Lexical and Discourse Semantics in a Reading-Time Corpus of English

Jakub Dotlačil, Laia Colina Fortuny, Li Kloostra, Johan Bos
Utrecht University, Utrecht University, Utrecht University, University of Groningen
j.dotlacil@uu.nl, laiacolina@gmail.com, l.kloostra@uu.nl, johan.bos@rug.nl

Abstract

We present a novel language resource that combines a reading-time corpus, constructed in psycholinguistics, with rich lexical, compositional, and discourse meaning representation annotations. While existing psycholinguistic corpora typically provide morphological and syntactic annotations, no comparable corpora with comprehensive semantic information have been made available until now. We enriched the UCL corpus (361 sentences of self-paced reading, eye-tracking, and EEG data) with annotations in the style of the Parallel Meaning Bank (PMB) project, including WordNet synsets, VerbNet thematic roles, Combinatory Categorial Grammar (CCG) parses, and Discourse Representation Theory (DRT) structures. We demonstrate the utility of this resource through two case studies examining (1) encoding interference effects due to gender similarity and (2) integration costs in semantic role assignment. Both studies reveal processing patterns consistent with established psycholinguistic theories and/or previous findings. This resource fills a significant gap in psycholinguistic research, enabling the evaluation of semantic processing theories on naturalistic corpus data and extending the existing pool of annotated reading-time corpora. It should be useful to psycholinguists, as well as to cognitive scientists interested in language processing.

Keywords: reading-time corpus, semantic annotation, discourse representation theory, sentence processing

1. Introduction

In psycholinguistics, it has become an increasingly common practice to evaluate processing models not just on individual experiments, but also on reading-time corpora, which include neural and behavioral measures for texts. While even just access to neuro-behavioral measures for various text collections is extremely useful for studies on processing, a lot of progress was made in particular with corpora enriched with linguistic information. A case in point are syntactically parsed corpora, which played an important role in validating or refuting syntactic theories of processing (Boston et al., 2008; Demberg and Keller, 2008; Smith and Levy, 2013; Dotlačil, 2021; Shain et al., 2022; Isono, 2024).

In this work, we provide a novel resource for psycholinguistics: a reading-time corpus with rich semantic annotations. We chose to use the annotation developed in the Parallel Meaning Bank (PMB) project (Abzianidze et al., 2017), which provides information about segmentation, syntactic parsing, and, most importantly for us, detailed information about semantics from the word level to the discourse level. To the best of our knowledge, no reading-time corpus comparable in its scope to ours yet exists. This arguably affects the current state of affairs of theorizing on human sentence processing, which heavily leans on syntactic theories, lacking the data and models on which semantic processing theories could be evaluated. Providing a reading-time corpus with rich (discourse) semantic annotations, developed independently in formal semantics, should help close this gap.

2. Related work

There are several psycholinguistic corpora that combine text with behavioral and/or neural measurements representing the processing information of the texts by native or second-language readers. Often, such corpora also include annotations with psycholinguistic or linguistic information.

Arguably, the most well-known corpus with reading-time data is the Dundee corpus (Kennedy, 2003), which provides eye-tracking measures for 51,501 word tokens in English. The corpus was later enriched with syntactic annotation using dependency parsing (Barrett et al., 2015).

Another popular reading-time corpus is the Natural Stories Corpus (Futrell et al., 2016), which provides self-paced reading measures for 10,245 word tokens in English. The corpus also includes syntactic information. It is annotated with manually corrected parses of the Penn Treebank phrase structures and Universal dependencies. The Penn Treebank syntactic information is also present in portions of the Corpus of Eye Movements in L1 and L2 English Reading (CELER; Berzak et al., 2022), with 320,360 tokens, which provides eye-tracking measures of second-language learners on texts from the Wall Street Journal.

Other reading-time corpora, notably the University College of London (UCL) corpus (Frank

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9592–9601
11–16 May 2026.

et al., 2013), the Ghent Eye-Tracking Corpus (GECO; Cop et al., 2017), the Zurich Cognitive Language Processing Corpus (ZuCo; Hollenstein et al., 2018), and the Multilingual Eye-Movements Corpus (MECO; Siegelman et al., 2022), include some (psycho)linguistic information, like part-of-speech tags, frequency and readability metrics, among others, but lack full syntax-level, semantic, or discourse-level information.

There is currently no reading-time corpus that has been annotated with the rich information of the PMB. Even more to the point, we are not aware of any psycholinguistic corpus that has been annotated with linguistic information on lexical and discourse-level semantics. The current work thus presents an entirely novel contribution to the field of (computational) psycholinguistics.

3. Method

3.1. The reading-time corpus

From the existing reading-time corpora, we selected the UCL corpus for the PMB annotation. In this section, we present basic details of the corpus. For further details, consult Frank et al. (2013).

