Revising with a Backward Glance: Regressions and Skips during Reading as Cognitive Signals for Revision Policies in Incremental Processing

Brielen Madureira[1] **Pelin Çelikkol**[1] **David Schlangen**[1,2]
[1]Computational Linguistics, Department of Linguistics
University of Potsdam, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
{madureiralasota,aynur.celikkol,david.schlangen}@uni-potsdam.de

Abstract

In NLP, incremental processors produce output in instalments, based on incoming prefixes of the linguistic input. Some tokens trigger revisions, causing edits to the output hypothesis, but little is known about why models revise when they revise. A policy that detects the time steps where revisions should happen can improve efficiency. Still, retrieving a suitable signal to train a revision policy is an open problem, since it is not naturally available in datasets. In this work, we investigate the appropriateness of regressions and skips in human reading eye-tracking data as signals to inform revision policies in incremental sequence labelling. Using generalised mixed-effects models, we find that the probability of regressions and skips by humans can potentially serve as useful predictors for revisions in BiLSTMs and Transformer models, with consistent results for various languages.

1 Introduction

"*Supreme court plans an attack on independent judiciary, says Labour*." This was the headline of a news article,[1] which sounds incongruous until one interprets it the way intended. That is a *crash blossom*,[2] a sentence that becomes ambiguous *e.g.* due to brevity. The correspondent later *revised* the headline to remove the ambiguity. You probably had to go back and read that sentence again. Such movement is called *regression* in the eye-tracking literature, when the eye makes a regressive, as opposed to progressive, saccade while reading a text.

In incremental NLP models, partial output hypotheses are built at each time step, based on incoming input prefixes, which renders revisability a desirable property to correct mistakes (Schlangen and Skantze, 2011). This mode takes place in interactive settings that require real-time processing, for instance disfluency detecion or reference resolution in dialogue (Hough and Schlangen, 2015; Kennington and Schlangen, 2017) and simultaneous translation (Cho and Esipova, 2016; Arivazhagan et al., 2020; Sen et al., 2023).

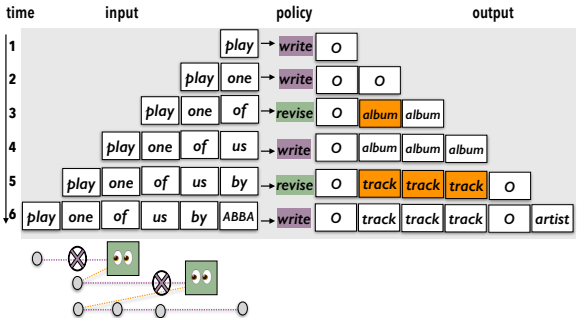

Figure 1: A constructed example of incremental sequence labelling where revisions occur at time steps 3 and 5. If tokens where humans initiate regressions in reading align with tokens that trigger revisions, it can be a cognitive signal to model a revision policy.

Figure 1 depicts a constructed example for sequence labelling. For each new token, the model either just extends the current output prefix with a new label, or also edits the output by changing previous labels (here at time steps 3 and 5). Modelling a policy that predicts when revisions should occur is an open research problem, because this signal is not naturally available in the training data (Köhn, 2018; Kahardipraja et al., 2023). Moreover, we currently lack evaluation methods to understand whether the revisions performed by a model are linguistically or cognitively motivated (*i.e.* being grounded in the linguistic input or resembling cognitive processes) or an idiosyncratic result of its internal processing patterns.

In eye-tracking experiments, many measures can be extracted per token while humans read texts (Rayner, 1998). Common data formats include variables representing whether each token, in first-pass reading, was skipped, fixated and left progressively

[1]Source: The Guardian, Nov 15, 2020. Retrieved from the Language Log blog.
[2]https://en.wiktionary.org/wiki/crash_blossom

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 335–351
December 6–7, 2023.

or triggered a regressive eye movement. In Figure 1, the constructed scanpath shows regressions at tokens *of* and *by* and skips at *one* and *us*. Various theories exist to account for why humans regress (see §3), but the fact that underlying cognitive processes cause the eyes to move forward or backward at each word (or skip it) lends itself as a cognitively motivated token-level signal.

In this paper, we bridge the concepts of *revisions* in incremental sequence labelling and *regressions* in human eye-tracking reading data. We investigate whether regressions and skips can aid the prediction of revisions in incremental processors, and conclude that eye-tracking measures are a potential cognitively-motivated learning signal to model revision policies.

2 Motivation

Currently on-trend models like Bi-LSTMs (Schuster and Paliwal, 1997) and Transformers (Vaswani et al., 2017) operate in a non-incremental fashion, relying on the availability of complete input sentences or texts to deliver output. One workaround to employ non-incremental encoders in real-time is applying a restart-incremental interface (Schlangen and Skantze, 2011), enabling outputs to be revised as a by-product of recomputations, as explored by Madureira and Schlangen (2020) and Kahardipraja et al. (2021). Although possible, it forces recomputation from scratch at every new piece of input, which increases the computational load and can become infeasible for long sequences (Kahardipraja et al., 2021). On the other hand, inherently incremental models like RNNs have the disadvantage of not being able to recover from mistakes via revisions (at least their prototypical versions).

The sweet spot would be a model that can detect the need to revise. Initiatives in this direction are HEAR (Kaushal et al., 2023), which has a module that predicts the need to *restart*, and TAPIR (Kahardipraja et al., 2023), which integrates an RNN with a Transformer-revisor, predicting whether to *recompute* or to just extend the current output. A difficulty encountered in the latter is how to obtain a ground-truth signal for the revision policy. They derived silver labels from the outputs of another Transformer, which is possibly too model-specific and its linguistic motivation is not explored. HEAR compares partial outputs to the non-incremental gold standard which, however, does not encode locally valid hypotheses (which only future input will rule out) and does not accommodate the fact that the gold standard may differ from its final output, thus penalising the incremental metrics with the model's non-incremental deficits (Baumann et al., 2011; Madureira et al., 2023).

We usually do not have corpora containing annotation for the incremental hypotheses for input prefixes by humans, only the annotated gold labels for the final output. But there is vast literature using human reading data as a supervision signal in NLP tasks (Barrett and Hollenstein, 2020; Mathias et al., 2021). Inspired by that, we ask ourselves whether a model's revisions coincide with human regressions in eye-tracking reading data. A positive answer would mean that human reading data could help modelling a dedicated policy for revisions (as opposed to naive recomputations or restarts), and would serve as a cognitively motivated yardstick to judge a models' revisions.

Among all revisions, some are *effective*, *i.e.* they edit the prefix into a better state, with respect to a gold standard or to the final output (Madureira et al., 2023). Identifying them can contribute to reducing undesired revisions, which cause instability without bringing the advantage of improvement in output quality. Therefore, if human reading behaviour can help perform only effective revisions, the signal is even more useful for incremental processing.

3 Related Literature

During reading, humans fixate the gaze on some words and make saccades that can be progressive or regressive with respect to the order of the words in the text, so that scanpaths and various measures regarding gaze position, direction and duration can be extracted with eye-tracking devices (Rayner et al., 2012), a technique that is becoming more accessible at scale (Ribeiro et al., 2023).

Research based on eye-tracking reading data often rely on the eye-mind hypothesis, which assumes that the eye remains fixated on a word as long as it is being processed (Just and Carpenter, 1980). Various research fields rely on the temporal and spatial dimensions of human reading data. We identify at least three (non-mutually exclusive) uses. A consolidated line of research involves studying human cognition and verifying linguistic theories of sentence processing (*e.g.* Demberg and Keller (2008) and Shain et al. (2016)). Another field is occupied with understanding to what extent com-

putational models like artificial neural networks resemble human cognition in how they process language, for example by estimating their psychometric predictive power (Wilcox et al., 2020; Hollenstein et al., 2021). A relationship commonly investigated is the surprisal of language models *versus* human reading time (Fernandez Monsalve et al., 2012; Goodkind and Bicknell, 2018; Wilcox et al., 2020). NLP has been incorporating eye-tracking data in recent years (Iida et al., 2013; Tokunaga et al., 2017), with the emerging use of human reading data both as input to enhance NLP models (see Barrett and Hollenstein (2020) and Mathias et al. (2021) for recent surveys) and as a means for their interpretability (Ikhwantri et al., 2023).

