At a Glance: The Impact of Gaze Aggregation Views on Syntactic Tagging

Sigrid Klerke
Department of Computer Science
ITU Copenhagen, Denmark
sigridklerke@gmail.com

Barbara Plank
Department of Computer Science
ITU Copenhagen, Denmark
bplank@itu.dk

Abstract

Readers' eye movements used as part of the training signal have been shown to improve performance in a wide range of Natural Language Processing (NLP) tasks. Previous work uses gaze data either at the type level or at the token level and mostly from a single eye-tracking corpus. In this paper, we analyze type vs token-level integration options with eye tracking data from two corpora to inform two syntactic sequence labeling problems: binary phrase chunking and part-of-speech tagging. We show that using globally-aggregated measures that capture the central tendency or variability of gaze data is more beneficial than proposed local views which retain individual participant information. While gaze data is informative for supervised POS tagging, which complements previous findings on unsupervised POS induction, almost no improvement is obtained for binary phrase chunking, except for a single specific setup. Hence, caution is warranted when using gaze data as signal for NLP, as no single view is robust over tasks, modeling choice and gaze corpus.

1 Introduction

Digital traces of human cognitive processing can provide valuable signal for Natural Language Processing (Klerke et al., 2016; Plank, 2016a,b). One emerging source of information studied within NLP is eye-tracking data (Barrett and Søgaard, 2015a; Klerke et al., 2016; Mishra et al., 2017a; Jaffe et al., 2018; Barrett et al., 2018b; Hollenstein et al., 2019). While ubiquitous gaze recording remains unavailable, NLP research has focused on exploring the value of including gaze information from large, mostly disjointly labeled gaze datasets in recurrent neural network models. This models the assumption that no new gaze data will be available at test time. The proposed approaches under this paradigm include gaze as auxiliary task

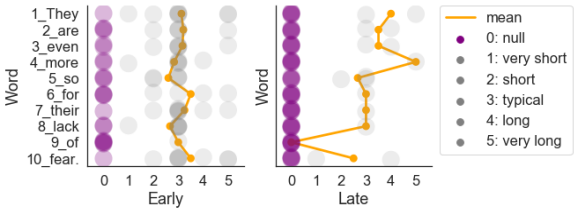


Figure 1: Gaze (binned) captured during reading.

in multi-task learning (Klerke et al., 2016; Hollenstein et al., 2019), gaze as word embeddings (Barrett et al., 2018b), gaze as type dictionaries (Barrett et al., 2016; Hollenstein and Zhang, 2019) and as attention (Barrett et al., 2018a). We follow this line of work and require no gaze data at test time.

Choosing a gaze representation means choosing what to consider as signal and what to consider as noise. Aggregation is a way to implement this choice; where the kind of aggregation typically depends on the modeling framework. In this work we investigate how different levels of aggregation and the kind of variability preserved in representations of gaze duration from early and late processing states interact with two low-level syntactic sequence labeling tasks. Specifically, we address the following questions:

RQ1 Is a *local* view of individual gaze trace beneficial for syntactic sequence labeling in comparison to an aggregate *global* view, where information is traced via i) the central tendency (mean) or ii) the variability (variance) of the gaze behavior?

RQ2 How well does learning from decontextualized gaze data represented at the *type*-level (as dictionary) perform in comparison to learning from contextualized gaze data, represented at the *token*-level (via multi-task learning)?

Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), pages 51–61
Hong Kong, China, November 3, 2019.

Contribution The main contribution of this paper is to provide a systematic overview of the influence of two independent levels of gaze data aggregation on low-level syntactic labeling tasks at two separate levels of complexity; i.e., a simple chunk boundary tagging and a supervised POS-tagging task.

Our results support the claim that learning from gaze information under maximal (global) aggregation is more helpful than learning from less aggregated gaze representations across two corpora, two gaze metrics and two modelling setups.

However, we find that caution is warranted, as no single view, model or even gaze corpus show consistent improvement and the influence of single measures is not robust enough to identify a reliably helpful configuration for practical applications under the explored setups.