The UCL corpus consists of 361 British English sentences, with an average length of 13.7 word tokens per sentence. The sentences in the corpus come from free online unpublished novels written by aspiring authors, and were hand-picked so that they would be interpretable out of context. Every sentence is treated as independent of the others, that is, they do not form a coherent narrative. To ensure and strengthen this independence, proper names in the sentences were changed from the original texts so that no proper name would appear more than twice across all stimuli.

The UCL corpus includes several reading measurements. First, there are measures collected from a self-paced reading experiment with the central window paradigm. In this experiment, 117 participants received a random subset of the 361 sentences from the corpus. In the self-paced reading study, each sentence starts with a fixation cross appearing in the center of the screen. After the participant presses the space bar, the first word appears in the center of the screen, and every subsequent press of the space bar removes the current word and displays the next word in the same position. The key press time stamps are recorded and represent the reading time measure in this method.

In addition to self-paced reading, the UCL corpus includes eye-tracking reading data. These were collected from 43 participants, who each read all 205 sentences from the corpus in random order. Only those sentences that could fit on one line were used for this data collection.

Finally, the UCL corpus was later enriched with EEG measurements (Frank et al., 2015). These were collected from 24 participants, who read the same subset of 205 sentences that were presented in the eye-tracking experiment.

3.2. The annotation

The annotations of the UCL corpus were carried out within the PMB project (Abzianidze et al., 2017), which provides rich, formal meaning representations for all language levels ranging from words to texts. The PMB annotation pipeline consists of six main steps, which can be seen as applying in sequence, see Figure 1. Each step allows for manual corrections, which we discuss in more detail below.

First, the text is segmented into sentences and the sentences are segmented into tokens, where multi-word expressions representing constituents are treated as single tokens. Second, universal semantic tagging assigns language-neutral semantic tags (semtags) to tokens. The semantic tagging generalizes over part-of-speech (POS) and named entity classes and includes more specific semantic information than is commonly denoted in POS tagging. The tagset comprises 80 fine-grained semtags divided into 13 coarse-grained classes (Abzianidze and Bos, 2017).

The third step, symbolization, performs some lexical disambiguation and unifies lemmatization to produce inherently consistent symbols for the meaning representations. For example, the pronoun *he* would be specified as *male* in this step, and function words like the modal *can* would be specified as the diamond logical symbol. Fourth, syntactic parsing is performed using Combinatory Categorial Grammar (CCG), which provides a transparent syntax-semantic interface suitable for compositional semantics (Bos et al., 2004). The fifth step specifies senses based on WordNet (Fellbaum, 1998). Finally, the sixth step provides compositional semantic interpretation using Discourse Representation Theory, DRT (Kamp and Reyle, 1993). The system Boxer (Bos, 2015) assigns lexical Discourse Representation Structures (DRSs) to each token based on its CCG category, semtag, and symbol, and then compositionally constructs the DRS for the entire sentence using lambda calculus with continuations. The resulting DRSs include VerbNet thematic roles to designate relations between events and their participants and also capture anaphora and quantifier scope resolution.

During the annotation process, we first had the sentences automatically annotated, relying on the tokenizer elephant (Evang et al., 2013), the TNT tagger for semantic tagging (Brants, 2000), the EasyCCG parser for syntactic annotations (Lewis and Steedman, 2014), and Boxer for producing semantic representations. Afterwards, we manu-

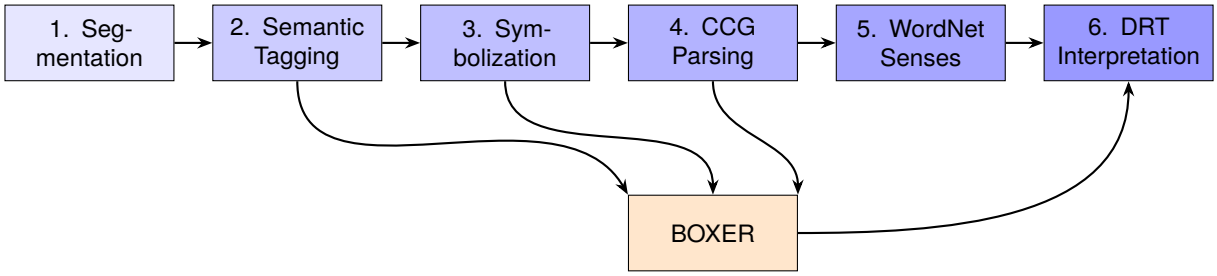

Figure 1: The PMB annotation pipeline consisting of six sequential steps. Boxer uses information from CCG parsing, semantic tagging, and symbolization to produce the DRT interpretation.

ally reviewed and corrected the annotations. This was always done by correcting the first five steps in the process. With the exception of very few cases, the resulting DRS is not corrected. This does not mean that the semantic discourse has not been manually changed—but whenever it was changed, this was done by changing the input that goes into Boxer (i.e., semantic tagging, symbolization, or CCG parse) with an eye to generating the correct DRT interpretation.