In this work, the phenomenon of interest is *regressions*, *i.e.* eye movements that move backwards in the text and can be shorter or longer-range (Rayner et al., 2012). They are a common topic in psycholinguistics research (Paape et al., 2022, 2021) and various hypotheses account for their role, such as comprehension or word identification difficulties, low-level visuomotor processes, rereading, memory cues and tools for language processing (see Vitu (2005), Lopopolo et al. (2019) and Booth and Weger (2013) for comprehensive discussions and references). Relevant measures are at which word a regression initiates, at which word it lands, regression path duration (how long the reader remains in past text before progressing to unseen text), and how many regressions are initiated for each word. We can also differentiate between first-pass and subsequent regressions.

Regressions in NLP Reading data has been used as a source of psycholinguistic information for various NLP tasks. When it comes to regressions, Barrett and Søgaard (2015a) used eye-movements to predict syntactic categories, an idea further explored in Barrett et al. (2016), who augmented PoS-taggers with various gaze features, among which was the number of regressions originating from a word. Barrett and Søgaard (2015b) used the number of regressions from and to a word as features to predict grammatical functions. The number of total regressions per word was also used as a feature by Mishra et al. (2016) for sarcasm understandability prediction. Regression duration, *i.e.* the total time spent on a word after the first pass over it, was a useful feature for sentence compression proposed by Klerke et al. (2016). Regressions during coreference resolution annotation were investigated by Cheri et al. (2016), who used it to propose a heuristic for pruning candidates in a coreference resolution model. In Hollenstein and Zhang (2019), the total duration of regressions from a word was used as a context feature in named-entity recognition.

We draw inspiration from the work by Lopopolo et al. (2019), who hypothesised that backward saccades are involved in online syntactic analysis, in which case regressions should coincide, at least partially, with the edges of the relations computed by a dependency parser. They found a significant effect of the number of left-hand side dependency relations on the number of backward saccades. While the authors were interested at predicting human regressions from a model instantiating a parsing theory, we are conversely interested in using human regressions as a signal to train an NLP model.[3]

4 Method

To perform the analysis, we use binomial generalised linear mixed models (GLMM) with a logit link function to predict model revisions. Similar to the approach by Lopopolo et al. (2019), for each combination of dataset and NLP model/task, we fit two GLMMs: The baseline model (1) only includes the token position variable as a fixed effect and texts as random effects. Since a model's revisions may vary depending on the word's position in the text, we add token position as a baseline predictor and include texts to account for any variability due to different types of texts. We fit model (2) with the same structure, adding the predictors of regression probability and skipping probability as fixed effects. The binary dependent variable is a token's revise/not-revise label.

$$\begin{aligned} model\ revision \sim\ & token\ position \\ & + (1|text) \end{aligned} \tag{1}$$

$$\begin{aligned} model\ revision \sim\ & token\ position \\ & + p(regression) \\ & + p(skip) \\ & + (1|text) \end{aligned} \tag{2}$$

We use likelihood ratio tests (LRT) between the null and the full models to evaluate the goodness of fit. LRTs are used to compare a baseline model to

[3] It is also worth investigating whether a model's revisions can predict human regression behaviour, but it is beyond the scope of this work.

(...)	That	night	there	was	scarcely	a	square	inch	of	earth	that	was	not	illuminated	by	aurora.
model	/	r	r	r	/	/	r	r	r	r	/	r	r	r	/	r
subject 1	-	👀	0	0	0	-	0	0	-	0	0	-	0	0	-	0
subject 2	👀	-	0	-	-	-	👀	👀	-	0	0	-	0	0	0	👀
subject 3	-	👀	👀	-	-	-	-	👀	-	-	-	0	-	👀	0	👀
subject 4	-	0	-	-	0	-	0	-	-	0	-	0	-	0	👀	👀
subject 5	-	0	0	-	0	-	0	-	-	0	-	-	0	0	-	👀

Figure 2: An example of our data structure for a portion of a text in the Provo corpus, processed by a restart-incremental Transformer predicting dependency relations. Each token is annotated with the reading variable for each subject (eyes: regressed, 0: not regressed, -: skipped) and the model's decision (r: revised, /: not revised).

a more complex one with more predictors and decide if certain predictors should be included, consequently selecting the model that fits the data better. To infer statistical significance, we obtain p-values using the χ^2 distribution.

We do not intend to make claims about *why* regressions occur. For our purposes, we take at face value that they *did* occur in the eye-tracking experiments (and when). We are interested in words at which regressions are initiated when they are first read, knowing that, for some reason, the reader went to past input before continuing (as a consequence, we also analyse words that are not fixated in the first pass). Still, the hypothesis that regressions occur due to reanalysis, when humans encounter garden path sentences (Altmann et al., 1992), is at our favour, since revisions represent updates in the current model's interpretation caused by input seen for the first time.

5 Data

In this section, we explain the data structure constructed for the analysis. We then introduce the eye-tracking corpora and the models selected for this study, and discuss how we extract the incremental outputs from non-incremental, pre-trained sequence labelling models.[4]

Procedure Our method requires knowing, for each token w in a text, what was the behaviour of the model while performing sequence labelling and of the humans while reading the text. More specifically, we need to know whether the model revised its hypothesis upon processing w and whether humans skipped w, fixated it but moved forward, or fixated it and regressed. We thus construct an annotation mapping tokens to human and model data as illustrated in Figure 2. The texts come from the eye-tracking corpora, from which we also extract the human skips or regressions. The revisions are retrieved by feeding the same texts to the NLP models, prefix by prefix in a restart-incremental fashion, and checking if labels change at each time step.

	language	tokens	texts	subjects
MECO-L1	Dutch	2,231	12	45
MECO-L2	English (L2)	1,658	12	538
Nicenboim	Spanish	791	48	71
PoTeC	German	1,895	12	62
Provo	English	2,743	55	84
RastrOS	Br. Portuguese	2,494	50	37

Table 1: Human reading eye-tracking corpora.

Human Regressions We analyse six eye-tracking human reading corpora: MECO-L1 (Siegelman et al., 2022), MECO-L2 (Kuperman et al., 2023), Nicenboim (no official name) (Nicenboim et al., 2015), PoTeC (Makowski et al., 2019; Jäger et al., 2020), Provo (Luke and Christianson, 2018) and RastrOS (Vieira, 2020; Leal et al., 2022). Table 1 presents their language and size. The distribution of regressions and skips (per token and per subject) is shown in Figure 3. Although many other corpora exist, we opted to use those that had first-pass regression and first-pass skip measures already available or easy to infer from other measures. For each interest area,[5] we retrieve the label for each subject as follows: If the token was skipped in the first-pass reading, we label it as skipped. Otherwise, we retrieve a variable which is 1 if a first-pass

[4]The pre-processing scripts and implementation code is available at https://github.com/briemadu/revreg.

[5]An interest area sometimes includes more than one token, *e.g* a word and punctuation, like *aurora.* in Figure 2.

		MECO (du)		MECO (enl2)		Nicenboim (es)		PoTeC (de)		Provo (en)		RastrOS (ptbr)	
		all-r	eff-r	all-r	eff-r	all-r	eff-r	all-r	eff-r	all-r	eff-r	all-r	eff-r
BiLSTM	deprel	58.45	47.20	60.74	54.52	55.75	50.32	53.56	44.27	60.99	53.70	54.01	46.75
	head	65.76	38.32	66.95	38.60	61.31	43.36	67.28	40.37	67.92	39.30	60.34	43.70
	pos	12.95	11.52	11.70	10.68	6.32	5.44	17.89	15.51	12.65	11.27	29.19	27.11
Transformer	deprel	63.92	52.44	67.97	57.66	48.93	44.37	73.67	56.36	66.68	58.77	52.81	44.23
	head	67.55	38.01	69.06	37.21	57.27	41.47	74.56	43.38	69.30	38.46	61.39	42.98
	pos	9.82	6.28	7.84	6.09	1.90	1.64	5.01	4.12	8.09	6.56	9.22	7.62

Table 2: % of timesteps that trigger revisions (all-r) and effective revisions (eff-r) for each model and task.

regression was initiated at that interest area, and 0 otherwise. Although regressions can occur later, we only consider what happens in the first-pass reading to approximate what the model does (revisions happen when a token is integrated for the first time in the sequence). The probabilities are estimated by computing the proportion of regressions and skips per token (excluding subjects with missing data), following existing literature in terms of using average human behaviour as a feature (Barrett et al., 2016; Hollenstein and Zhang, 2019). We checked that they are only moderately (negatively) correlated ($-0.59 < \rho < -0.44$). See Appendix for details about the measures and pre-processing.