2 Background and Motivation

Eye movements during reading consist of fixations which are short stationary glances on individual words. These are interrupted by saccades, which are the ballistic movements between fixations. The gaze loosely traces the sequence of words in a text and gaze research in reading has at its basis the understanding that deviations from a monotone eye movement progression tend to occur when the reader's cognitive processing is being challenged by the text.

The raw gaze signal is a time series of (x, y)-coordinates mapped to word positons on the screen and clustered into consecutive fixations. This data must necessarily be pre-processed and radically filtered to fit the shape of any NLP problem (Holmqvist et al., 2011). NLP researchers therefore need to decide how to meaningfully aggregate and filter gaze corpora in order to construct a mapping of the time series onto the meaningful unit of the problem at hand, such as a sequence of words or sentences.

The most commonly applied feature extraction approach is based on counting durations of fixations, visits and re-visits per word as pioneered in the psycholinguistic tradition and most commonly aggregating to the global mean across multiple readers (see orange line in Figure 1).

An alternative paradigm to psycholinguistics-based feature extraction is to instead represent raw recorded scanpaths over entire word sequences as 2D or 3D matrices and images (von der Malsburg et al., 2012; Martínez-Gómez et al., 2012; Benfatto et al., 2016; Mishra et al., 2017a). However, this paradigm has only been explored in a jointly labeled setting where gaze data is assumed to be available at test time. This requirement can not yet be met in most practical NLP-applications.

Positive results have emerged from a range of diverging representations. In some cases, including tens of gaze features show a benefit (Mishra et al., 2017a; Barrett and Søgaard, 2015b) while other studies report successful experiments using a single gaze feature (Barrett et al., 2018a; Klerke et al., 2016).

The extraction of multiple features from the same bit of a raw recording can in theory allow to represent multiple distinct views on the same data; the number of visits, the order of visits and the durations of visits are examples of distinct perspectives. However, when features partially or entirely subsume other features[1] the inclusion of multiple views effectively performs a complex implicit weighting of the available gaze information through partial duplication. In order to eliminate these subtle effects, this work follow the single-metric approach, using a strict split of the recorded fixation durations into an Early and a Late measure with no overlap (see Section 4 for details). This allows us to isolate the target dimensions of inquiry, namely *effects of the level of aggregation*.

The two candidate gaze metrics used in this work are the first pass fixation duration as our Early metric and regression duration for our Late metric, which are the same two metrics as employed for sentence compression by Klerke et al. (2016). While they studied only a multi-task learning setup and one level of aggregation, we focus on multiple levels of aggregation and two NLP tasks.

The latter study represents a group of studies where individual readers' records are available at training time (i.e., multiple copies of the data with annotations obtained from different reading behaviour) rather than learning from the aggregate of multiple readers. This approach which involves a minimal level of aggregation is frequently applied where individual readers' cognition is of primary interest, such as categorizing individual language skill level or behaviour (Martínez-Gómez et al., 2012; Matthies and Søgaard, 2013;

[1] E.g. total reading time subsumes first pass reading time entirely.

Augereau et al., 2016; Bingel et al., 2018). Noticeably, the opposite approach of using maximally aggregated type-level representations which average all readings across all occurrences and all participants, has also been shown to contribute to improvements (Barrett et al., 2016; Bingel et al., 2018; Hollenstein et al., 2019). The effect of these two different views (global vs local) on the same task hence remained unexplored and is a gap we seek to fill in this paper.

We focus on the use of gaze for syntax-oriented NLP tasks, because human readers' eye movements reflect necessary language processing work, including syntax parsing, to reach comprehension of a text[2] (Rayner et al., 2006; Demberg and Keller, 2008). Multiple concurrent triggers within the reader as well as within the text may affect unconscious eye movement planning and execution. For this reason, psycholinguistic research favors maximal averaging, seeking to eliminate as much noise as possible. In contrast, NLP-systems primarily suffer from, and seek to handle, specific hard cases. This indicates that the variability in the gaze signal is valuable to retain for learning patterns in the data for disambiguation.