The manual control is highly labor-intensive for such a complex annotation of sentences, even with the help of automation tools. This is even more visible in the case of complex sentences, which are often employed in psycholinguistic corpora. The manual checks stretched over a period of several months. We split the work among us and four student annotators. Each annotator corrected the annotations they were sure about. Any unclear issues were brought up in joint meetings where we agreed on a solution. Two annotators at the end of the project walked through the whole corpus and checked all the sentences for any remaining discrepancies and inconsistencies. Any remaining issues were again discussed and resolved jointly in group meetings.

3.3. The created dataset

The annotations are accessible on the PMB explorer website: https://pmb.let.rug.nl/explorer, under the index of the document IDs 30/0349–99/0349, 00/0350–99/0350, 00/0351–38/0351, 51/0351–99/0351, 00/0352–99/0352, 00/0353–07/0353. The PMB representations merged with the UCL reading-time data (along with other psycholinguistic measures, such as frequencies) can be downloaded at the following link: `https://doi.org/10.17605/OSF.IO/PES63`.

An example of a fully annotated sentence is given in Figures 2 and 3, which show two meaning representations: a box notation (Figure 2) and a sequence notation (Figure 3). The former is used standardly in DRT, the latter has been developed in Bos (2023). Both carry the same information, but we find the sequence notation easier to read. Furthermore, the sequence notation allows for a straightforward connection of the discourse semantic information to reading data; hence, we focus on the latter.

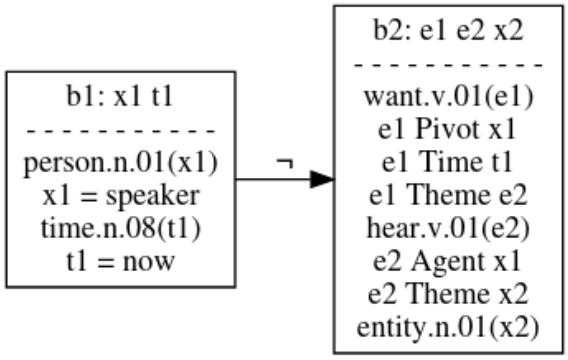

Figure 2: Box notation (standard DRT representation) for "I don't want to hear it."

The sequence notation in PMB consists of three columns. These are the first three columns in Figure 3. The first column specifies WordNet synsets (e.g., `person.n.01`) and constants, used with proper names, numerical values, dates, times, and deixis. The second column specifies roles like `Agent`, `Theme` (connected to the VerbNet database, see Kipper et al., 2008). The indices signal how many positions away the bearer of the role is. When the number is positive, the bearer of the role follows the role assigner; when it is negative, the bearer precedes the role assigner (e.g., -1 is the previous element; +1 is the following element); when the number is preceded by a greater-than or less-than sign, the bearer spans across the indicated elements. Capitalized words (like NEGATION) signal discourse relations (separators between DRSs), EQU is the equality statement. Some other common elements in the second column, not shown in this example, are ANA, which indicates the antecedent of an anaphor, and time relations, like TPR, which signals the tense interpretation as prior to now, i.e., past. The third column

shows the words that bring about the meaning presented in the previous columns. For more details on the notation, see Bos (2023) and references therein. For more details on DRT, see Kamp and Reyle (1993).

The last two columns in Figure 3 show how the sequence notation can be seamlessly combined with reaction-time data. We show two columns (more are present in the actual dataset): the first column provides the participant identification number, and the second column provides their reading times per word in milliseconds.[1]

Sometimes, multiple words correspond to one row in the sequence notation. This is particularly the case for idioms, phrasal verbs or verbs followed by an infinitive marker, like `want ~~to~~` in our example. Since the reading-time corpus reports measures per word, there are several options for how researchers could match the reading times to the semantic information. Only the reading times for the first full content word could be included. Alternatively, we could sum up over all the relevant words or take their mean. There might be other possibilities. The second and third options seem plausible, but they are complicated by eye-tracking measures, since region reading measures are standardly not calculated as a sum of by-word reading measures. For instance, the first pass RT over the region of two words, which measures the time spent on the two words until the eye fixation leaves the region, is not the same as the sum of the first pass RT on each of those words.

For this reason, and for simplicity of calculation, we currently opt for using only the RTs of the first content word in multiple-word cases.[2]

Inversely, it can also happen, albeit much more rarely, that PMB assigns meaning to a subpart of a word. Such an example is *don't*, whose meaning contribution is split across two rows: time specification, linked to *do*, and the negation separator, linked to *n't*. We currently leave sub-words unaligned with the reading-time data.

Crucially, along with the dataset, we provide Python scripts for merging the PMB annotation and the reading data from the UCL corpus. The scripts are included in the linked osf repository. Researchers can explore other options for matching reading times to multi-word and sub-word instances.

[1]This is the case for self-paced reading data. In the case of eye-tracking, we combine the sequence notation with the standard eye-tracking measures, as they are given in Frank et al. (2013). The measures are: first fixation RT, first pass RT, right bounded RT, and go-past RT. For the definition of the eye-tracking measures, see Frank et al. (2013). EEG data are, at the moment, not yet matched to the sequence notation.

