine the optimality of various heuristics posited for the language system (Chapter 2). An initial application of the theory we developed to a real parser with a restricted training set yielded somewhat disappointing performance and showed a great sensitivity to certain free parameters (i.e. feature model specification). More research will be required to determine sensible values for these parameters, which made apparent the necessity of larger training corpora for future research. Nonetheless, some results were promising in their human-like performance characteristics, which suggests that this method may yet have more insights to offer and that the effort required for developing training corpora may be worth the investment.

Using the tools of **linear algebra and real analysis**, we provided a simple yet rigorous formalism for ideas from **psycholinguistics** and **cognitive psychology** such as prominence and distinctness (Chapter 3). We connected these measures to real human EEG data and provided a quantitative account of the influence of prominence on the EEG signal. Additionally, we demonstrated the compatibility of this mathematical model with recent suggestions of attractor-basins as a fundamental processing mechanism.

The development of this mathematical model unfortunately depended on a number of free parameters, but suggested a way to estimate them. Using **regression models**, we were able to estimate parameters even at the single-subject level (Chapter 4). Plugging these estimates back into the models from Chapter 3 showed the previous results to be tenable and not just based on favorably chosen parameter values.

Previous attempts to analyze EEG data from naturalistic contexts have suffered from the interaction of high temporal sensitivity of EEG and the limits of traditional statistical methods. Inspired by our success in Chapter 3, we demonstrated the feasibility of analyzing EEG in a rich, naturalistic context using **hierarchical models** (Chapter 5). This approach suggested that the usual intuition about ERP components being discrete events was misguided.

Moreover, examining the complex interactions of prominence features in a rich context emphasized the importance of a graded, holistic perspective on language processing instead of a categorial, parametric view.

Finally, this last attempt inspired a neurocomputational theory of cognitive ERP components based on two fundamental mechanisms, model updates and evaluation. We suggest that more complicated processes such as prediction and conflict monitoring arise from basic properties of these two fundamental mechanisms. Attractor basins, and hence the actor heuristic, fit cleanly within this framework.

Together, these approaches present a many-faceted exploration of the properties of both the actor heuristic and its neurocognitive implementation. None of the methods presented here were revolutionary, but each model revealed new insights, both in its successes and in its failures. The complexity of language far exceeds all of our models, but even simple models can deepen our understanding, and the more specific the model is — i.e. the more quantitative it is or the more fully computationally implemented — the more helpful it is. Combining simple models from multiple approaches allows us to close the gap that much faster.

At the beginning of this dissertation, we suggested that quantitative methods were the key to the future. Looking back, we were not always explicitly quantitative, but in general we provided for computationally supported methods and models that will serve as the foundation for future quantitative work, with each implementation revealing hidden assumptions and underspecifications. The transition from qualitative stories to quantitative predictions will take time. In many ways, we do not yet know what "quantitative" looks like in a particular problem domain. But we will never know if we do not try. And even if we get it wrong, well, then we just have one more data point to work with.

Bibliography

Alday, P. M. (2010). "Regularisierungstendenzen in der Morphologie germanischer Sprachen". MA thesis. Philipps-Universität Marburg.

Alday, P. M., A. Nagels, et al. (2011). *Actor Identification in Natural Stories: Qualitative Distinctions in the Neural Bases of Actor-related Features*. Talk presented at the Neurobiology of Language Conference. Annapolis.

Alday, P. M., J. Sassenhagen, and I. Bornkessel-Schlesewsky (2014a). *Neural Signatures of Incremental Text Processing Correlate with Word Entropy in a Natural Story Context*. Poster presented at the Society for Neurobiology of Language Conference. Amsterdam.

– (2014b). *Tracking the Emergence of Meaning in the Brain during Natural Story Comprehension*. Poster presented at the International Conference on Cognitive Neuroscience (ICON). Brisbane.

Alday, P. M., M. Schlesewsky, and I. Bornkessel-Schlesewsky (2012). *Towards a Computational Model of Actor-based Language Comprehension*. Poster presented at the Neurobiology of Language Conference. San Sebastian.

– (2014). "Towards a Computational Model of Actor-based Language Comprehension". In: *Neuroinformatics* 12.1, pp. 143–179. DOI: 10.1007/s12021-013-9198-x.

– (in press). "Discovering Prominence and its Role in Language Processing: An Individual (Differences) Approach". In: *Linguistic Vanguard*. DOI: 10.1515/lingvan-2014-1013.

– (submitted). "Electrophysiology Reveals the Neural Dynamics of Naturalistic Auditory Language Processing: Event-Related Potentials Reflect Continuous Model Updates". In: *Journal of Neuroscience*.

Baayen, R. H., D. J. Davidson, and D. M. Bates (2008). "Mixed-effects Modeling with Crossed Random Effects for Subjects and Items". In: *Journal of Memory and Language* 59, pp. 390–412.

Barr, D. J. (2008). "Analyzing 'Visual World' Eyetracking Data using Multilevel Logistic Regression". In: *Journal of Memory and Language* 59.4, pp. 457–474. DOI: 10.1016/j.jml.2007.09.002.

– (2013). "Random Effects Structure for Testing Interactions in Linear Mixed-effects Models". In: *Frontiers in Psychology* 4.328. DOI: 10.3389/fpsyg.2013.00328.

Barr, D. J. et al. (2013). "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal". In: *Journal of Memory and Language* 68, pp. 255–278. DOI: 10.1016/j.jml.2012.11.001.

Basten, U. et al. (2010). "How the Brain Integrates Costs and Benefits during Decision Making." In: *Proceedings of the National Academy of Sciences* 107.50, pp. 21767–21772.

Bates, E. and B. MacWhinney (1989). "Cross-linguistic Research in Language Acquisition and Language Processing". In: *Proceedings of the World Conference on Basque Language and Culture*. San Sebastian: Basque Regional Government.

Bates, E., S. McNew, et al. (1982). "Functional Constraints on Sentence Processing: A Cross-linguistic Study". In: *Cognition* 11, pp. 245–299.

Bornkessel, I. (2002). "The Argument Dependency Model: A Neurocognitive Approach to Incremental Interpretation". PhD thesis. University of Potsdam.