To answer the research questions, we first split the gaze duration data into an Early and a Late measure which form two distinct views of the gaze data. We operationalize between-subject variation as a local and a global aggregate as described in Section 4. We then relate gaze variation to the token and type-level context-modeling distinction afforded with a multi-task learning setup and a type dictionary setup, respectively, as detailed in Section 3. We evaluate on both a simplified and a typical low-level syntactic sequence labeling task described in Section 5. Finally we report our results and draw perspectives to related work in Sections 6 and 7 and conclude.

3 Token and type modelling – as multi-task learning and dictionary supervision

We contrast the impact of learning from token-level gaze information with learning from a type-level aggregated representation. A compelling argument for the token-level representation is that preserving context-specific information may allow a model to distinguish words and contexts which elicit more and less predictable gaze behavior. However, direct comparisons have demonstrated that the type-level global mean, which discards information on ambiguity, may be preferable (Barrett et al., 2016; Hollenstein and Zhang, 2019), which is somewhat surprising as the tasks require token-level disambiguation. Hence, we test this distinction for several aggregation ways and corpora, to shed more light on this modeling choice. The following describes the neural network modelling options which allow this comparison.

Multi-task learning (MTL) trains a neural network model to predict gaze data as an auxiliary task to the target task (Caruana, 1997). At training time, an input sentence is drawn from one of the task specific training sets. The relevant output is evaluated against the gold labeling and if a loss is recorded, parameters are updated accordingly. By forcing the two tasks to share parameters, the gaze information floats into the shared parameters through back-propagation. In this way, updates caused by losses on one task affect the activations and output on the other task.

Dictionary modelling trains a neural network model on a target task where the base representation of each word of an input sentence is concatenated with type-level gaze-derived features stored as a separate set of word embeddings, as further detailed in Section 5.2.

4 Eye-tracking Data

We extract gaze data from two large-scale eye-tracking corpora, the English part of the Dundee corpus (Kennedy et al., 2003) and the monolingual English reading part of the Ghent Eye-Tracking Corpus (GECO)[3] (Cop et al., 2017). Statistics of the corpora are provided in Table 1. The GECO corpus is more recent and contains more tokens (and utterances). The average sentence length is shorter compared to the Dundee corpus.

We use the English portion of the Dundee corpus which consists of recordings of 10 native English speakers' reading of 20 newspaper articles from *The Independent*. The text was presented for self-paced reading on a screen with a maximum of 5 lines at a time. The experimental setup included a set of comprehension questions after each article, re-calibration at every three screens and a

[2] For other tasks, non-linguistic aspects such as the reader's personal interest or emotional response to the reading material may be a primary argument for using gaze data.

[3] http://expsy.ugent.be/downloads/geco/

Genre:	GECO novel		Dundee news	
Readers:	14		10	
	Sents	Tokens	Sents	Tokens
Train	4,200	45,004	1,896	41,618
Dev	547	5,614	231	5,176
Test	574	5,792	243	5,206
Types	–	11,084	–	8,608

Table 1: Overview of the eye-tracking corpora. Type information is extracted from the training partition using the original white-space tokenization.

chin rest and bite bar to fixate head position during reading (Kennedy and Pynte, 2005). The extraction of our Early and Late metric use the original white-space tokenisation.[4]

From the GECO corpus we use the English monolingual reading data. This portion consists of recordings of 14 native English speakers' reading of the full novel *The Mysterious Affair at Styles* (Christie, 1920). The text was presented for self-paced reading on a screen with one paragraph or a maximum of 145 characters at a time. The novel was read in four sessions with comprehension questions and breaks after each chapter. Re-calibration was performed every 10 minutes or when drift was detected. The extraction of first pass duration is the `WORD_SELECTIVE_GO_PAST_TIME` feature in the published data and total regression duration is calculated as the total reading time from the feature `WORD_TOTAL_READING_TIME` minus the first pass duration.

No official train–dev–test split has been established for the corpora. We use the split of the Dundee corpus provided by Barrett et al. (2016): a training set containing 46,879 tokens/1,896 sentences, a development set containing 5,868 tokens/230 sentences, and a test set of 5,832 tokens/241 sentences. For GECO, we reserve the first and the last of every ten paragraphs as test and development sets, respectively.