[2]The content word is signaled in the PMB annotation: the words that do not contribute to synsets (like particles or infinitive markers) are crossed out. Thus, in multiple-word instances, we link the first non-crossed-out word to reading-time data.

4. Case studies

We now present two case studies, showcasing how the semantically annotated reading-time corpus can be used to bring novel evidence for particular claims and theories in psycholinguistics. As far as the case studies match previous findings in the research on processing, they can also be seen as a validation of the dataset.

4.1. Case study A: Encoding interference due to gender

Encoding interference refers to the processing difficulty that arises when similar representations compete in working memory (Oberauer and Kliegl, 2006). Research in psycholinguistics has explored whether semantic similarity between discourse referents leads to such interference effects during sentence processing. However, as far as we know, the effect of interference in processing has been solely established using data from individual experiments in which the interference is often manipulated in carefully constructed minimal pairs. Finding evidence in reading-time corpus data that was not designed for this purpose would significantly strengthen the theory. The semantically annotated corpus provides a good opportunity to directly test the hypothesis.

We exploit the fact that the corpus carries the lexical information from WordNet synsets. We explore one particular case: can we detect encoding interference between male-referent words?

We addressed the question in two specific analyses. First, we checked whether male-referent words (such as male proper names or nouns specifying stereotypical male professions) are harder to process if they appeared in a discourse that already contained one or more male referents. Second, we selected discourse contexts that introduced two or more entities and examined whether there were processing difficulties if the later introduced entity was a second male, compared to a case when it was not (because it was the first male or it was a female). We relied on the WordNet database for the male/female attribution to nouns.

For the analysis, we employed Bayesian hierarchical models with covariates for log-frequency, word position, and word length, along with maximal random effect structures for participants and word types. The dependent variables included right-bounded reading times and skipping rates on the target word from eye-tracking data. Skipping rates were coded as 0=no skipping, 1=skipping. From

Sequence notation			Reading-time info	
Synsets & constants	Roles, separators...	Words	Participants	RTs
person.n.01	EQU speaker	% I [0-1]	pp01	395
time.n.08	EQU now	% do~~n't~~ [2-7]	pp01	-
	NEGATION <1	%		
want.v.01	Pivot -2 Time -1 Theme +1	% want ~~to~~ [8-15]	pp01	377
hear.v.01	Agent -3 Theme +1	% hear [16-20]	pp01	362
entity.n.01		% it~~.~~ [21-24]	pp01	377

Figure 3: Sequence notation (SBN) for the sentence “I don’t want to hear it.”, word-aligned with the information from the reading-time corpus.

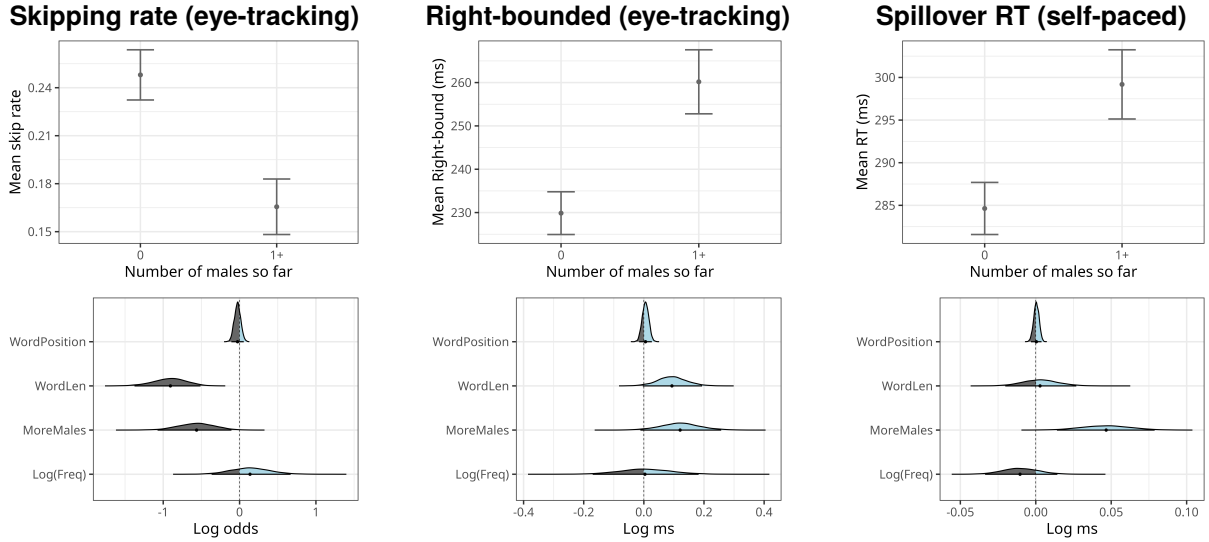

Figure 4: Descriptive summaries and posteriors for reading measures on words introducing a male (Case Study A). The top three graphs: means and standard errors per number of males introduced in the previous discourse. The bottom three graphs: posterior distributions. The thick horizontal lines represent 95% credible intervals. MoreMales encodes how many males appeared in the previous discourse, before the current male word (with two levels: either none or at least 1 male previously; a higher number of males appeared in only very few sentences, these instances were collapsed with the 1-male level). Left: the skipping rate of the target word. Middle: right-bounded reading times on the target word. Right: reading times on the spillover word.