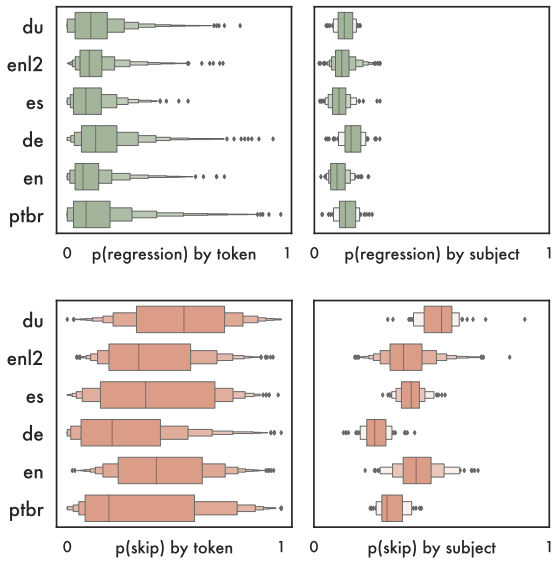

Figure 3: Distributions of the probabilities of regression and skips, by token (left) and by subject (right) estimated from the human reading data for each dataset.

Models' Revisions We opt to evaluate pre-trained sequence labelling models with a restart-incremental paradigm. Models were selected according to the availability of languages to match the eye-tracking corpora. We evaluate Stanza's BiLSTM models (Qi et al., 2020)[6] and Explosion's pre-trained multi-task Transformer architectures.[7] These families of models were selected due to the availability of all languages and comparability in terms of similar training data, as both were trained on the Universal Dependencies corpora (de Marneffe et al., 2021). The model checkpoints for each language are listed in Table 3. We extract the incremental outputs for dependency parsing (prediction of the head position and the relation) and POS-tagging. We also inspected NER, but revisions were extremely sparse in these datasets (possibly due to the genres of the texts), so we did not analyse it further. The same texts from the eye-tracking data are fed to each model, one prefix after another, as illustrated in Figure 1, following previous works (Madureira and Schlangen, 2020; Kahardipraja et al., 2021). At each time step, we extend the input with one interest area (*i.e.*, sometimes it means more than one token). If the output prefix at time t (apart from the recently added label(s), which refer to the last interest area) differs from the output at time $t-1$, a revision occurred. If more labels match the final output than in the previous prefix, the revision is effective. The percentage of (effective) revisions over tokens/timesteps is shown in Table 2.

	Explosion	Stanza
MECO-L1	`nl_udv25_dutchalpino_trf`	nl
MECO-L2	`en_udv25_englishewt_trf`	en
Nicenboim	`es_udv25_spanishancora_trf`	es
PoTeC	`de_udv25_germanhdt_trf`	de
Provo	`en_udv25_englishewt_trf`	en
RastrOS	`pt_udv25_portuguesebosque_trf`	pt

Table 3: Specification of the pre-trained NLP models.

[6] https://github.com/stanfordnlp/stanza.

[7] Release documented in https://explosion.ai/blog/ud-benchmarks-v3-2 and available at their model hub on Hugging Face https://huggingface.co/explosion.

			estimate		SE		z		p	
			MECO-L2	Provo	MECO-L2	Provo	MECO-L2	Provo	MECO-L2	Provo
BiLSTM	deprel	intercept	1.29***	1.22***	0.05	0.05	24.18	24.29	<0.001	<0.001
		p(reg)	3.41***	3.30***	0.05	0.09	73.39	38.56	<0.001	<0.001
		p(skip)	-2.80***	-3.68***	0.02	0.03	-178.47	-133.52	<0.001	<0.001
		position	-0.03***	0.21***	0.00	0.01	-8.94	38.87	<0.001	<0.001
	head	intercept	1.59***	1.76***	0.06	0.05	27.44	33.12	<0.001	<0.001
		p(reg)	4.32***	2.18***	0.05	0.10	81.05	21.84	<0.001	<0.001
		p(skip)	-3.23***	-4.92***	0.02	0.03	-193.35	-155.18	<0.001	<0.001
		position	-	0.40***	-	0.01	-	68.85	-	<0.001
	pos	intercept	-2.62***	-1.92***	0.07	0.08	-36.21	-22.77	<0.001	<0.001
		p(reg)	1.25***	1.42***	0.05	0.08	27.53	18.61	<0.001	<0.001
		p(skip)	-1.16***	-0.66***	0.02	0.04	-52.26	-18.63	<0.001	<0.001
		position	0.20***	-	0.00	-	42.18	-	<0.001	-
Transformer	deprel	intercept	1.22***	1.28***	0.09	0.05	14.28	24.39	<0.001	<0.001
		p(reg)	4.39***	3.26***	0.05	0.09	82.91	34.39	<0.001	<0.001
		p(skip)	-2.53***	-3.75***	0.02	0.03	-154.71	-129.34	<0.001	<0.001
		position	0.03***	0.30***	0.00	0.01	11.37	54.95	<0.001	<0.001
	head	intercept	1.45***	1.45***	0.08	0.05	18.13	29.17	<0.001	<0.001
		p(reg)	4.40***	2.27***	0.05	0.10	82.24	23.76	<0.001	<0.001
		p(skip)	-2.64***	-4.01***	0.02	0.03	-160.14	-133.24	<0.001	<0.001
		position	-	0.37***	-	0.01	-	64.92	-	<0.001
	pos	intercept	-2.64***	-2.69***	0.17	0.14	-15.28	-19.71	<0.001	<0.001
		p(reg)	-0.62***	3.00***	0.06	0.10	-9.49	31.11	<0.001	<0.001
		p(skip)	-0.77***	0.80***	0.03	0.04	-29.33	18.07	<0.001	<0.001
		position	0.08***	-0.25***	0.01	0.01	15.56	-30.18	<0.001	<0.001

Table 4: Overview of the GLMM results, showing the estimated coefficients for each variable and their statistical significance, for the English corpora. See Appendix for the the complete table.

6 Results

We summarise the full GLMM results in Table 4 for Provo and MECO-L2 datasets. Due to a large number of experiments, we only present results for the English models in this table; the complete results are in the Appendix. In every (dataset, NLP model, task) combination, the likelihood ratio test between the baseline and full models revealed that the full model, including the two predictors of interest, is a better fit to the data than the baseline model with only token position and text.

The token position was a significant predictor of revisions in most models. For the few cases in which it did not significantly affect revisions (*i.e.*, MECO-L2-Transformer-head and BiLSTM-head, MECO-L1-BiLSTM-head, Provo-BiLSTM-pos), we fitted models without this predictor instead.

We found that average human gaze patterns, namely the estimated word's regression and skip probability, were significant predictors of revisions. This was a consistent result across all eye-tracking corpora, for the BiLSTM and the Transformer, both for dependency parsing and POS-tagging. On the one hand, human regressions were often positively related to revisions, so that words with a higher regression probability were more likely to be revised by models (MECO-L2-Transformer-pos was the only exception where regression probability negatively affected revisions). Conversely, a word's skip probability decreased the probability of it triggering a revision in most cases (with the exceptions of Potec and Provo-Transformer-pos and Nicenboim-BiLSTM-pos). These relationships are illustrated in Figure 4. The magnitude of the regression coefficient did not follow a general pattern for the tasks, but the skip coefficient was more often larger for the task of predicting the head than for the dependency relation, which was usually larger than for POS-tagging (exceptions to this is RastrOS-Transformer and MECO-L1-BiLSTM).

In a further analysis, we repeated the same procedure to predict only the effective revisions and observed the same trend in regression and skip coefficients when predicting effective revisions, in terms of direction and significance, in all experiments. However, the magnitude of the coefficients differed, sometimes being larger in one or the other, which does not allow us to draw general conclusions at this point. The coefficient of token position