4.1 Early and Late measures

For our Early metric of gaze processing we extract first pass fixation duration which is the total time spent looking on a word when it is first encountered, and be fore any subsequent words have been fixated, as the reader's gaze passes over the text. First pass duration may consist of several fixations accumulated up until the gaze leaves the word for the first time. When words are occasionally skipped on the first pass, null-values occur in the Early measure.

For our Late measure we use regression duration which is defined as the total time spent looking at the word on all later passes. We compute it as the total fixation time on the word minus the first pass duration, our Early measure. All words that were visited at most once will receive null values for this measure. These two metrics effectively split the total recorded reading time at every word with no overlap.

The duration measures are recorded in milliseconds with a minimum value defined by the lower fixation detection threshold as defined in the in the recording software (most commonly 40-50ms). This built-in offset and the reading speed variation between slow and fast readers, means that comparable relative variation (e.g., doubling in reading speed) within different readers contribute with different weight to the raw measure. In order to represent relative changes in reading speed consistently we standardize the raw durations of both metrics to each individual's average duration for each measure without counting null-values.

The standardization translates the raw measures into a record of how far a recorded duration on a given word is from the reader's typical time per word, expressed in standard deviations. Once standardized, the values are aggregated as described next.

4.2 Global and local views

As detailed in Section 2, Klerke et al. (2016) used each individual's gaze record as representation, which is a minimally aggregated gaze representation view that preserves the full width of participant's individual measures. Drawing from this view, we include a similar *local* view of the data. The local view collects the set of observed values for each type as a dictionary.

In contrast, the global view aggregates over the readings of all individuals. In particular, while the commonly-used mean is an estimate of a central tendency and produces a smoothed aggregate, variance is an estimate of how well the mean models the data and this measure is particularly sensi-

[4] We follow the extraction procedure described for the metrics *first pass duration* and *total regression to* in Barrett et al. (2016)

tive to outliers. We use these two aggregates as *global* views; one representing a hypothetical typical reader; our other novel aggregate is representing the extent to which the eye movements of multiple readers agreed on an underlying sample.

4.3 Binning

The local and global measures are split into 6 distinct categorical bins following Klerke et al. (2016) and outlined below. One bin is reserved for the *null* values while the central standard deviation is considered the *typical* duration and an additional half of a standard deviation on each side denotes *short* and *long* duration spans. Values outside the central two standard deviations are binned as *very short* and *very long*, respectively.

0. $x = null$, not seen.
1. $x < -1$ SD, very short duration.
2. -1 SD $\leq x < -0.5$ SD, short duration.
3. -0.5 SD $\leq x < 0.5$ SD, typical duration.
4. 0.5 SD $\leq x < 1$ SD, long duration.
5. 1 SD $\leq x$, very long duration.

The binned values assigned for two example sentences from each corpus are shown in Figure 2a–2d. Each subject's (local) Early and Late measures are shown as a translucent large dot: several dots in the same category are represented as darker dots, and between-subject mean (global) is included as small connected orange dots. The null values (purple) are not included in the global aggregate. Practically, this decision means that as long as a single participant spend time fixating or re-fixating a word, the information about any participants who do not spend time on this particular word is lost in the global aggregates.

Inspecting the figure reveals how the Early measure is the most variable: we observe many grey dots per word, and fewer words with no attention on first pass (pale purple dots). In contrast, the Late measure is frequently recorded as null, reflecting how most words are not revisited. Interestingly, the GECO data (right figures), even though it has more (14) participants, noticeably it shows more agreement and less spread of the Late measure compared to the Dundee data. This difference may be attributable to the difference in text genre, readability, reading task formulation or sample differences.[5]

The robust effects of word length, frequency and wrap-up are discernible in the examples shown in Figure 2. Specifically, long words such as "visiting" in Figure 2c and the sentence boundary for example at the end of Figure 2d have received more attention than surrounding words. The wrap-up effect occur at boundaries of syntactically meaningful units and mostly reflects the time needed for the cognitive processing to catch up to the eyes (Rayner, 1998).

5 Experiments

Our experiments focus on two levels of syntactic influence on gaze data. In order to leverage the wrap-up effect, we design a simplified chunk boundary detection task, modelled as a binary sequence labeling problem. The second task is the classic supervised POS-tagging task.