self-paced reading data, we considered reading times on the spillover (the word following the gendered noun). For the skipping rate, we used the Bernoulli likelihood with a logit link function. For the reading time, we used a shifted log-normal likelihood. As is common in processing studies, we did not include first and last word reading times in the analysis, since these are often outliers, and we removed too short and too excessive reading times (shorter than 100 ms and longer than 10,000 ms). Details about the prior structure of the models and the sampling method are provided in the appendix.

Descriptive summaries and posterior distributions, summarizing the results for the manipulation and the covariates, are given in Figures 4 and 5. Both figures have descriptive summaries in the top row and posterior distributions in the bottom row. Figure 4 shows evidence of increased processing difficulty in eye-tracking and self-paced reading when additional males are introduced into the discourse. Skipping rates clearly decrease in eye-tracking, indicating that readers are less likely to skip words introducing male referents if other male referents are already present in the discourse. Furthermore, right-bounded reading times on the target male word in eye-tracking data and reading times on the spillover in self-paced reading increase if other males are already present. The effect of the greater number of males on the processing of next male word is also visible in the descriptive summaries in the same figure.

Figure 5 provides additional evidence showing that person words introducing a second male referent are harder to process than those introducing either females in the mixed-gender discourse, or males in discourses in which no other male is present. The effect is clear in the case of eye-tracking reading data, but not so for self-paced

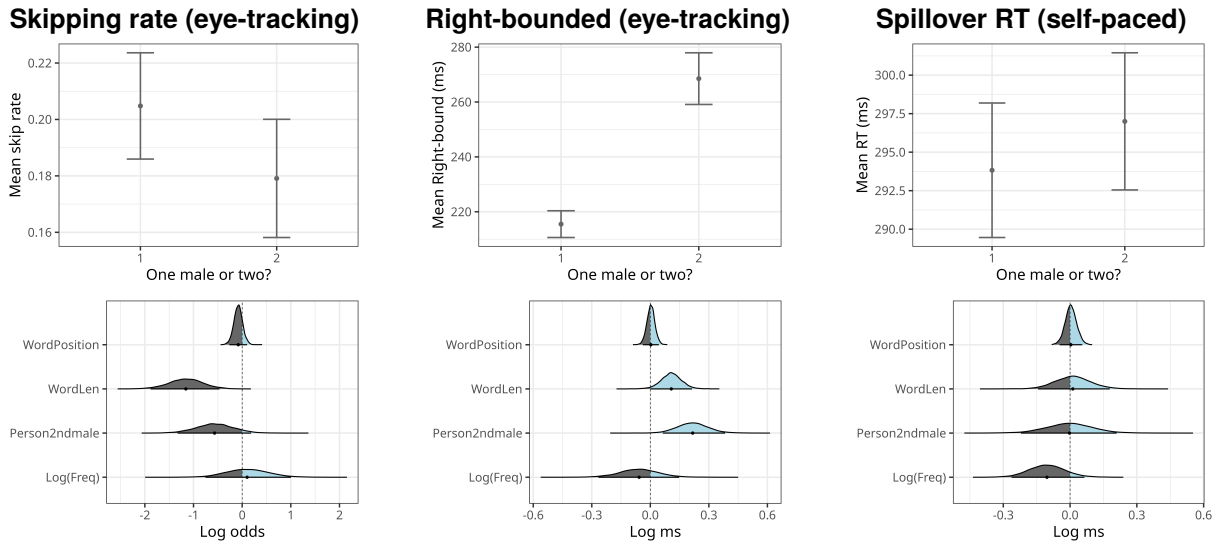

Figure 5: Descriptive summaries and posteriors for reading measures on words introducing a person (Case Study A). The top three graphs: means and standard errors per number of males introduced in the previous discourse. The bottom three graphs: posterior distributions. Person2ndMale codes whether the second entity word appears is the first male (the reference level) or it appears in a 2-male discourse. Left: skipping rates on the target word. Middle: right-bounded reading times on the target word. Right: reading times on the spillover word.

reading data, whose credible intervals for Person2ndmale spread around zero.

In sum, these results provide novel support for theories of encoding interference (Oberauer and Kliegl, 2006), demonstrating that semantic similarity between discourse referents leads to measurable processing costs. This is particularly true for eye-tracking data; self-paced reading data are less conclusive. Of course, the effect could be explored further: while we focus here on male referent nouns for illustration, the same pattern could be studied for female-gendered nouns, or persons vs. entities, etc. We leave such explorations, which could significantly strengthen the interference theory of processing, for future research.