Bornkessel, I. and M. Schlesewsky (2006). "The Extended Argument Dependency Model: A Neurocognitive Approach to Sentence Comprehension Across Languages". In: *Psychological Review* 113.4, pp. 787–821.

Bornkessel, I., M. Schlesewsky, and A. D. Friederici (2002). "Grammar Overrides Frequency: Evidence from Online Processing of Flexible Word Order". In: *Cognition* 85, B21–B30.

Bornkessel-Schlesewsky, I., F. Kretzschmar, et al. (2011). "Think Globally: Cross-linguistic Variation in Electrophysiological Activity during Sentence Comprehension". In: *Brain and Language* 117.3. First Neurobiology of Language Conference: NLC 2009, Neurobiology of Language Conference, pp. 133–152. DOI: 10.1016/j.bandl.2010.09.010.

Bornkessel-Schlesewsky, I. and M. Schlesewsky (2009). "The Role of Prominence Information in the Real-Time Comprehension of Transitive Constructions: A Cross-Linguistic Approach". In: *Language and Linguistics Compass* 3.1, pp. 19–58. DOI: 10.1111/j.1749-818x.2008.00099.x.

– (2013). "Reconciling Time, Space and Function: A New Dorsal-Ventral Stream Model of Sentence Comprehension". In: *Brain and Language* 125, pp. 60–76. DOI: 10.1016/j.bandl.2013.01.010.

– (2014). "Competition in Argument Interpretation: Evidence from the Neurobiology of Language". In: *Competing Motivations in Grammar and Usage*. Ed. by B. MacWhinney, A. Malchukov, and E. Moravcsik. Oxford: Oxford University Press.

– (in press). "The Argument Dependency Model". In: *Neurobiology of Language*. Ed. by G. Hickok and S. Small. Academic Press. Chap. 30. DOI: 10.1016/B978-0-12-407794-2.00030-4.

Bornkessel-Schlesewsky, I., M. Schlesewsky, et al. (in press). "Neurobiological Roots of Language in Primate Audition: Common Computational Properties". In: *Trends in Cognitive Sciences*. DOI: 10.1016/j.tics.2014.12.008.

Bornkessel-Schlesewsky, I. and M. Schlesewsky (2008). "An Alternative Perspective on "Semantic P600" Effects in Language Comprehension". In: *Brain Research Reviews* 59, pp. 55–73. DOI: 10.1016/j.brainresrev.2008.05.003.

Brouwer, H., H. Fitz, and J. Hoeks (2012). "Getting Real about Semantic Illusions: Rethinking the Functional Role of the P600 in Language Comprehension". In: *Brain Research* 1446, pp. 127–143.

Comrie, B. (2011). "Alignment of Case Marking of Full Noun Phrases". In: *The World Atlas of Language Structures Online*. Ed. by M. S. Dryer and M. Haspelmath. Munich: Max Planck Digital Library.

Coulson, S., J. W. King, and M. Kutas (1998a). "ERPs and Domain Specificity: Beating a Straw Horse". In: *Language and Cognitive Processes* 13.6, pp. 653–672.

– (1998b). "Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations". In: *Language and Cognitive Processes* 13.1, pp. 21–58.

Crepaldi, D. et al. (2011). "A Place for Nouns and a Place for Verbs? A Critical Review of Neurocognitive Data on Grammatical-class Effects". In: *Brain and Language* 116.1, pp. 33–49. DOI: 10.1016/j.bandl.2010.09.005.

Cumming, G. (2013). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Multivariate Applications Series. Taylor and Francis.

– (2014). "The New Statistics: Why and How". In: *Psychological Science* 20.10, pp. 1–23. DOI: 10.1177/0956797613504966.

Dawson, M. (1998). *Understanding Cognitive Science*. Blackwell Publishing.

Deco, G., E. T. Rolls, L. Albantakis, et al. (2013). "Brain Mechanisms for Perceptual and Reward-related Decision-making". In: *Progress in Neurobiology* 103.194–213. DOI: 10.1016/j.pneurobio.2012.01.010.

Deco, G., E. T. Rolls, and R. Romo (2009). "Stochastic Dynamics as a Principle of Brain Function". In: *Progress in Neurobiology* 88.1, pp. 1–16. DOI: 10.1016/j.pneurobio.2009.01.006.

DeWitt, I. and J. P. Rauschecker (2012). "Phoneme and Word Recognition in the Auditory Ventral Stream". In: *Proceedings of the National Academy of Sciences* 109.8, E505–E514. DOI: 10.1073/pnas.1113427109.

Donchin, E. (1981). "Surprise!... Surprise?" In: *Psychophysiology* 18.5, pp. 493–513. DOI: 10.1111/j.1469-8986.1981.tb01815.x.

Donchin, E. and M. G. H. Coles (1988). "Is the P300 Component a Manifestation of Context Updating?" In: *Behavioral and brain sciences* 11.03, pp. 357–374. DOI: 10.1017/S0140525X0005802.

Dowty, D. (1991). "Thematic Proto-Roles and Argument Selection". In: *Language* 67.3, pp. 547–619.

Dröge, A., M. Schlesewsky, and I. Bornkessel-Schlesewsky (2012). *Separable Effects of Lexical Association and Plausibility on the N400*. Poster presented at the Architectures and Mechanisms for Language Processing (AMLaP) Conference. Riva del Garda, Italy.

Federmeier, K. D. et al. (2000). "Brain Responses to Nouns, Verbs and Class-ambiguous Words in Context". In: *Brain* 123.12, pp. 2552–2566. DOI: 10.1093/brain/123.12.2552.

Fidler, F. et al. (2004). "Statistical Reform in Medicine, Psychology and Ecology". In: *The Journal of Socio-Economics* 33, pp. 615–630. DOI: 10.1016/j.socec.2004.09.035.

Frank, S. L. et al. (2015). "The ERP Response to the Amount of Information Conveyed by Words in Sentences". In: *Brain and Language* 140, pp. 1–11. DOI: 10.1016/j.bandl.2014.10.006.