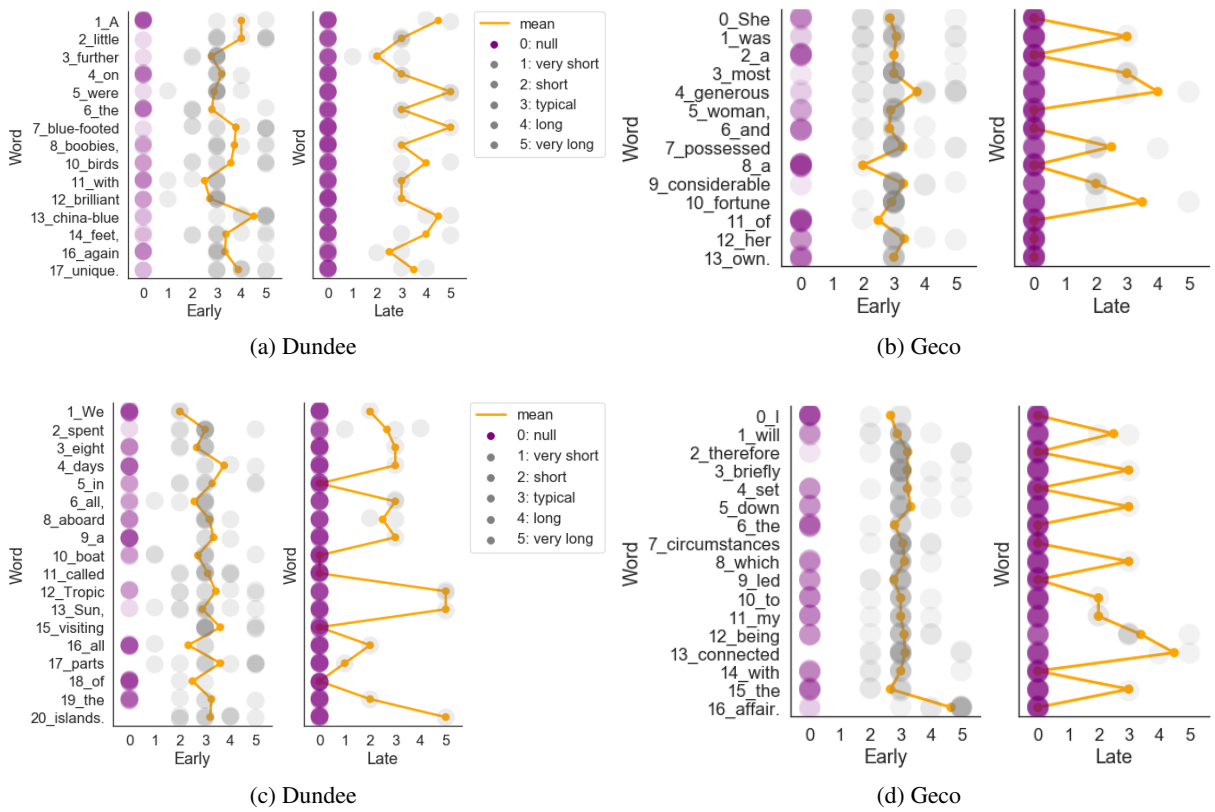
5.1 Data

Chunking data The chunk boundary data was extracted from the CoNLL2000 chunking data (Sang and Buchholz, 2000) which consists of 8,936 training sentences. Punctuation is not treated as separate tokens in gaze data, which is why we augment the CoNLL2000 data by combining punctuation with any preceding character and dropping its label. To isolate the boundary detection problem, we retain only the chunk prefixes B and I. That is, 315 distinct tokens were labeled O originally. Of these, O-labeled coordinating conjunctions were found in 2,803 sentences. We re-label these as B, positing that these conjunctions act as unary chunks between the boundaries of existing chunks. We proceed to drop all remaining sentences with any remaining O-labels, which leaves a dataset of 8,204 sentences, 91,8% of the original sentences. The new binary labels show a slightly un-balanced label distribution of 58.8% tokens labeled B. The test set after binarization has 1881 sentences corresponding to 93.5% of the original test set with the label B accounting for 58.5% of the tokens.