4.2. Case study B: Integration cost

Dependency Locality Theory (Gibson, 2000) proposes that processing difficulty is correlated with the so-called integration cost. As the name suggests, the integration cost is a measure of how costly it is to integrate a role assigner with its arguments. The theory predicts that the integration cost grows with the distance, measured in the number of intervening elements, between a head (e.g., a verb) and its syntactic dependents. In our semantically annotated corpus, we can test whether similar integration costs arise when measuring the distance between a role assigner and its thematic arguments. Specifically, we hypothesized that if the distance between a role assigner and its most distant past argument grows, the role assigner should be harder to process. To be sure, this is not identical to the integration cost proposed in Dependency Locality Theory, but the measures should be correlated.

Consider the example sentence *I don't want to hear it* from Figure 3. The verb *want* assigns the Pivor role to *I* (indicated as `Pivot -2`, meaning that the role bearer is two positions back), while *hear* assigns the Agent role three positions back. Since the role assigner needs to integrate with an argument three positions back, this should potentially increase the processing difficulty.

We employed the same Bayesian hierarchical modeling approach as in Case Study A, with the studied effect of Integration Cost, measured as the distance between the role assigner and the most remote, previously introduced argument. The model included covariates for log-frequency, word position, and word length, along with maximal random effect structures. We analyzed right-bounded reading times from eye-tracking data. The prior structure and the sampling were the same as in Case Study A, see Appendix.

The results summarized in Figure 6 present an interesting pattern. The left panel shows that when all data are included, processing difficulty appears to *decrease* with increased distance—a surprising finding that contradicts the integration cost hypothesis. However, the right panel reveals that the negative effect disappears and, in fact, it trends towards more positive values when we exclude cases where an argument is immediately adjacent to its role assigner (that is, distance of 1 is removed). This is also visible in the descriptive summaries, shown in

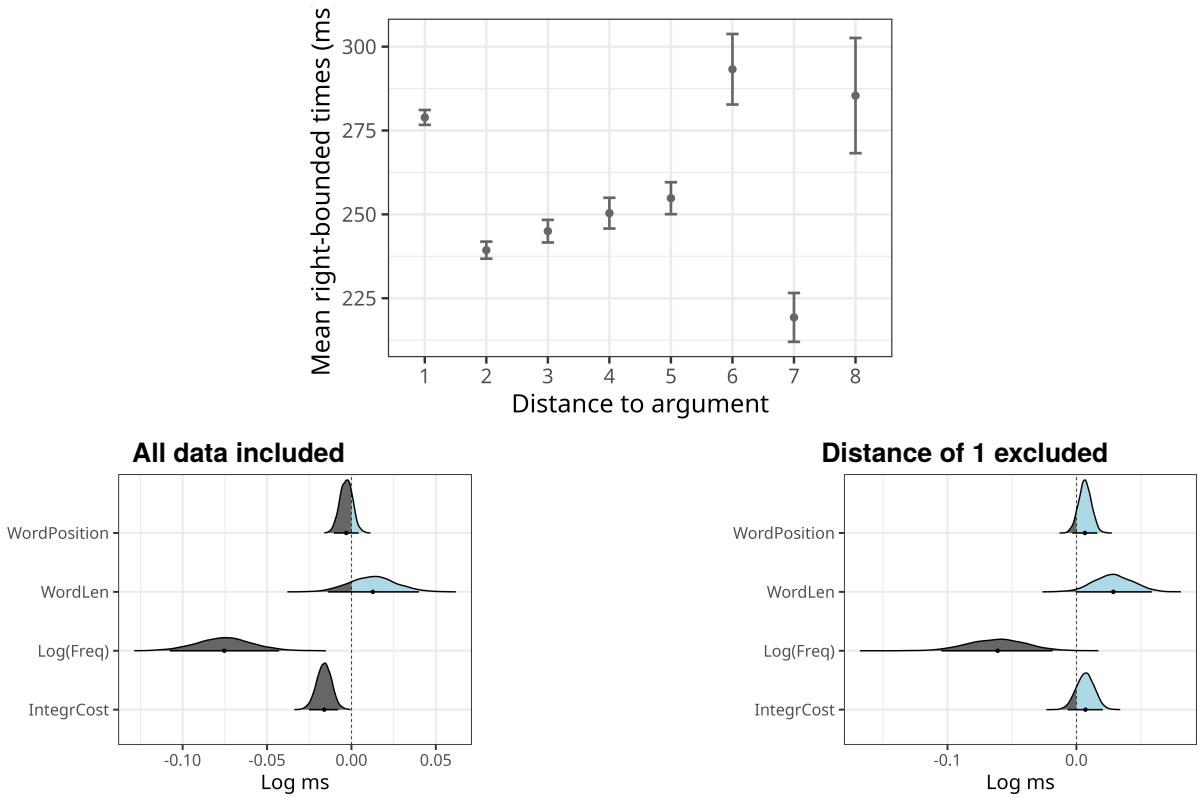

Figure 6: Descriptive summaries and posteriors for right-bounded reading times on words assigning thematic roles (Case Study B). The top graph shows means and standard errors of right-bounded reading times as a function of distance from the verb to the most remote argument. The bottom graphs show posteriors of two Bayesian models. IntegrCost measures the distance between the role assigner and its most remote argument. Left: All data included. Right: Cases where the most remote argument is adjacent (distance of 1) are excluded.