Frenzel, S., M. Schlesewsky, and I. Bornkessel-Schlesewsky (2015). "Two Routes to Actorhood: Lexicalized Potency to Act and Identification of the Actor Role". In: *Frontiers in Psychology* 6.1. DOI: 10.3389/fpsyg.2015.00001.

Friederici, A. D. (2002). "Towards a Neural Basis of Auditory Sentence Processing". In: *Trends in Cognitive Sciences* 6.2, pp. 78–84. DOI: 10.1016/S1364-6613(00)01839-8.

– (2011). "The Brain Basis of Language Processing: from Structure to Function". In: *Physiological Reviews* 91.4, pp. 1357–1392. DOI: 10.1152/physrev.00006.2011.

Frisch, S. and M. Schlesewsky (2001). "The N400 Reflects Problems of Thematic Hierarchizing". In: *NeuroReport* 12.15, pp. 3391–3394.

Friston, K. (2005). "A Theory of Cortical Responses". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1456, pp. 815–836. DOI: 10.1098/rstb.2005.1622.

Friston, K. (2009). "The Free-energy Principle: a Rough Guide to the Brain?" In: *Trends in Cognitive Sciences* 13.7, pp. 293–301. DOI: 10.1016/j.tics.2009.04.005.

Friston, K. and S. Kiebel (2009). "Predictive Coding under the Free-energy Principle". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1521, pp. 1211–1221. DOI: 10.1098/rstb.2008.0300.

Garrido, M. I. et al. (2009). "The Mismatch Negativity: A Review of Underlying Mechanisms". In: *Clinical Neurophysiology* 120.3, pp. 453–463. DOI: 10.1016/j.clinph.2008.11.029.

Gelman, A. and E. Loken (2013). "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There is No "Fishing Expedition" or "p-hacking" and the Research Hypothesis was Posited Ahead of Time". In:

Gibson, E. (1998). "Linguistic complexity: Locality of syntactic dependencies". In: *Cognition* 68, pp. 1–76. DOI: 10.1016/S0010-0277(98)00034-1.

– (2000). "The Dependency Locality Theory: A Distance-based Theory of Linguistic Complexity". In: *Image, Language, Brain*. Ed. by Y. Miyashita, A. Marantz, and W. O'Neil. Cambridge, MA: MIT Press, pp. 95–126.

Giraud, A.-L. and D. Poeppel (2012). "Cortical Oscillations and Speech Processing: Emerging Computational Principles and Operations". In: *Nature Neuroscience* 15.4, pp. 511–517. DOI: 10.1038/nn.3063.

Haupt, F. S. et al. (2008). "The Status of Subject–Object Reanalyses in the Language Comprehension Architecture". In: *Journal of Memory and Language* 59, pp. 54–96. DOI: 10.1016/j.jml.2008.02.003.

Heekeren, H. et al. (2004). "A General Mechanism for Perceptual Decision-making in the Human Brain". In: *Nature* 431.7010, pp. 859–862. DOI: 10.1038/nature02966.

Honnibal, M., Y. Goldberg, and M. Johnson (2013). "A Non-Monotonic Arc-Eager Transition System for Dependency Parsing". In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 163–172.

Howes, A., R. L. Lewis, and A. Vera (2009). "Rational Adaptation under Task and Processing Constraints: Implications for Testing Theories of Cognition and Action". In: *Psychological Review* 116.4, pp. 717–751. DOI: 10.1037/a0017187.

Jaeger, T. F. (2008). "Categorical Data Analysis: Away from ANOVAs (Transformation or Not) and Towards Logit Mixed Models". In: *Journal of Memory and Language* 59.4, pp. 434–446. DOI: 10.1016/j.jml.2007.11.007.

Jonides, J. et al. (2008). "The Mind and Brain of Short-Term Memory". In: *Annual Reviews* 59, pp. 193–224. DOI: 10.1146/annurev.psych.59.103006.093615.

Kempe, V. and B. MacWhinney (1999). "Processing of Morphological and Semantic Cues in Russian and German". In: *Language and Cognitive Processes* 14.2, pp. 129–171.

Kliegl, R., M. E. J. Masson, and E. M. Richter (2010). "A Linear Mixed Model Analysis of Masked Repetition Priming". In: *Visual Cognition* 18.5, pp. 655–681. DOI: 10 . 1080 / 13506280902986058.

Kliegl, R., P. Wei, et al. (2010). "Experimental Effects and Individual Differences in Linear Mixed Models: Estimating the Relationship between Spatial, Object, and Attraction Effects in Visual Attention". In: *Frontiers in Psychology* 1.238. DOI: 10.3389/fpsyg.2010.00238.

Knoeferle, P., T. P. Urbach, and M. Kutas (2014). "Different Mechanisms for Role Relations versus Verb–action Congruence Effects: Evidence from ERPs in Picture–sentence Verification". In: *Acta Psychologica* 152, pp. 133–148. DOI: 10.1016/j.actpsy.2014.08.004.

Kok, A. (2001). "On the Utility of P3 Amplitude as a Measure of Processing Capacity". In: *Psychophysiology* 38.3, pp. 557–577. DOI: 10.1017/S0048577201990559.

Kretzschmar, F. (2010). "The Electrophysiological Reality of Parafoveal Processing: On the Validity of Language-related ERPs in Natural Reading". PhD thesis. Philipps-Universität Marburg.

Kretzschmar, F., I. Bornkessel-Schlesewsky, and M. Schlesewsky (2009). "Parafoveal versus Foveal N400s Dissociate Spreading Activation from Contextual Fit". In: *NeuroReport* 20.18, pp. 1613–1618. DOI: 10.1097/WNR.0b013e328332c4f4.

Kutas, M. and K. D. Federmeier (2011). "Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP)". In: *Annual Review of Psychology* 62.1, pp. 621–647. DOI: 10.1146/annurev.psych.093008.131123.

Kutas, M. and S. A. Hillyard (1980). "Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity". In: *Science* 207.4427, pp. 203–205.

Kutas, M., G. McCarthy, and E. Donchin (1977). "Augmenting Mental Chronometry: The P300 as a Measure of Stimulus Evaluation Time". In: *Science* 197.4305, pp. 792–795.

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes*. Cambridge University Press.