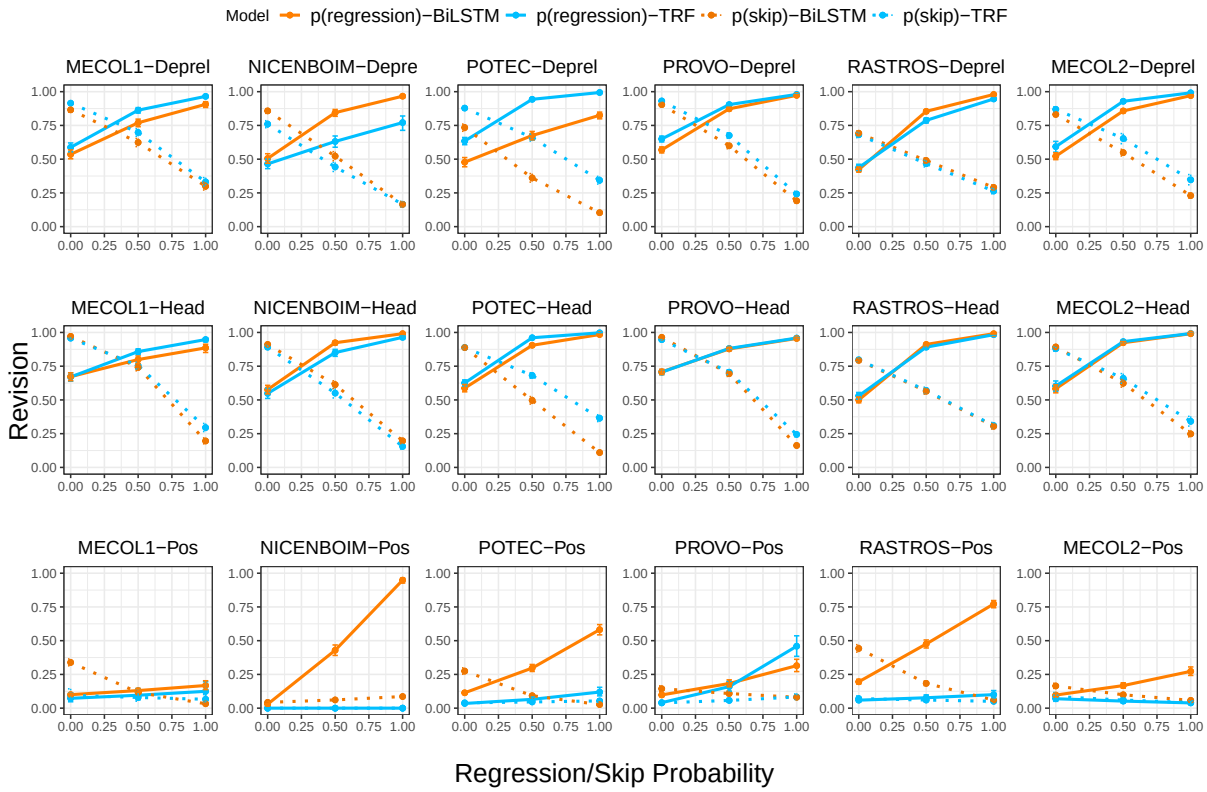

Figure 4: The full GLMM predictions of the revision probability are shown. Each plot presents the predictions for BiLSTM and Transformer models given regression and skip probability in the corresponding dataset. Error bars represent 95% confidence interval.

was, in most cases, smaller in the model that predicts effective revisions. Similarly, in many models the magnitude of the coefficient of skips was larger for models predicting effective revisions.

To assess the fit of the model to the data in more detail, we evaluated its predictions by running permutation tests with the null hypothesis that the probabilities assigned to (effective) revisions and to not-revisions are randomly sampled from the same distribution. Besides, we computed the area under the ROC curve in each model. As we can see in Table 5, most of the differences were significant (except for many cases in POS-tagging), but their magnitude was relatively small. The AUC was around 0.7 for all datasets, and in some experiments the models of effective revisions had higher AUC. Examples with considerable improvements are RastrOS-head and Nicenboim-head.

7 Do models revise when humans regress?

We have gathered evidence that there is a relationship between NLP restart-incremental models' revisions and human gaze behaviour in reading, which manifests as the probability of revision at a given token being partially predictable from it being often skipped or triggering regressions, when token position and text are accounted for. Interestingly, the overall findings hold for BiLSTM and Transformers, even though their encoding mechanisms are different, and also for all five languages, despite the eye-tracking data having been collected from different text genres and the readers having performed different tasks (or no additional task beyond reading for comprehension, as in Provo).

For this conclusion, we did not rely on any assumptions for the connection between human regressions and incremental models' revisions beyond the analogy of what we factually know: When seeing text areas for the first time, humans made decisions to skip or fixate, and possibly to revisit past text, and at some words, models "decided" to revisit past decisions.

Some exceptions to the general trend in predicting model revisions occurred in POS-tagging, for which relatively fewer revisions occur (see Table 2). The sparsity of revisions may cause the signal to be harder to model well without more data. For dependency parsing, more revisions are expected, especially because in the beginning of the sentence

			abs. mean diff		AUC	
			all-r	eff-r	all-r	eff-r
MECO-L1	deprel	BiLSTM	0.13*	0.16*	0.71	0.74
		Trfmer	0.15*	0.14*	0.73	0.72
	head	BiLSTM	0.22*	0.26*	0.78	0.80
		Trfmer	0.18*	0.21*	0.76	0.77
	pos	BiLSTM	0.05*	0.05*	0.69	0.71
		Trfmer	0.03*	0.02	0.68	0.66
MECO-L2	deprel	BiLSTM	0.12*	0.12*	0.70	0.69
		Trfmer	0.14*	0.10*	0.72	0.68
	head	BiLSTM	0.15*	0.20*	0.73	0.76
		Trfmer	0.12*	0.22*	0.70	0.77
	pos	BiLSTM	0.02*	0.02*	0.63	0.62
		Trfmer	0.03*	0.01*	0.67	0.64
Nicenboim	deprel	BiLSTM	0.27*	0.28*	0.79	0.80
		Trfmer	0.19*	0.19*	0.74	0.74
	head	BiLSTM	0.31*	0.45*	0.81	0.88
		Trfmer	0.31*	0.41*	0.81	0.87
	pos	BiLSTM	0.03*	0.04*	0.69	0.73
		Trfmer	0.06	0.06	0.89	0.89
PoTeC	deprel	BiLSTM	0.14*	0.12*	0.71	0.70
		Trfmer	0.14*	0.11*	0.74	0.69
	head	BiLSTM	0.23*	0.28*	0.79	0.81
		Trfmer	0.15*	0.22*	0.75	0.77
	pos	BiLSTM	0.08*	0.08*	0.70	0.71
		Trfmer	0.01	0.00	0.62	0.61
Provo	deprel	BiLSTM	0.20*	0.19*	0.76	0.75
		Trfmer	0.20*	0.17*	0.76	0.74
	head	BiLSTM	0.25*	0.21*	0.79	0.77
		Trfmer	0.20*	0.22*	0.76	0.77
	pos	BiLSTM	0.02	0.01	0.64	0.64
		Trfmer	0.04	0.02	0.72	0.70
RastrOS	deprel	BiLSTM	0.17*	0.18*	0.74	0.74
		Trfmer	0.16*	0.16*	0.73	0.74
	head	BiLSTM	0.22*	0.32*	0.77	0.83
		Trfmer	0.21*	0.31*	0.76	0.82
	pos	BiLSTM	0.16*	0.17*	0.76	0.76
		Trfmer	0.05*	0.02	0.71	0.68

Table 5: Left block: Absolute difference of sample means in the predictions of the models between time steps with and without revisions. * means p-value < 0.001. Right block: Area Under the ROC Curve when the fitted models' predictions are used for binary classification of revision time steps in the data.

the model has to wait until the root is processed to make good predictions. There may also be a difference in processing, since the humans could regress to previous sentences in the text, whereas the NLP models depend on their internal tokenisation and sentence boundary detection.

This suggests that eye-tracking measures can be transformed into a useful signal to inform the decision of when to revise in mixed restart-incremental processors, especially when the model's task entails more syntactic tasks with frequent revisions to the input.

Still, preliminary investigation of the revision probabilities predicted by the model did not yield a straightforward threshold for binary classification, despite the difference in means being statistically significant. This invites a more detailed extrinsic evaluation, by incorporating the human predictors into a revision controller like TAPIR (Kahardipraja et al., 2023), and assessing the revisions with the evaluation methods discussed by Madureira et al. (2023). One approach is to train an incremental sequence labelling model whose revision policy relies on eye-tracking data as part of the input and comparing its performance against a model without it. Since skips had a negative effect, it may also be possible to use other variables that relate to the probability of a token being skipped, like POS-tags or word frequency and length, as additional input, which are cheaper to obtain. The analysis should also be done with larger datasets and other models and tasks.

The usefulness of our findings presupposes the availability of eye-tracking measures during inference on truly unseen data, which is an open problem because such signal is not always available in real time. One possibility is to use pretrained eye-tracking models to predict regressions and skips, as in approaches discussed in the literature (Engbert et al., 2005; Deng et al., 2023).

Down the road, a revision policy should not only detect times to revise, but times to revise *effectively*, since wrong revisions make the partial outputs less reliable for downstream processors. Our experiments showed that regressions and skips are also good predictors for effective revisions. Identifying ways to filter this more specific signal demands further investigation. An immediate next step is to evaluate the predictions of each model in unseen data for all revisions and for effective revisions.