POS data We use the English part of the Universal Dependencies (UD version 2.1) POS tagging data built over the source material of the English Web Treebank.[6] The tagset spans the 17 universal POS tags. We use the standard splits provided

[5] The more recent GECO sample population is likely more accustomed to screen reading

[6] https://github.com/UniversalDependencies/UD_English-EWT



(a) Dundee (b) Geco

(c) Dundee (d) Geco

Figure 2: Example sentences from the two eye-tracking corpora.

by UD, which contains 12,543 sentences (204k tokens) for training, 2,002 sentences (25k tokens) for development and 2,077 sentences (25k tokens) for final evaluation. The development data is used for early stopping.

5.2 Model

In all our experiments, we use a bidirectional long short-term memory network (bi-LSTM) (Graves and Schmidhuber, 2005; Hochreiter and Schmidhuber, 1997; Plank et al., 2016) with a word encoding model which consists of a hierarchical model that combines pre-trained word embeddings with subword-character representations obtained from a recurrent character-based bi-LSTM.

For chunking and the MTL setup, we use the cascaded model proposed by (Klerke et al., 2016): it predicts the chunks at the outermost stacked-bi-LSTM layer of a 3-layer stacked network; and it predicts the gaze label at the first bi-LSTM layer. Note that our model differs from theirs in that we add a subword bi-LSTM at the character level, which has shown to be effective for POS tagging. Moreover, for POS we use a single bi-LSTM layer, hence the MTL setup reduces to a setup in which both tasks are predicted from the the single bi-LSTM layer. For dictionary modeling, we use the model proposed by (Plank and Agić, 2018) which integrates type-level information as lexicon embeddings concatenated to the word and sub-word level representations.

Hyperparameters For both tasks we tune model parameters on the development data for the respective task. We keep word embedding inputs fixed, which are set to 64 (size of the pre-trained Polyglot embeddings). We tune LSTM dimensions, character representations and hidden dimensions on the dev data. Early stopping was important to avoid overfitting of the auxiliary task.[7]

For binary chunking, the hyperparameters are: character and hidden dimensions 64, hidden layer size 100, cascaded model for MTL. For POS tagging, the parameters are: character input and hid-

[7]The gaze representation splits the data into unbalanced classes. The preliminary results indicated a tendency for the multi-task setup auxiliary task to learn only the majority class. With the implementation of a patience threshold for early model stopping, this was eliminated in all but the local view of the late measure which coincides with the experiments where gaze data is most detrimental to target task performance.

Baseline		Token-level		Type-level	
		94.93			
		Early	Late	Early	Late
Dundee	G: mean	94.81	94.48	94.89	94.94
	G: var	94.78	94.67	**95.01**	**94.98**
	L: union	93.79	94.13	94.80	94.93
GECO	G: mean	94.61	94.70	94.93	94.80
	G: var	94.40	94.57	94.91	94.91
	L: union	94.06	93.87	94.74	94.93

Table 2: F1 scores for binary chunking task training with an early or late gaze metric as an auxiliary task or as a type-level lexicon. G: global, L: local. Underlined: above baseline. Best per early/late: boldfaced.

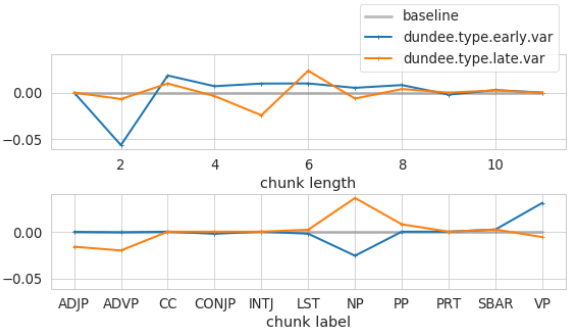


Figure 3: Relevance-weighted difference in F1 from baseline performance over chunk lengths and chunk labels for the Dundee data.

den dimension 64, hidden layer size 150. Both models were trained with Stochastic Gradient Decent (SGD) using a learning rate of 0.1, word dropout 0.25, and patience 2. The lexicon embedding size was tuned on Dundee data using the development data for both the Early and Late measure. For POS tagging the 40-dimensional lexical embeddings worked best for both Early and Late measure, similar to what was found for cross-lingual POS tagging (Plank and Agić, 2018). For chunking, the best result was obtained with 40 for Early and 70 for Late, respectively. In all experiments and tuning setups we average over 3 runs.