Lakatos, I. and A. Musgrave (1970). *Criticism and the Growth of Knowledge*. Cambridge University Press.

Lau, E. F., C. Phillips, and D. Poeppel (2008). "A Cortical Network for Semantics: (De)constructing the N400". In: *Nature Reviews Neuroscience* 9.12, pp. 920–933. DOI: 10.1038/nrn2532.

Lewis, R. L. and S. Vasishth (2005). "An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval". In: *Cognitive Science* 29, pp. 375–419. DOI: 10 . 1207 / s15516709cog0000_25.

Lewis, R. L., S. Vasishth, and J. A. Van Dyke (2006). "Computational Principles of Working Memory in Sentence Comprehension". In: *Trends in Cognitive Sciences* 10.10, pp. 447–454. DOI: 10.1016/j.tics.2006.08.007.

Lotze, N. et al. (2011). "Meaningful Physical Changes Mediate Lexical-Semantic Integration: Top-Down and Form-based Bottom-Up Information Sources Interact in the N400". In: *Neuropsychologia* 49, pp. 3573–3582. DOI: 10.1016/j.neuropsychologia.2011.09.009.

Luck, S. J. and E. K. Vogel (2013). "Visual Working Memory Capacity: from Psychophysics and Neurobiology to Individual Differences". In: *Trends in Cognitive Sciences* 17.8, pp. 391–400. DOI: 10.1016/j.tics.2013.06.006.

MacWhinney, B. and E. Bates, eds. (1989). *The Crosslinguistic Study of Sentence Processing*. Cambridge University Press.

MacWhinney, B., E. Bates, and R. Kliegl (1984). "Cue Validity and Sentence Interpretation in English, German and Italian". In: *Journal of Verbal Learning and Verbal Behavior* 23.2, pp. 127–50.

Mahalanobis, P. C. (1936). "On the Generalised Distance in Statistics". In: *Proceedings of the National Institute of Sciences of India* 2.1, pp. 49–55.

Marr, D. and T. Poggio (1976). "From Understanding Computation to Understanding Neural Circuitry". In: *AI Memos* 357.

McElree, B. (2006). "Accessing Recent Events". In: *The Psychology of Learning and Motivation*. Ed. by B. H. Ross. Vol. 46. San Diego: Academic Press.

Meyer, W. S. et al. (1909). "Sanskrit Literature". In: *The Imperial Gazetteer of India* 2.

Nivre, J. (2009). "Parsing Indian Languages with MaltParser". In: *Proceedings of the ICON09 NLP Tools Contest: Indian Language Dependency Parsing*, pp. 12–18.

– (2010). "Dependency Parsing". In: *Language and Linguistics Compass* 4.3, pp. 138–152. DOI: 10.1111/j.1749-818X.2010.00187.x.

Nivre, J. et al. (2007). "MaltParser: A Language-Independent System for Data-Driven Dependency Parsing". In: *Natural Language Engineering* 13.2, pp. 95–135.

O'Connell, R. G., P. M. Dockree, and S. P. Kelly (2012). "A Supramodal Accumulation-To-Bound Signal that Determines Perceptual Decisions in Humans". In: *Nature Neuroscience* 15.12, pp. 1729–1735. DOI: 10.1038/nn.3248.

Philipp, M. et al. (2008). "The Role of Animacy in the Real Time Comprehension of Mandarin Chinese: Evidence from Auditory Event-Related Brain Potentials". In: *Brain and Language* 105.2, pp. 112–133. DOI: 10.1016/j.bandl.2007.09.005.

Popper, K. (1934). *Logik der Forschung*. Mohr Siebeck.

Primus, B. (1999). *Cases and Thematic Roles*. Tübingen: Niemeyer.

Rauschecker, J. P. and S. K. Scott (2009). "Maps and Streams in the Auditory cortex: Nonhuman Primates Illuminate Human Speech Processing". In: *Nature Neuroscience* 12.6, pp. 718–724. DOI: 10.1038/nn.2331.

Roehm, D. et al. (2004). "Fractionating Language Comprehension via Frequency Characteristics of the Human EEG". In: *NeuroReport* 15.3, pp. 409–412.

Sassenhagen, J. (2014). "Evoked Potentials during Language Processing as Neurophysiological Phenomena". PhD thesis. Philipps-Universität Marburg.

Sassenhagen, J., M. Schlesewsky, and I. Bornkessel-Schlesewsky (2014). "The P600-as-P3 Hypothesis Revisited: Single-trial Analyses Reveal that the Late EEG Positivity Following Linguistically Deviant Material is Reaction Time Aligned". In: *Brain and Language* 137, pp. 29–39. DOI: 10.1016/j.bandl.2014.07.010.

Schumacher, P. B. (2011). "The Hepatitis Called ... Electrophysiological Evidence for Enriched Composition". In: *Experimental Pragmatics/Semantics*. Ed. by J. Meibauer and M. Steinbach. John Benjamins, pp. 199–219.

Seidenberg, M. S. and M. F. Joanisse (2003). "Show Us the Model". In: *Trends in Cognitive Sciences* 7.3, pp. 106–107. DOI: 10.1016/S1364-6613(03)00020-2.

Shapiro, K. and A. Caramazza (2003). "The Representation of Grammatical Categories in the Brain". In: *Trends in Cognitive Sciences* 7.5, pp. 201–206. DOI: 10.1016/S1364-6613(03)00060-3.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011). "False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". In: *Psychological Science* 22.11, pp. 1359–1366. DOI: 10.1177/0956797611417632.

Tune, S. et al. (2014). "Cross-linguistic Variation in the Neurophysiological Response to Semantic Processing: Evidence from Anomalies at the Borderline of Awareness". In: *Neuropsychologia* 56, pp. 147–166. DOI: 10.1016/j.neuropsychologia.2014.01.007.

Tyler, L. K. et al. (2004). "Neural processing of nouns and verbs: the role of inflectional morphology". In: *Neuropsychologia* 42.4, pp. 512–523. DOI: 10.1016/j.neuropsychologia.2003.10.001.

Van Petten, C. and B. J. Luka (2012). "Prediction during Language Comprehension: Benefits, Costs, and ERP Components". In: *International Journal of Psychophysiology* 83.2, pp. 176–190. DOI: 10.1016/j.ijpsycho.2011.09.015.