8 Conclusion

Let us conclude with a *backward glance* to our contribution. We have addressed the open question of whether pre-trained sequence labelling models, when employed incrementally, perform revisions in a similar fashion as humans skip words or make regressive eye movements while reading. We have

found a significant effect in all the experiments, supporting the use of human reading data as a cognitive signal to inform revision policies. This is a valuable finding: BiLSTMs and Transformers are bidirectional, trained on full sequences, but if we make them process linguistic input incrementally, their revisions can be partially predicted by human reading behaviour. This is also a step forward towards understanding why these models change hypotheses at some tokens, when only partial prefixes are available.

Besides advancing the research on eye-tracking-augmented NLP, this study also opens the door to exploring other cognitive perspectives with restart-incremental NLP models. We see a potential to go the other direction and investigate to what extent a "mixed incrementality" model (architectures relying on an incremental processor with occasional restarts) would capture the patterns of human gaze in reading, and hence function as a model of that. In this case, revisions would serve as predictors of human regressions, with control variables like word frequency, surprisal and word length. Other possibility for future work is to investigate whether other measures, like number of fixations or regressions *to* a token, are related to the edits per label.

Limitations

Here we summarise a few known limitations that we have mentioned throughout the text. We have analysed various datasets which differ both in the ways they were collected (the task humans were performing, *e.g.* only reading or also answering to comprehension questions) as well as the length and genre of the texts. The size of the eye-tracking datasets is, in general, small. Ideally, larger amounts of data are necessary to train a revision policy than what we had available for the analysis. Some preprocessing steps had to be made; in particular, some decisions were necessary on had how to merge tokens and interpret documentation, so that a mapping could be created. This is documented in the Appendix, but alternative ways are also possible. We limited the study to families of pre-trained models and tasks for which all languages were available. There can be a mismatch between the humans having the full text available at any point and the models performing sentence segmentation internally in different ways. For models that are trained on sequence level, it may be better if the human reading is also performed the same way. Further research expanding these aspects is desired. Other models beyond GLLMs, *e.g.* with non-linearity, may be examined, because the probability of regression is within a narrow range in most of the cases. Using models' revisions to predict human behaviour is also a possible research question which was not addressed in this work.

Acknowledgements

We thank Nora Hollenstein for some helpful advice on using eye-tracking measures, as well as the authors of the eye-tracking datasets who replied to our clarification requests. Thanks to Patrick Kahardipraja for initial discussions on using surprisal as a signal for revisions policies. We are also thankful for the valuable feedback and suggestions provided by the anonymous reviewers.

References

Gerry TM Altmann, Alan Garnham, and Yvette Dennis. 1992. Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, 31(5):685–712.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584, Berlin, Germany. Association for Computational Linguistics.

Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14(11):1–16.

Maria Barrett and Anders Søgaard. 2015a. Reading behavior predicts syntactic categories. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 345–349, Beijing, China. Association for Computational Linguistics.

Maria Barrett and Anders Søgaard. 2015b. Using reading behavior to predict grammatical functions. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, pages 1–5, Lisbon, Portugal. Association for Computational Linguistics.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Timo Baumann, Okko Buß, and David Schlangen. 2011. Evaluation and Optimisation of Incremental Processors. *Dialogue and Discourse*, 2(1):113–141.

Robert W Booth and Ulrich W Weger. 2013. The function of regressions in reading: Backward eye movements allow rereading. *Memory & cognition*, 41:82–97.

Joe Cheri, Abhijit Mishra, and Pushpak Bhattacharyya. 2016. Leveraging annotators' gaze behaviour for coreference resolution. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 22–26, Berlin. Association for Computational Linguistics.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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

Shuwen Deng, David R. Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A. Jäger. 2023. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. *Proc. ACM Hum.-Comput. Interact.*, 7(ETRA).

Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. 2005. Swift: a dynamical model of saccade generation during reading. *Psychological review*, 112(4):777.

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Hough and David Schlangen. 2015. Recurrent Neural Networks for Incremental Disfluency Detection. In *Interspeech 2015*, pages 849–853.

Ryu Iida, Koh Mitsuda, and Takenobu Tokunaga. 2013. Investigation of annotator's behaviour using eye-tracking data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 214–222, Sofia, Bulgaria. Association for Computational Linguistics.

Fariz Ikhwantri, Jan Wira Gotama Putra, Hiroaki Yamada, and Takenobu Tokunaga. 2023. Looking deep in the eyes: Investigating interpretation methods for neural models on reading tasks using human eye-movement behaviour. *Information Processing and Management*, 60(2):103195.

Lena A Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler, and Tobias Scheffer. 2020. Deep eyedentification: Biometric identification using micro-movements of the eye. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 299–314. Springer.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2021. Towards incremental transformers: An empirical analysis of transformer models for incremental NLU. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1178–1189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2023. Tapir: Learning adaptive revision for incremental natural language understanding with a two-pass model. In *Findings of the Association for Computational Linguistics: ACL 2023 (forthcoming)*, Toronto, Canada. Association for Computational Linguistics.

Ayush Kaushal, Aditya Gupta, Shyam Upadhyay, and Manaal Faruqui. 2023. Efficient encoders for streaming sequence tagging. *arXiv preprint arXiv:2301.09244*.

Casey Kennington and David Schlangen. 2017. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language*, 41:43–67.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Arne Köhn. 2018. Incremental natural language processing: Challenges, strategies, and evaluation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2990–3003, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2023. Text reading in english as a second language: Evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*, 45(1):3–37.

Sidney Evaldo Leal, Katerina Lukasova, Maria Teresa Carthery-Goulart, and Sandra Maria Aluísio. 2022. Rastros project: Natural language processing contributions to the development of an eye-tracking corpus with predictability norms for brazilian portuguese. *Language Resources and Evaluation*, pages 1–40.

Alessandro Lopopolo, Stefan L. Frank, Antal van den Bosch, and Roel Willems. 2019. Dependency parsing with your eyes: Dependency structure predicts eye regressions during reading. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 77–85, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Brielen Madureira, Patrick Kahardipraja, and David Schlangen. 2023. The road to quality is paved with good revisions: A detailed evaluation methodology for revision policies in incremental sequence labelling. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 156–167, Prague, Czechia. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2020. Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 357–374, Online. Association for Computational Linguistics.

Silvia Makowski, Lena A Jäger, Ahmed Abdelwahab, Niels Landwehr, and Tobias Scheffer. 2019. A discriminative model for identifying readers and assessing text comprehension from eye movements. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 209–225. Springer.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. 2021. A survey on using gaze behaviour for natural language processing. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4907–4913.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Bruno Nicenboim, Shravan Vasishth, Carolina Gattei, Mariano Sigman, and Reinhold Kliegl. 2015. Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, 6:312.

Dario Paape, Shravan Vasishth, and Ralf Engbert. 2021. Does local coherence lead to targeted regressions and illusions of grammaticality? *Open Mind*, 5:42–58.

Dario Paape, Shravan Vasishth, Dario Paape, and Shravan Vasishth. 2022. Is reanalysis selective when regressions are consciously controlled? *Glossa Psycholinguistics*, 1(1).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading*. Psychology Press.

Tiago Ribeiro, Stephanie Brandl, Anders Søgaard, and Nora Hollenstein. 2023. Webqamgaze: A multilingual webcam eye-tracking-while-reading dataset. *arXiv preprint arXiv:2303.17876*.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue and Discourse*, 2(1):83–111.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Sukanta Sen, Rico Sennrich, Biao Zhang, and Barry Haddow. 2023. Self-training reduces flicker in retranslation-based simultaneous translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3734–3744, Dubrovnik, Croatia. Association for Computational Linguistics.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka, Japan. The COLING 2016 Organizing Committee.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, pages 1–21.

Takenobu Tokunaga, Hitoshi Nishikawa, and Tomoya Iwakura. 2017. An eye-tracking study of named entity annotation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 758–764, Varna, Bulgaria. INCOMA Ltd.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

João Marcos Munguba Vieira. 2020. *The Brazilian Portuguese eye tracking corpus with a predictability study focusing on lexical and partial prediction.* Master thesis, Linguistics Department, ederal University of Ceará, Fortaleza.

Françoise Vitu. 2005. Visual extraction processes and regressive saccades in reading. *Cognitive processes in eye guidance*, pages 1–32.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

A Appendix

B Pre-processing Human Data

We pre-process all datasets to combine the measures into a common format, with one token per row and one column for each subject. If no data was available for a subject, the cell is filled with a NaN value, so that it is later ignored. We partition the measure into three groups: interest areas that were skipped in the first-pass reading (and, consequently, also interest areas that were skipped altogether) are assigned a `skipped` category (label -1). For the remaining interest areas, *i.e.* those that had a first-pass fixation, we extract either a `regressed` (label 1) or `not regressed` (label 0) category. Here we document some necessary decisions. The measures we rely on are documented in Table 6 and the pre-processing scripts are available at https://github.com/briemadu/revreg. For further details about the data collections, please refer to the original publications.