The chunking task is evaluated using phrase-based F1-score implemented by the conlleval script.[8] For POS tagging, performance is reported as tagging accuracy.

6 Results

6.1 Binary Phrase Chunking Results

Table 2 presents the results for the binary phrase chunking. Gaze data seems to provide little signal for this task. Over 2x12 setups, only the global (yet novel) view using variance at the type level provides a small increase in F1, but only on one gaze corpus.

Token vs type-level In more detail, for the chunking task the results show no benefit from learning gaze at token level in a multi-task setup (left two columns in Table 2). In all twelve MTL setups (two corpora, 2 gaze measures and three aggregations), no improvement is obtained over the baseline. In contrast, the type level dictionary-based gaze information is in all cases better than the token-level MTL, yet, results largely fall below baseline. In one specific setup the novel variance aggregation way, which holds over both the early and the late measure, results in the best gaze-based model (boldface). It results in a modest improvement, but it is not robust: the specific setup only works for Dundee, it does not carry over to the Geco corpus. We return to these results below.

Global vs local What is interesting to note is a clear negative effect of using un-aggregated (local) data: The local view consistently fails to improve over the no-gaze baseline on the chunk boundary detection task. This is in marked contrast to results on sentence compression (Klerke et al., 2016) (where Dundee local union helped in addition to integrating CCG tagging). Here, keeping individual readers' gaze information confuses the model and taking an aggregated view is more beneficial.

Analysis To assess the impact of the experimental conditions, we compare the performance of the two best setups across chunk length and over the underlying kinds of chunks (by relating predicted binary chunks back to their original labeling). Figure 3 depict the differences between these two models as the difference in F1-score relative to the baseline and weighted by the proportion of data accounted for by each subgroup.

The figures show how differences in performance on medium length chunks separate the two best chunk boundary detection models, despite their overall similar performance (95.01 vs 94.98). Early performs worse on short chunks (2 words long), while this is the case for longer chunks (5 words) for the regression-based Late measure.

With respect to chunk label, there is an interesting difference in performance with respect to

[8] github.com/spyysalo/conlleval.py

chunk category: the early measure outperforms the baseline on VP's; the late measure outperforms it on NP's (for which the two result in near mirror images). Note that this difference in the Early and Late metric is observed despite the fact that the chunk type information was not part of training. This points to the importance of analyzing performance in more detail to reveal differences between overall similarly-performing models.

6.2 Part-of-Speech Tagging Results

Table 3 shows the results for Part-of-Speech tagging. In contrast to binary chunking, gaze data provides signal for supervised POS tagging. There are several take-aways.

Token vs type-level Integrating the gaze data as type-level dictionary is the most beneficial and aids Part-of-Speech tagging, more than multi-task learning does. In particular, for the dictionary-based approach, we observe improvements in 9 out of 12 cases, yielding to up to +.23 absolute accuracy improvement. This shows that gaze data aids POS tagging also in our high-resource supervised POS tagging setup, which complements earlier findings restricted to unsupervised POS induction with naturally lower baselines (Barrett and Søgaard, 2015a; Barrett et al., 2016). MTL leads to a few but not consistent improvements for POS.

Global vs local Again, using the globally aggregated gaze measures is better than taking the local view of the individual readers. For both Dundee and GECO corpora, results for the local view approach fall below baseline in almost all cases. This holds for the local view in both setups, dictionary and MTL.

Analysis We analyzed the types of tags for which we observed the most improvement (or drop) in performance on the best model per corpus relative to proportion in data. For Dundee (G: mean) we observe that the model using the Late measure improves the most on content tags (adj, nouns) and misses the most on function words (pron, sym). Similarly for Geco (Early) most improvements are observed for content words including subordinate conjunctions (adj, sconj) while largest drops are on pronouns and numerals.