the top graph in Figure 6. That figure confirms that the distance of 1 is an outlier, and when we ignore it, we see a general trend of increase in reading times with the increase of distance – with the exception of very large distances, which, however, are based on few data, as visible in their larger standard errors. The pattern that we uncover here is reminiscent of a previous corpus study by Demberg and Keller (2008), who used syntactic dependency annotations and also observed negative effects due to integration cost on all data, which reversed or disappeared after the cases with zero integration cost were removed. Our results replicate this finding and, importantly, extend it to the semantic level.

5. Conclusions

We presented a novel language resource that combines reading-time measurements with rich semantic and discourse-level annotations. The resource is, to our knowledge, the first psycholinguistic corpus that provides such comprehensive semantic information, filling a significant gap in the field where existing corpora typically offer only morphological and syntactic annotations.

We demonstrated the utility of this resource through two case studies. Case Study A focused on exploring the processing cost of encoding interference due to gender similarity. Case Study B investigated integration costs in semantic role assignment, examining whether processing difficulty increases with the distance between a role assigner and its arguments.

These case studies served a dual purpose. First, they demonstrated the practical applications of the created dataset, showing how researchers can leverage the rich semantic annotations to test theories of language processing. Second, they tested whether the dataset exhibits expected behavioral patterns consistent with established psycholinguistic theories and/or previous findings. As far as the second goal is concerned, the demonstration confirmed the reliability and usefulness of the dataset for future research on semantic and discourse processing.

For Case Study A, we have seen that encoding interference, predicted by psycholinguistic theories and confirmed in reading experiments, can be observed in the annotated corpus. Needless to say, by looking at interference due to gender similarity,

we only touched the tip of the iceberg: the dataset could be used to explore other cases of interference.

For Case Study B, we have seen that the dataset does not provide direct evidence of integration cost in semantic role assignment. While surprising, our findings go in line with past corpus research, done on syntactic annotations. Just as in Case Study A, future research on the annotated corpus could further explore the state of affairs, for instance, by considering other ways of calculating integration costs.

We hope that in the future, the annotated corpus will be useful for semantic research in psycholinguistics beyond the presented studies. Ideally, the corpus should be of use not just to confirm existing and well-established theories, but also to generate and test novel, semantic-based hypotheses about sentence processing.

6. Acknowledgements

We would like to thank the student annotators involved in the project who annotated and cleaned the dataset together with us. These were: Bram Emanuel Pramono, Abel Koffeman, Florence Liu and Madalina Zgreaban. We would also like to thank the technical support of the Parallel Meaning Bank project, especially Xiao Zhang. The research reported in this paper was supported by the European Research Council (ERC), grant 101088098 - MEMLANG. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

7. Bibliographical References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik Van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 242–247, Valencia, Spain.

Lasha Abzianidze and Johan Bos. 2017. Towards universal semantic tagging. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017) – Short Papers*, pages 1–6, Montpellier, France.

Maria Barrett, Zeljko Agic, and Anders Søgaard. 2015. The dundee treebank. In *The 14th international workshop on treebanks and linguistic theories (TLT 14)*, pages 242–248.

Johan Bos. 2015. Open-domain semantic parsing with boxer. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 301–304.

Johan Bos. 2023. The sequence notation: Catching complex meanings in simple graphs. In *15th International Conference on Computational Semantics*, pages 195–208. Association for Computational Linguistics (ACL).

Johan Bos, Stephen Clark, Mark Steedman, James R Curran, and Julia Hockenmaier. 2004. Wide-coverage semantic representations from a ccg parser. In *COLING 2004: Proceedings of the 20th international conference on computational linguistics*, pages 1240–1246.

Marisa Ferrara Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2:1–12.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231, Seattle, Washington, USA. Association for Computational Linguistics.

Paul-Christian Bürkner. 2021. Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Jakub Dotlačil. 2021. Parsing as a cue-based retrieval model. *Cognitive science*, 45:e13020.

Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. 2013. Elephant: Sequence labeling for word and sentence segmentation. In *EMNLP 2013*, pages 1422–1426.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain*, pages 95–126. MIT Press.

Shinnosuke Isono. 2024. Category locality theory: A unified account of locality effects in sentence comprehension. *Cognition*, 247:105766.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic. Introduction to Model theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.