▷ **RastrOS**: Participants read paragraphs, one by one in a random order, from journalistic, literary and popular science sources. There was a yes/no comprehension question after 20 of the paragraphs. We get the tokens from the columns `Word` and `IA_LABEL`. We solve inconsistencies as follows: if `Word` contains a comma and `IA_LABEL` contains a full stop, we use the former (in accordance to personal communication with the author). If there are mismatches in quotation marks, we also use the former. For other inconsistencies (33 tokens), we use the latter.

▷ **PoTeC**: Participants read scientific texts on biology and physics from textbooks. Three multiple-choice comprehension questions were presented after each text in a separate screen. We use the negation of `FPF` as an auxiliary to detect tokens that were skipped in the first pass. The raw text files do not contain punctuation in a straightforward format. We thus only extract commas, and final sentence punctuation is considered to be always a full stop, except for two cases that we noticed were not end of sentences, so a ; was used. We follow the list of 13 subjects ids (in the original script `mergeFixationsWordFeatures.py`) that were removed due to poor calibration (according to Jäger et al. (2020)) and exclude them from our sample.

▷ **Provo**: Participants read the texts from various sources in a random order, without any additional task. For the tokens, we rely on `IA_LABEL`, due to inconsistencies in the `Word` column. Four tokens do not match the raw texts (apparently due to encoding), so we use the text instead of the `IA_LABEL`.

▷ **MECO-L1**: Wikipedia style texts, each on a separate screen. After each text, there were four yes/no comprehension questions. We could only use the Dutch version, as the other languages had mismatches between the source texts and the interest area column.

▷ **MECO-L2**: Texts are from training materials for English tests. Participants answered four yes/no questions after each text. 5 subjects were excluded due to unexplained repetitions.

▷ **Nicenboim**: Participants read stimuli (sentences). True/false statements appeared randomly after half of them. We use the filler sentences (as the others had varying conditions across participants). We use `FPRT`, assuming it is first-pass reading time, to infer first-pass fixations: if it is NaN, we consider it to be a skip (because otherwise it is always a number higher than 0).

C Pre-processing Models' Data

We use off-the-shelf implementations of sequence labelling models. To extract the outputs, we loop over the interest areas for each text in the eye-tracking corpus for the corresponding language. At each time step t, a string is created with the interest areas up to position t, joined with a blank space. The models output a list of labels, which we take to be the output prefix for that time step. Due to the internal tokenization, it can happen in a few cases that tokenization changes slightly or that more than one new label is added. We use the number of labels in the previous time step as a reference, all new labels beyond that length are considered an addition and do not affect revisions. A revision happens if the output prefix at time t differs from the output at time $t-1$; and it is effective if the number of labels that match the final output labels up to that time step increased. For Stanza BiLSTM, we extract the labels from the attributes `upos_`, `deprel`, `head`. For Explosion's Transformers, we extract the labels from the attributes `pos_`, `dep_`, `head_i`.

D Modelling Details

We fit generalized linear mixed models using the lme4 (Bates et al., 2015) package in the R statistical

	regression	description	skip	description
MECO-L1 and **MECO-L2**	`firstrun.reg.out`	*Variable indicating whether there was a regression from the IA during first-pass reading*	`firstrun.skip`	*Variable indicating whether the IA was skipped during first-pass reading*
Nicenboim	`fp_reg`	no description	`FPRT`	no description
PoTeC	`FPReg`	*1 if a regression was initiated in the first-pass reading of the word, otherwise 0 (sign(RPD exc))*	negation of `FPF`	*1 if the word was fixated in the first-pass, otherwise 0*
Provo	`IA_REGRESSION_OUT`	*Whether the current interest area received at least one regression from later interest areas (e.g., later parts of the sentence). 1 if interest area was entered from a higher IA_ID (from the right in English); 0 if not. (...) Note that IA_REGRESSION_OUT only considers first-pass regressions.*	`IA_SKIP`	*An interest area is considered skipped (i.e., IA_SKIP = 1) if no fixation occurred in first-pass reading.*
RastrOS	`IA_REGRESSION_OUT`	*Whether regression(s) was made from the current interest area to earlier interest areas (e.g., previous parts of the sentence) prior to leaving that interest area in a forward direction. 1 if a saccade exits the current interest area to a lower IA_ID (to the left in English) before a later interest area was fixated; 0 if not. (...) Note that IA_REGRESSION_OUT only considers first-pass regressions.*	`IA_SKIP`	*An interest area is considered skipped (i.e.,IA_SKIP = 1) if no fixation occurred in first-pass reading.*

Table 6: Measures used for each eye-tracking corpus and their definition according to the available documentation.

computing environment (R Core Team, 2022). All baseline and full models were initially fit with the same structure described in the Methods section. We made changes to the model structure in 6 cases to tackle with convergence issues: Model fits to the Nicenboim-TRF-Pos and Nicenboim-BiLSTM-Pos datasets revealed low text-level variance and random effects were excluded in these datasets in further analyses. Token position was not a significant predictor of model revision in MECOL1-BiLSTM-Head, MECOL2-TRF-Head, MECOL2-BiLSTM-Head, and Provo-BiLSTM-Pos models, thus, we refitted these models without the token positions variable.

E Detailed Results

Tables 7 and 8 show all the estimated coefficients, standard errors, z and p-values for all models. Table 9 presents the results of the likelihood ratio tests for the full models in relation to their corresponding null model. All results in the paper have been rounded to to decimal places programatically.