Van Valin, R. D. (2005). *Exploring the Syntax-semantics Interface*. Cambridge: Cambridge University Press.

Vigliocco, G. et al. (2011). "Nouns and Verbs in the Brain: A Review of Behavioural, Electrophysiological, Neuropsychological and Imaging Studies". In: *Neuroscience and Biobehavioral Reviews* 35.3, pp. 407–426. DOI: 10.1016/j.neubiorev.2010.04.007.

Vogel, E. K. and M. G. Machizawa (2004). "Neural Activity Predicts Individual Differences in Visual Working Memory Capacity". In: *Nature* 428, pp. 748–751.

Wagenmakers, E.-J. et al. (2012). "An Agenda for Purely Confirmatory Research". In: *Perspectives on Psychological Science* 7.6, pp. 632–638. DOI: 10.1177/174569161246307.

Weckerly, J. and M. Kutas (1999). "An Electrophysiological Analysis of Animacy Effects in the Processing of Object Relative Sentences". In: *Psychophysiology* 36, pp. 559–570. DOI: 10.1017/S0048577299971202.

Whitney, C. et al. (2009). "Neural Correlates of Narrative Shifts during Auditory Story Comprehension". In: *NeuroImage* 47, pp. 360–366.

Winkler, I., S. Haufe, and M. Tangermann (2011). "Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals". In: *Behavioral and Brain Functions* 7.1.

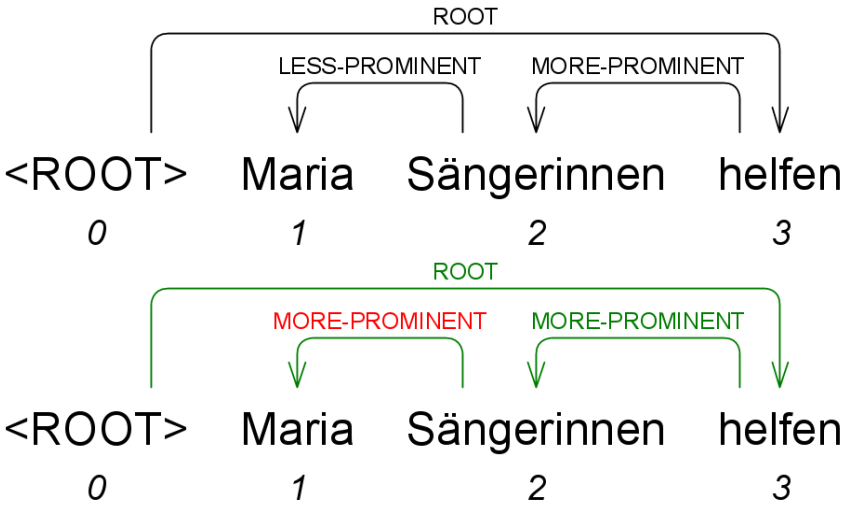
Wolpert, D. M. (1997). "Computational Approaches to Motor Control". In: *Trends in Cognitive Sciences* 1.6, pp. 209–216. DOI: 10.1016/S1364-6613(97)01070-X.