Mike Lewis and Mark Steedman. 2014. A* CCG parsing with a supertag-factored model. In *2014 Conference on Empirical Methods in Natural Language Processing*, pages 990–1000. Association for Computational Linguistics.

Bruno Nicenboim, Daniel J Schad, and Shravan Vasishth. 2025. *Introduction to Bayesian data analysis for cognitive science*. CRC Press.

Klaus Oberauer and Reinhold Kliegl. 2006. A formal model of capacity limits in working memory. *Journal of Memory and Language*, 55(4):601–626.

Cory Shain, Idan Asher Blank, Evelina Fedorenko, Edward Gibson, and William Schuler. 2022. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *Journal of Neuroscience*, 42(39):7412–7430.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

8. Language Resource References

Berzak, Yevgeni and Nakamura, Chie and Smith, Amelia and Weng, Emily and Katz, Boris and Flynn, Suzanne and Levy, Roger. 2022. *CELER: A 365-participant corpus of eye movements in L1 and L2 English reading*. MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA. PID https://doi.org/10.1162/opmi_a_00054.

Cop, Uschi and Dirix, Nicolas and Drieghe, Denis and Duyck, Wouter. 2017. *Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading*. Springer. PID https://doi.org/10.3758/s13428-016-0734-0.

Frank, Stefan L. and Monsalve, Irene Fernandez and Thompson, Robin L. and Vigliocco, Gabriella. 2013. *Reading time data for evaluating broad-coverage models of English sentence processing*. PID https://doi.org/10.3758/s13428-012-0313-y.

Frank, Stefan L and Otten, Leun J and Galli, Giulia and Vigliocco, Gabriella. 2015. *The ERP response to the amount of information conveyed by words in sentences*. Elsevier. PID https://doi.org/10.1016/j.bandl.2014.10.006.

Futrell, Richard and Gibson, Edward and Tily, Harry J. and Blank, Idan and Vishnevetsky, Anastasia and Piantadosi, Steven T. and Fedorenko, Evelina. 2016. *The Natural Stories Corpus*. PID http://github.com/languageMIT/naturalstories.

Hollenstein, Nora and Rotsztejn, Jonathan and Troendle, Marius and Pedroni, Andreas and Zhang, Ce and Langer, Nicolas. 2018. *ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading*. Nature Publishing Group. PID https://doi.org/10.1038/sdata.2018.291.

Kennedy, Alan. 2003. *The Dundee Corpus*. The University of Dundee, Psychology Department. CD-ROM.

Siegelman, Noam and Schroeder, Sascha and Acartürk, Cengiz and Ahn, Hee-Don and Alexeeva, Svetlana and Amenta, Simona and Bertram, Raymond and Bonandrini, Rolando and Brysbaert, Marc and Chernova, Daria and others. 2022. *Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO)*. Springer. PID https://doi.org/10.3758/s13428-021-01772-6.

9. Optional Supplementary Materials: Appendices, Software, and Data

9.1. Appendices

9.1.1. Appendix A. Prior structure of the Bayesian models

The priors in the model were specified as follows.

For reading-time data: the intercept was assumed to come from a normal distribution ($\mu = 0, \sigma = 10$). The slopes for the fixed effects were set to be a normal distribution with the parameters $\mu = 0, \sigma = 1$, the standard deviation of the random effects was a truncated normal distribution ($\mu = 0, \sigma = 1$), and we used the LKJ distribution with $\eta = 2$ for the correlation between random effects.

For skipping-rate data: the intercept was assumed to come from a normal distribution ($\mu = 0, \sigma = 2$). The slopes for the fixed effects were set to be a normal distribution with the parameters $\mu = 0, \sigma = 1$, the standard deviation of the random effects was a truncated normal distribution ($\mu = 0, \sigma = 1$), and we used the LKJ distribution

with $\eta = 2$ for the correlation between random effects.

Both models had the maximal random-effect structure for subjects and word types.

The priors used are so-called principled priors, commonly used in psycholinguistics (see Nicenboim et al., 2025 for a discussion on priors in the domain of cognitive science and psycholinguistics).

9.2. Appendix B. Sampling and convergence of the Bayesian models

We ran the Bayesian model in R, using the `brms` package (Bürkner, 2021), which interfaces the probabilistic programming language `Stan`. The models were sampled from 4 chains, with 2,500 iterations per chain, 1,250 for warm-up. All Rhat values were below 1.02, indicating chain convergence.

9.3. Software and Data

The PMB representations merged with the UCL reading-time data (along with other psycholinguistic measures, such as frequencies) are in the supplementary material. They can also be downloaded at the following link: https://doi.org/10.17605/OSF.IO/PES63. Using the osf link is preferable – it includes the most up-to-date version of the data.

The annotations are also accessible on the PMB explorer website: https://pmb.let.rug.nl/explorer. They can be found under the index of the document IDs 30/0349–99/0349, 00/0350–99/0350, 00/0351–38/0351, 51/0351–99/0351, 00/0352–99/0352, 00/0353–07/0353.