				estimate		SE		z		p	
				all-r	eff-r	all-r	eff-r	all-r	eff-r	all-r	eff-r
MECO-L1 (du)	BiLSTM	deprel	intercept	1.47***	1.52***	0.08	0.07	17.33	23.32	<0.001	<0.001
			p(reg)	2.13***	1.60***	0.12	0.11	17.04	13.97	<0.001	<0.001
			p(skip)	-2.71***	-3.48***	0.05	0.05	-52.0	-67.81	<0.001	<0.001
			position	0.03**	0.0	0.01	0.01	3.10	0.49	0.002	0.622
		head	intercept	3.34***	1.98***	0.08	0.08	40.29	25.40	<0.001	<0.001
			p(reg)	1.34***	1.31***	0.16	0.11	8.51	11.47	<0.001	<0.001
			p(skip)	-4.93***	-5.14***	0.07	0.06	-75.22	-91.52	<0.001	<0.001
			position	-	-	-	-	-	-	-	-
		pos	intercept	0.07	0.02	0.08	0.08	0.86	0.25	0.388	0.805
			p(reg)	0.59***	0.63***	0.12	0.13	4.77	4.89	<0.001	<0.001
			p(skip)	-2.72***	-2.96***	0.07	0.07	-41.33	-42.59	<0.001	<0.001
			position	-0.18***	-0.18***	0.01	0.01	-15.50	-14.60	<0.001	<0.001
	Transformer	deprel	intercept	1.70***	1.13***	0.09	0.07	19.25	16.35	<0.001	<0.001
			p(reg)	2.97***	2.38***	0.15	0.12	20.42	19.60	<0.001	<0.001
			p(skip)	-3.08***	-2.87***	0.06	0.05	-54.30	-56.21	<0.001	<0.001
			position	0.07***	0.05***	0.01	0.01	7.31	6.11	<0.001	<0.001
		head	intercept	2.17***	1.67***	0.09	0.08	23.66	21.33	<0.001	<0.001
			p(reg)	2.18***	1.16***	0.16	0.11	13.87	10.48	<0.001	<0.001
			p(skip)	-3.99***	-4.49***	0.06	0.05	-63.76	-83.52	<0.001	<0.001
			position	0.15***	-0.0	0.01	0.01	15.93	-0.27	<0.001	0.789
		pos	intercept	-2.11***	-2.22***	0.25	0.20	-8.55	-11.08	<0.001	<0.001
			p(reg)	0.59***	0.91***	0.16	0.19	3.65	4.86	<0.001	<0.001
			p(skip)	-0.40***	-0.66***	0.08	0.09	-5.08	-7.05	<0.001	<0.001
			position	-0.05***	-0.10***	0.01	0.02	-3.45	-6.13	<0.001	<0.001
MECO-L2 (en-l2)	BiLSTM	deprel	intercept	1.29***	1.04***	0.05	0.04	24.18	26.86	<0.001	<0.001
			p(reg)	3.41***	3.10***	0.05	0.04	73.39	72.32	<0.001	<0.001
			p(skip)	-2.80***	-2.80***	0.02	0.02	-178.47	-182.23	<0.001	<0.001
			position	-0.03***	-0.03***	0.0	0.0	-8.94	-10.02	<0.001	<0.001
		head	intercept	1.59***	0.72***	0.06	0.06	27.44	12.28	<0.001	<0.001
			p(reg)	4.32***	3.27***	0.05	0.04	81.05	85.21	<0.001	<0.001
			p(skip)	-3.23***	-4.49***	0.02	0.02	-193.35	-257.08	<0.001	<0.001
			position	-	-	-	-	-	-	-	-
		pos	intercept	-2.62***	-2.72***	0.07	0.06	-36.21	-44.03	<0.001	<0.001
			p(reg)	1.25***	1.18***	0.05	0.05	27.53	25.28	<0.001	<0.001
			p(skip)	-1.16***	-1.23***	0.02	0.02	-52.26	-53.20	<0.001	<0.001
			position	0.20***	0.21***	0.0	0.0	42.18	42.45	<0.001	<0.001
	Transformer	deprel	intercept	1.22***	1.17***	0.09	0.07	14.28	16.69	<0.001	<0.001
			p(reg)	4.39***	2.56***	0.05	0.04	82.91	60.24	<0.001	<0.001
			p(skip)	-2.53***	-2.70***	0.02	0.02	-154.71	-176.92	<0.001	<0.001
			position	0.03***	-0.01***	0.0	0.0	11.37	-5.27	<0.001	<0.001
		head	intercept	1.45***	0.81***	0.08	0.05	18.13	17.05	<0.001	<0.001
			p(reg)	4.40***	3.11***	0.05	0.04	82.24	81.62	<0.001	<0.001
			p(skip)	-2.64***	-4.93***	0.02	0.02	-160.14	-270.17	<0.001	<0.001
			position	-	-	-	-	-	-	-	-
		pos	intercept	-2.64***	-2.62***	0.17	0.14	-15.28	-18.69	<0.001	<0.001
			p(reg)	-0.62***	-1.35***	0.06	0.08	-9.49	-17.23	<0.001	<0.001
			p(skip)	-0.77***	-0.40***	0.03	0.03	-29.33	-13.58	<0.001	<0.001
			position	0.08***	0.01*	0.01	0.01	15.56	2.10	<0.001	0.035
Nicenboim (es)	BiLSTM	deprel	intercept	0.42***	0.22**	0.07	0.07	5.61	3.04	<0.001	0.002
			p(reg)	3.35***	4.83***	0.18	0.18	18.17	26.59	<0.001	<0.001
			p(skip)	-3.42***	-3.32***	0.05	0.05	-70.86	-69.87	<0.001	<0.001
			position	0.46***	0.32***	0.02	0.02	28.17	19.62	<0.001	<0.001
		head	intercept	0.55***	0.90***	0.07	0.08	7.97	11.03	<0.001	<0.001
			p(reg)	4.37***	3.22***	0.21	0.19	20.91	16.87	<0.001	<0.001
			p(skip)	-3.74***	-6.59***	0.05	0.06	-73.13	-102.23	<0.001	<0.001
			position	0.59***	0.47***	0.02	0.02	34.30	23.23	<0.001	<0.001
		pos	intercept	-4.49***	-4.82***	0.07	0.08	-60.34	-58.02	<0.001	<0.001
			p(reg)	6.40***	6.58***	0.22	0.23	28.83	28.66	<0.001	<0.001
			p(skip)	0.74***	0.39***	0.09	0.10	8.43	4.10	<0.001	<0.001
			position	0.31***	0.43***	0.03	0.03	10.26	12.47	<0.001	<0.001
	Transformer	deprel	intercept	0.18*	0.07	0.08	0.07	2.34	0.99	0.019	0.32
			p(reg)	1.36***	1.89***	0.16	0.15	8.68	12.29	<0.001	<0.001
			p(skip)	-2.77***	-2.80***	0.04	0.04	-62.04	-62.64	<0.001	<0.001
			position	0.37***	0.30***	0.02	0.02	24.60	19.58	<0.001	<0.001
		head	intercept	0.55***	0.82***	0.08	0.08	6.96	10.46	<0.001	<0.001
			p(reg)	3.09***	3.89***	0.19	0.18	16.32	21.24	<0.001	<0.001
			p(skip)	-3.79***	-5.90***	0.05	0.06	-76.03	-97.66	<0.001	<0.001
			position	0.55***	0.30***	0.02	0.02	32.66	15.40	<0.001	<0.001
		pos	intercept	-2.87***	-2.72***	0.08	0.08	-34.99	-32.63	<0.001	<0.001
			p(reg)	2.92***	2.52***	0.41	0.45	7.07	5.54	<0.001	<0.001
			p(skip)	-0.54***	-0.76***	0.13	0.14	-4.06	-5.41	<0.001	<0.001
			position	-0.68***	-0.79***	0.04	0.05	-16.06	-17.45	<0.001	<0.001

Table 7: Overview of all results (part I).