A. Selected Leave-One-Out Results for a Non-Replicable Feature Model

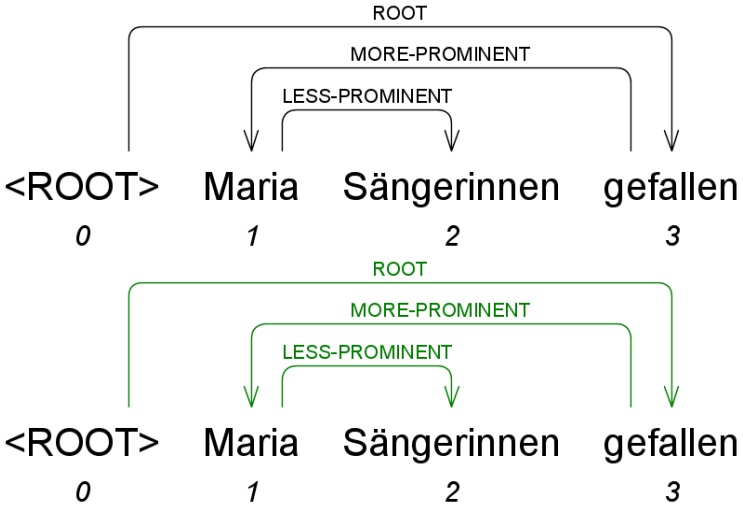
A.1. Ambiguous

A.1.1. Prominence Labels

Active Verbs

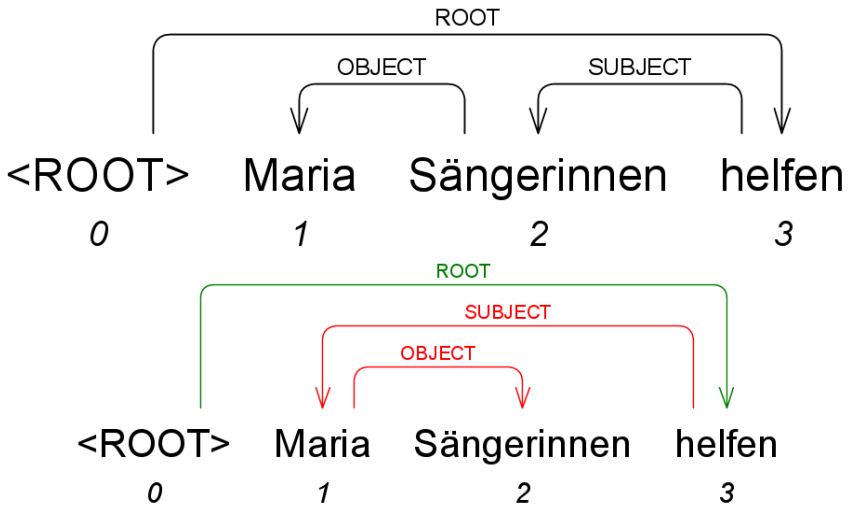


Object-Experiencer Verbs

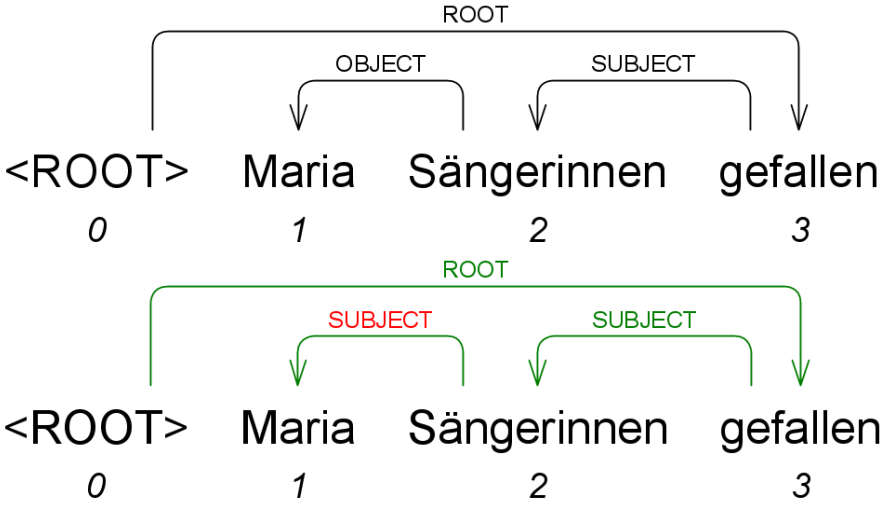


A.1.2. Grammatical-Relation Labels

Active Verbs



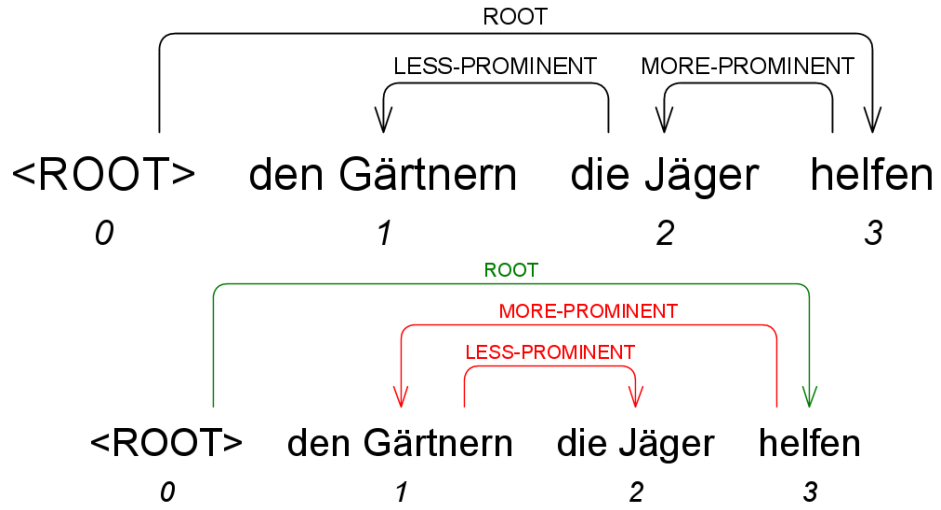
Object-Experiencer Verbs



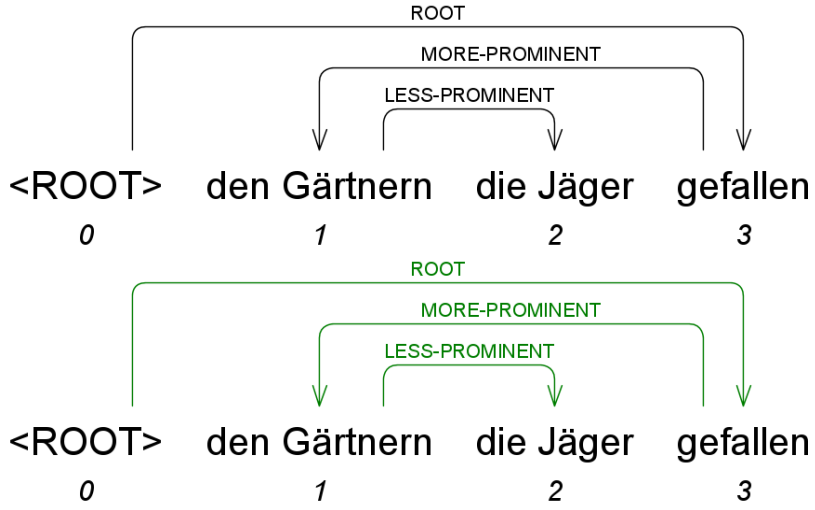
A.2. Unambiguous

A.2.1. Prominence Labels

Active Verbs

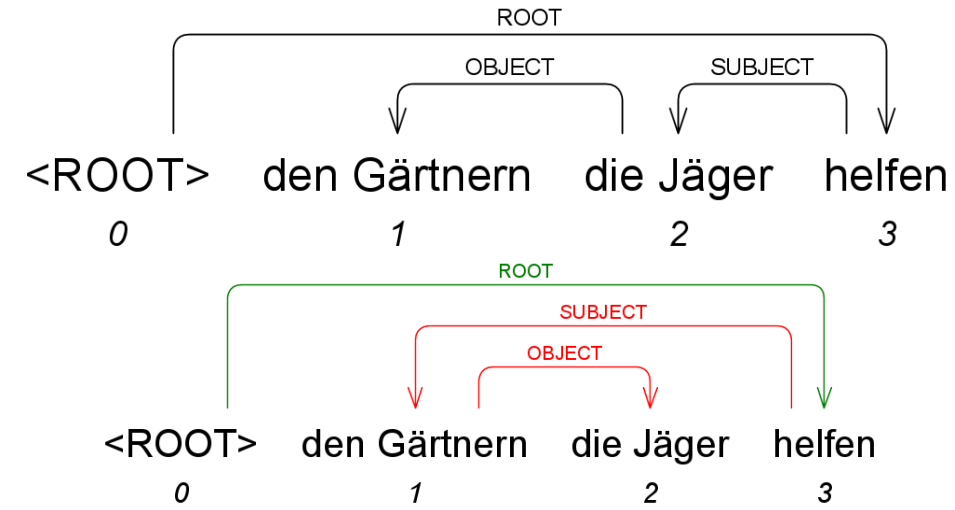


Object-Experiencer Verbs

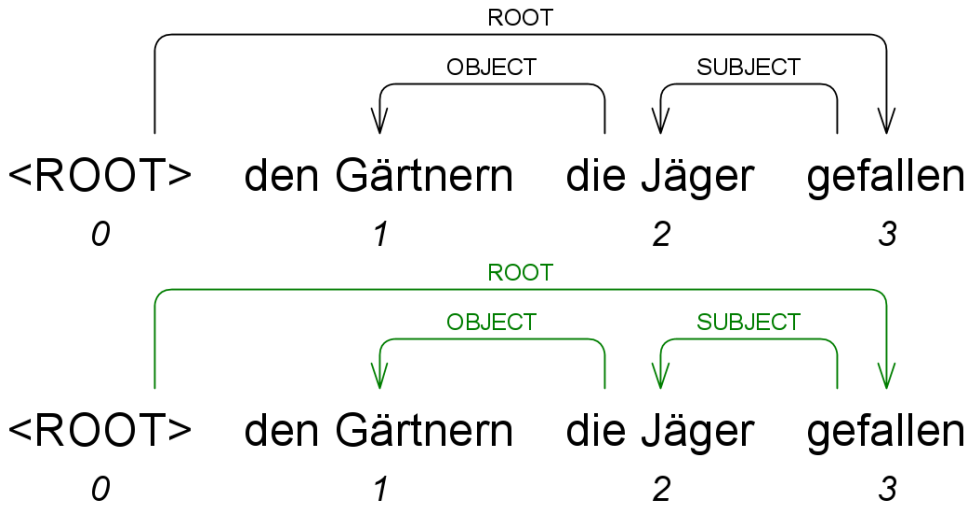


A.2.2. Grammatical-Relation Labels

Active Verbs



Object-Experiencer Verbs



Eidesstattliche Versicherung

Hiermit versichere ich, dass ich die vorgelegte Dissertation mit dem Titel

Quantity and Quality: Not a Zero-Sum Game. A computational and neurocognitive examination of human language processing

selbst und ohne fremde Hilfe verfasst, keine andere als die in ihr angegebenen Quellen oder Hilfsmittel benutzt (einschließlich des World Wide Web und anderen elektronischen Text- und Datensammlungen), alle vollständig oder sinngemäß übernommene Zitate als solche gekennzeichnet sowie die Dissertation in der vorliegenden oder einer ähnlichen Form noch keiner anderen in- oder ausländischen Hochschule anlässlich eines Promotionsgesuches oder zu anderen Prüfungszwecken eingereicht habe.

Marburg, den 10. Februar 2015 .....
Ort, Datum Unterschrift

Summary in English

Quantification of existing theories is a great challenge but also a great chance for the study of language in the brain. While quantification is necessary for the development of precise theories, it demands new methods and new perspectives. In light of this, four complementary methods were introduced to provide a quantitative and computational account of the extended Argument Dependency Model from Bornkessel-Schlesewsky and Schlesewsky.

First, a computational model of human language comprehension was introduced on the basis of dependency parsing. This model provided an initial comparison of two potential mechanisms for human language processing, the traditional "subject" strategy, based on grammatical relations, and the "actor" strategy based on prominence and adopted from the eADM. Initial results showed an advantage for the traditional "subject" model in a restricted context; however, the "actor" model demonstrated behavior in a test run that was more similar to human behavior than that of the "subject" model.

Next, a computational-quantitative implementation of the "actor" strategy as weighted feature comparison between memory units was used to compare it to other memory-based models from the literature on the basis of EEG data. The "actor" strategy clearly provided the best model, showing a better global fit as well as better match in all details.

Building upon the success modeling EEG data, the feasibility of estimating free parameters from empirical data was demonstrated. Both the procedure for doing so and the necessary software were introduced and applied at the level of individual participants. Using empirically estimated parameters, the models from the previous EEG experiment were calculated again and yielded similar results, thus reinforcing the previous work.

In a final experiment, the feasibility of analyzing EEG data from a naturalistic auditory stimulus was demonstrated, which conventional wisdom says is not possible. The analysis suggested a new perspective on the nature of event-related potentials (ERPs), which does not contradict existing theory yet nonetheless goes against previous intuition. Using this new perspective as a basis, a preliminary attempt at a parsimonious neurocomputational theory of cognitive ERP components was developed.

Zusammenfassung in deutscher Sprache

Die Quantifizierung stellt für die Theoriebildung in der Neurolinguistik eine große Herausforderung und eine große Chance dar. Denn die Quantifizierung ist nötig für die Entwicklung von präzisen Theorien, jedoch verlangt sie neue Methoden und Perspektiven. Zu diesem Zweck wurden vier komplementäre Versuche im Rahmen des extended Argument Dependency Model von Bornkessel-Schlesewsky und Schlesewsky eingeführt.

Im Bereich der kognitiven Modellierung wurde ein computationelles Model des menschlichen Sprachverstehens anhand eines Dependency Parser entwickelt. Diese Modellierung dient dem Vergleich der Optimalität der Lösungen eines herkömmlichen Sprachverarbeitungsmodells (Subjekt-Strategie) und des eADMs (Actor-Strategie). Die Ergebnisse zeigten gewisse Vorteile für herkömmliche Modelle in einer eingeschränkten Umgebung. Dennoch zeigte sich das Modell für die Actor-Strategie in einem Testlauf als dem Menschen ähnlicher als das Modell für die Subjekt-Strategie.

Im Bereich der Elektrophysiologie wurde anhand von EEG-Daten eine computationell-quantitative Implementierung der Actor-Strategie als gewichteter Eigenschaftsvergleich zwischen Gedächtniseinheiten mit anderen gedächtnisbasierten Ansätzen aus der Literatur verglichen. Die Actor-Strategie erwies sich als die eindeutig beste Strategie mit einer besseren Anpassung sowohl in der globalen Form als auch in allen Einzelheiten.