				estimate		SE		z		p	
				all-r	eff-r	all-r	eff-r	all-r	eff-r	all-r	eff-r
PoTeC (de)	BiLSTM	deprel	intercept	0.45***	0.14*	0.08	0.06	5.98	2.17	<0.001	0.03
			p(reg)	1.64***	1.77***	0.06	0.06	25.48	29.24	<0.001	<0.001
			p(skip)	-3.18***	-2.98***	0.04	0.04	-86.61	-81.10	<0.001	<0.001
			position	0.07***	0.02***	0.01	0.01	10.20	3.51	<0.001	<0.001
		head	intercept	0.95***	-0.17***	0.07	0.05	14.36	-3.59	<0.001	<0.001
			p(reg)	3.80***	3.99***	0.09	0.07	41.28	57.10	<0.001	<0.001
			p(skip)	-4.16***	-4.99***	0.04	0.04	-100.28	-113.17	<0.001	<0.001
			position	0.12***	0.07***	0.01	0.01	15.72	9.71	<0.001	<0.001
		pos	intercept	-1.88***	-1.89***	0.07	0.07	-26.25	-26.58	<0.001	<0.001
			p(reg)	2.38***	2.66***	0.06	0.07	37.30	40.46	<0.001	<0.001
			p(skip)	-2.63***	-2.78***	0.05	0.05	-52.65	-51.57	<0.001	<0.001
			position	0.12***	0.07***	0.01	0.01	13.44	7.61	<0.001	<0.001
	Transformer	deprel	intercept	0.62***	0.28***	0.07	0.05	9.23	5.88	<0.001	<0.001
			p(reg)	4.53***	2.86***	0.10	0.07	45.75	41.76	<0.001	<0.001
			p(skip)	-2.62***	-2.29***	0.04	0.04	-64.82	-64.33	<0.001	<0.001
			position	0.14***	0.04***	0.01	0.01	19.63	5.69	<0.001	<0.001
		head	intercept	0.46***	-0.20***	0.06	0.05	7.90	-4.53	<0.001	<0.001
			p(reg)	5.36***	3.31***	0.11	0.07	50.06	49.28	<0.001	<0.001
			p(skip)	-2.63***	-4.08***	0.04	0.04	-63.93	-101.16	<0.001	<0.001
			position	0.17***	0.10***	0.01	0.01	23.09	13.98	<0.001	<0.001
		pos	intercept	-2.40***	-2.47***	0.13	0.11	-18.18	-21.86	<0.001	<0.001
			p(reg)	1.32***	1.12***	0.12	0.13	11.36	8.57	<0.001	<0.001
			p(skip)	0.32***	0.36***	0.08	0.08	4.29	4.37	<0.001	<0.001
			position	-0.23***	-0.25***	0.01	0.01	-18.32	-18.61	<0.001	<0.001
Provo (en)	BiLSTM	deprel	intercept	1.22***	0.80***	0.05	0.04	24.29	18.20	<0.001	<0.001
			p(reg)	3.30***	2.95***	0.09	0.08	38.56	38.54	<0.001	<0.001
			p(skip)	-3.68***	-3.66***	0.03	0.03	-133.52	-137.30	<0.001	<0.001
			position	0.21***	0.24***	0.01	0.01	38.87	43.77	<0.001	<0.001
		head	intercept	1.76***	0.41***	0.05	0.04	33.12	10.44	<0.001	<0.001
			p(reg)	2.18***	1.36***	0.10	0.07	21.84	20.64	<0.001	<0.001
			p(skip)	-4.92***	-4.57***	0.03	0.03	-155.18	-161.06	<0.001	<0.001
			position	0.40***	0.31***	0.01	0.01	68.85	54.05	<0.001	<0.001
		pos	intercept	-1.92***	-2.02***	0.08	0.08	-22.77	-25.78	<0.001	<0.001
			p(reg)	1.42***	1.58***	0.08	0.08	18.61	20.38	<0.001	<0.001
			p(skip)	-0.66***	-0.77***	0.04	0.04	-18.63	-20.72	<0.001	<0.001
			position	-	-	-	-	-	-	-	-
	Transformer	deprel	intercept	1.28***	0.93***	0.05	0.04	24.39	23.44	<0.001	<0.001
			p(reg)	3.26***	2.69***	0.09	0.08	34.39	33.70	<0.001	<0.001
			p(skip)	-3.75***	-3.41***	0.03	0.03	-129.34	-127.32	<0.001	<0.001
			position	0.30***	0.24***	0.01	0.01	54.95	45.93	<0.001	<0.001
		head	intercept	1.45***	0.46***	0.05	0.04	29.17	11.59	<0.001	<0.001
			p(reg)	2.27***	1.69***	0.10	0.07	23.76	25.60	<0.001	<0.001
			p(skip)	-4.01***	-4.66***	0.03	0.03	-133.24	-163.42	<0.001	<0.001
			position	0.37***	0.28***	0.01	0.01	64.92	48.09	<0.001	<0.001
		pos	intercept	-2.69***	-2.89***	0.14	0.13	-19.71	-23.06	<0.001	<0.001
			p(reg)	3.00***	3.15***	0.10	0.10	31.11	30.24	<0.001	<0.001
			p(skip)	0.80***	0.93***	0.04	0.05	18.07	19.09	<0.001	<0.001
			position	-0.25***	-0.27***	0.01	0.01	-30.18	-29.77	<0.001	<0.001
RastrOS (pt-br)	BiLSTM	deprel	intercept	-0.22***	-0.32***	0.05	0.05	-4.68	-7.01	<0.001	<0.001
			p(reg)	4.16***	3.62***	0.08	0.07	51.32	49.69	<0.001	<0.001
			p(skip)	-1.70***	-2.05***	0.03	0.03	-56.48	-65.25	<0.001	<0.001
			position	0.14***	0.12***	0.01	0.01	17.09	14.20	<0.001	<0.001
		head	intercept	-0.15**	-0.19***	0.05	0.05	-2.90	-3.69	0.004	<0.001
			p(reg)	4.66***	4.03***	0.10	0.08	48.70	50.46	<0.001	<0.001
			p(skip)	-2.17***	-4.13***	0.03	0.04	-69.37	-102.76	<0.001	<0.001
			position	0.26***	0.19***	0.01	0.01	30.43	20.73	<0.001	<0.001
		pos	intercept	-0.99***	-0.98***	0.06	0.06	-15.45	-16.45	<0.001	<0.001
			p(reg)	2.63***	2.61***	0.06	0.06	43.02	42.69	<0.001	<0.001
			p(skip)	-2.52***	-2.91***	0.04	0.04	-65.44	-69.72	<0.001	<0.001
			position	0.12***	0.11***	0.01	0.01	13.14	11.68	<0.001	<0.001
	Transformer	deprel	intercept	-0.10	-0.26***	0.06	0.05	-1.79	-5.77	0.073	<0.001
			p(reg)	3.12***	3.07***	0.07	0.07	42.88	45.31	<0.001	<0.001
			p(skip)	-1.78***	-2.06***	0.03	0.03	-59.32	-65.27	<0.001	<0.001
			position	0.13***	0.08***	0.01	0.01	16.10	10.07	<0.001	<0.001
		head	intercept	-0.07	-0.26***	0.06	0.05	-1.14	-4.81	0.255	<0.001
			p(reg)	3.95***	3.74***	0.09	0.08	43.65	48.84	<0.001	<0.001
			p(skip)	-2.17***	-3.93***	0.03	0.04	-69.58	-100.10	<0.001	<0.001
			position	0.28***	0.20***	0.01	0.01	31.87	21.72	<0.001	<0.001
		pos	intercept	-1.97***	-2.12***	0.14	0.12	-14.37	-17.71	<0.001	<0.001
			p(reg)	0.57***	0.78***	0.09	0.09	6.32	8.43	<0.001	<0.001
			p(skip)	-0.36***	-0.65***	0.05	0.06	-7.14	-11.52	<0.001	<0.001
			position	-0.21***	-0.18***	0.01	0.01	-16.20	-13.25	<0.001	<0.001

Table 8: Overview of all results (part 2).

			BIC		χ^2		Df		p	
			all-r	eff-r	all-r	eff-r	all-r	eff-r	all-r	eff-r
MECO-L1 (du)	BiLSTM	deprel	83546.50	82400.75	7670.49	11010.84	2	2	<0.001	<0.001
		head	72210.57	71300.38	14407.59	18647.46	2	2	<0.001	<0.001
		pos	48702.71	44738.04	3086.69	3257.63	2	2	<0.001	<0.001
	Transformer	deprel	78422.59	84143.92	9326.01	9114.93	2	2	<0.001	<0.001
		head	73396.78	74729.62	11051.14	15030.07	2	2	<0.001	<0.001
		pos	40292.37	30157.89	104.20	188.25	2	2	<0.001	<0.001
MECO-L2 (en-l2)	BiLSTM	deprel	786910.36	813266.48	80770.80	80710.35	2	2	<0.001	<0.001
		head	721580.34	719579.16	99023.63	143061.55	2	2	<0.001	<0.001
		pos	453665.09	428746.07	6319.52	6111.91	2	2	<0.001	<0.001
	Transformer	deprel	733070.61	810697.01	71546.51	69594.78	2	2	<0.001	<0.001
		head	721451.14	700433.46	74786.38	156206.75	2	2	<0.001	<0.001
		pos	339878.15	289129.13	891.56	320.93	2	2	<0.001	<0.001
Nicenboim (es)	BiLSTM	deprel	62038.56	61705.57	12760.80	14065.35	2	2	<0.001	<0.001
		head	57808.46	47871.65	14703.78	28111.09	2	2	<0.001	<0.001
		pos	-	-	-	-	2	2	<0.001	<0.001
	Transformer	deprel	67874.18	67156.03	7930.09	8486.24	2	2	<0.001	<0.001
		head	59810.38	50487.63	14257.23	25095.62	2	2	<0.001	<0.001
		pos	-	-	-	-	2	2	<0.001	<0.001
Potec (de)	BiLSTM	deprel	145892.59	146622.24	15375.96	14146.19	2	2	<0.001	<0.001
		head	121763.91	123406.25	26247.07	35086.70	2	2	<0.001	<0.001
		pos	101606.89	92595.02	8237.35	8481.05	2	2	<0.001	<0.001
	Transformer	deprel	119666.13	148027.04	15136.50	12789.51	2	2	<0.001	<0.001
		head	116103.04	133947.43	16464.67	26794.85	2	2	<0.001	<0.001
		pos	45668.14	39611.90	122.34	69.33	2	2	<0.001	<0.001
Provo (en)	BiLSTM	deprel	265555.00	274893.44	38215.03	39111.65	2	2	<0.001	<0.001
		head	235978.14	258861.40	47478.23	46669.46	2	2	<0.001	<0.001
		pos	166971.65	155560.74	1373.30	1651.36	2	2	<0.001	<0.001
	Transformer	deprel	252130.08	275598.83	35534.99	33376.38	2	2	<0.001	<0.001
		head	243848.75	255250.75	34344.11	49079.47	2	2	<0.001	<0.001
		pos	118204.33	103652.19	886.80	843.64	2	2	<0.001	<0.001
RastrOS (pt-br)	BiLSTM	deprel	106976.51	106122.58	13333.03	14659.62	2	2	<0.001	<0.001
		head	99489.29	89499.43	16638.70	30213.54	2	2	<0.001	<0.001
		pos	92050.16	87854.00	12436.14	13744.63	2	2	<0.001	<0.001
	Transformer	deprel	108432.79	106773.67	11382.04	13219.43	2	2	<0.001	<0.001
		head	99987.62	91261.27	15027.19	27893.16	2	2	<0.001	<0.001
		pos	49397.83	44548.61	169.15	371.72	2	2	<0.001	<0.001

Table 9: Overview of likelihood ratio tests, showing how each full model compares to the null model.