Aufbauend auf dem erfolgreichen Modellierungsversuch mit EEG wurde die Umsetzbarkeit der empirischen Parameterschätzung für die freien Parameter im computationellen Modell gezeigt. Eine Vorgehensweise und zugehörige Software zur Schätzung auf Einzelversuchspersonenbasis wurden eingeführt. Anhand der empirisch geschätzten Parameter wurden die Modelle aus dem EEG-Experiment erneut berechnet. Die Ergebnisse aus dem vorherigen Versuch lassen sich mit den neuen Parametern bestätigen.

Im letzten Experiment wurde die Machbarkeit der Auswertung elektrophysiologischer Daten in einer natürlichen Umgebgung nachgewiesen, welche bisher als nicht machbar betrachtet wurde. Aus dieser Auswertung ergab sich eine neue Perspektive auf die Natur ereigniskorrelierter Potenziale (EKPs), die bestehenden Theorien nicht widersprach jedoch der gängigen Intuition. Aufbauend auf diesen Überlegungen wurde ein vorläufiger Ansatz für eine neurocomputationelle Theorie kognitiver EKP-Komponenten vorgeschlagen.

Curriculum Vitae (English)

Education

since 10.2010	Dr. phil. (PhD-equivalent) Linguistics, Philipps-Universität Marburg
2010	MA Linguistics of German, Philipps-Universität Marburg
2008	BS Mathematics Honors, University of Notre Dame du Lac
2008	BA German Language and Literature with Honors, University of Notre Dame du Lac

Publications

P. M. Alday (to appear). "Be Careful When Assuming the Obvious: Commentary on "The Placement of the head that minimizes online memory: a complex systems approach"". In: *Language Dynamics and Change* 5.1

P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (in press). "Discovering Prominence and its Role in Language Processing: An Individual (Differences) Approach". In: *Linguistic Vanguard*. DOI: 10.1515/lingvan-2014-1013

P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (2014). "Towards a Computational Model of Actor-based Language Comprehension". In: *Neuroinformatics* 12.1, pp. 143–179. DOI: 10.1007/s12021-013-9198-x

P. M. Alday (2012). "Province on a Hill: South Tyrol as a Mircocosm of European Federalism". In: *Ideas of | for Europe. An Interdisciplinary Approach to European Identity*. Ed. by T. Pinheiro, B. Cieszynska, and E. Franco. Frankfurt am Main: Peter Lang, pp. 271–284

Conference Presentations

P. M. Alday, J. Sassenhagen, and I. Bornkessel-Schlesewsky (2014c). *Tracking the Emergence of Meaning in the Brain during Natural Story Comprehension*. Poster presented at the International Conference on Cognitive Neuroscience (ICON). Brisbane

P. M. Alday, J. Sassenhagen, and I. Bornkessel-Schlesewsky (2014a). *Neural Signatures of Incremental Text Processing Correlate with Word Entropy in a Natural Story Context*. Poster presented at the Society for Neurobiology of Language Conference. Amsterdam

J. Sassenhagen and P. M. Alday (2014a). *Reliability of Gamma Activity during Semantic Integration*. Poster presented at the Society for the Neurobiology of Language Conference. San Diego

P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (2012a). *Towards a Computational Model of Actor-based Language Comprehension*. Poster presented at the Neurobiology of Language Conference. San Sebastian

P. M. Alday et al. (2011a). *Actor Identification in Natural Stories: Qualitative Distinctions in the Neural Bases of Actor-related Features*. Talk presented at the Neurobiology of Language Conference. Annapolis

Curriculum Vitae (Deutsch)

Ausbildung

seit 10.2010	Promotion im Fach Linguistik, Philipps-Universität Marburg
2010	MA Germanistische Linguistik, Philipps-Universität Marburg
2008	BS Mathematics Honors (theoretische Mathematik), University of Notre Dame du Lac
2008	BA German Language and Literature with Honors (Deusche Sprache und Literatur mit ausgezeichneter Abschlussarbeit), University of Notre Dame du Lac

Veröffentlichungen

P. M. Alday (to appear). "Be Careful When Assuming the Obvious: Commentary on "The Placement of the head that minimizes online memory: a complex systems approach"". In: *Language Dynamics and Change* 5.1

P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (in press). "Discovering Prominence and its Role in Language Processing: An Individual (Differences) Approach". In: *Linguistic Vanguard*. DOI: 10.1515/lingvan-2014-1013

P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (2014). "Towards a Computational Model of Actor-based Language Comprehension". In: *Neuroinformatics* 12.1, pp. 143–179. DOI: 10.1007/s12021-013-9198-x

P. M. Alday (2012). "Province on a Hill: South Tyrol as a Mircocosm of European Federalism". In: *Ideas of | for Europe. An Interdisciplinary Approach to European Identity*. Ed. by T. Pinheiro, B. Cieszynska, and E. Franco. Frankfurt am Main: Peter Lang, pp. 271–284

Tagungen

P. M. Alday, J. Sassenhagen, and I. Bornkessel-Schlesewsky (2014d). *Tracking the Emergence of Meaning in the Brain during Natural Story Comprehension*. Poster präsentiert auf der Interna-

tional Conference on Cognitive Neuroscience (ICON). Brisbane

P. M. Alday, J. Sassenhagen, and I. Bornkessel-Schlesewsky (2014b). *Neural Signatures of Incremental Text Processing Correlate with Word Entropy in a Natural Story Context*. Poster präsentiert auf der Society for Neurobiology of Language Conference. Amsterdam

J. Sassenhagen and P. M. Alday (2014b). *Reliability of Gamma Activity during Semantic Integration*. Poster präsentiert auf der Society for the Neurobiology of Language Conference. San Diego

P. M. Alday, M. Schlesewsky, and I. Bornkessel-Schlesewsky (2012b). *Towards a Computational Model of Actor-based Language Comprehension*. Poster präsentiert auf der Neurobiology of Language Conference. San Sebastian

P. M. Alday et al. (2011b). *Actor Identification in Natural Stories: Qualitative Distinctions in the Neural Bases of Actor-related Features*. Vortrag gehalten auf der Neurobiology of Language Conference. Annapolis