7 Related Work and Discussion

Klerke et al. (2016) proposed the applicability of single gaze metrics for improving sentence compression. Using the first pass duration and a regression duration measure in a multi-task learning setup, their study is, to the best of our knowledge, the only one to report a benefit from using the un-aggregated (local) data. Our study contributes to this research lacuna, where our results show that un-aggregated data is inferior (RQ1)—the detrimental effect might be partly due to a possible higher noise-to-signal ratio, disfavoring such setups.

		Token-level		Type-level	
Baseline		95.25			
		Early	Late	Early	Late
Dundee	G: mean	95.30	95.37	**95.35**	**95.48**
	G: var	95.21	95.33	**95.35**	95.44
	L: union	95.01	95.23	95.30	95.17
GECO	G: mean	95.23	95.27	95.34	95.35
	G: var	95.31	95.22	95.41	95.23
	L: union	94.97	95.23	95.14	95.26

Table 3: Accuracy scores for POS tagging with an early or late gaze metric as type-level lexicon or as an auxiliary task. G: global aggregation, L: local. Underlined: above baseline. Best per measure: boldfaced.

In contrast, Barrett and Søgaard (2015a) report benefits from aggregating the individual view of the data away, at first, and later demonstrate positive influence from aggregating also the individual tokens' context away, proposing the type-level view of gaze data for NLP (Barrett et al., 2016). Our results show that these type-level aggregates aid also supervised POS tagging, supporting further this type-level view. We here proposed a type-level view with novel global aggregation metrics and leveraging dictionary-based embeddings (Plank and Agić, 2018).

Recent related work on gaze in NLP rely to a greater extent on the strong emotional and affective gaze responses associated with the semantic content of a text. These works include the classification of sentiment (Mishra et al., 2017b; Hollenstein et al., 2019), coreferences (Jaffe et al., 2018; Cheri et al., 2016), named entities (Hollenstein and Zhang, 2019), sarcasm (Mishra et al., 2016) and multi-word detection (Yaneva et al., 2017; Rohanian et al., 2017).

8 Conclusions

We analyzed to which degree types of gaze aggregation over two distinct gaze measures impact on

syntactic tagging.

Our results show that gaze data from a single feature is informative for supervised POS tagging, complementing previous findings on unsupervised POS induction. Results on binary phrase chunking are however largely negative; only one specific setup led to a modest improvement. This points to the importance of evaluating across tasks, aggregation method and gaze corpus.

In particular, we found (RQ1) that the local view of gaze interaction traces was not helpful in comparison to a global view of either the mean or the variance computed over multiple participants. We could observe a clear detrimental effect of the local view for both tasks. To the best of our knowledge, only one prior study report a benefit for this view (cf. Section 7).

Regarding RQ2, our results show that the type-level dictionary-based learning from an aggregated representation leads to better representations than the token-based multi-task learning setup. Overall, our results support that POS-tagging benefits more from the gaze signal than the simplified chunk-boundary detection task. Inspection of the best models further indicated that the improvement was based on particular sensitivity to content word classes and phrases. These observations collectively agree well with the emerging picture that particular aspects of some content words are reflected more reliably in gaze data, compared to less semantically rich aspects of text.

The two corpora we use show quite different results which may be an effect of a number of differences, pointing to important future directions of work. The difference in genre and typical sentence length and, not least in number of unique entities, as discussed in Hollenstein and Zhang (2019), would very likely have affected readers to optimize their reading strategy, and thereby their oculomotor program, accordingly. The distance in time and technological maturity is likely to have some effects as well, albeit those are less testable. Overall, our findings point to the importance of analyzing overall performance measures in more detail and evaluating impact across different corpora and NLP tasks.

Acknowledgments

The authors would like to thank the current and previous anonymous reviewers for their thorough reviews. This research has been supported by the Department of Computer Science at IT University of Copenhagen and NVIDIA cooperation.

References

Olivier Augereau, Kai Kunze, Hiroki Fujiyoshi, and Koichi Kise. 2016. Estimation of English skill with a mobile eye tracker. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1777–1781. ACM.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018a. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312.

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, pages 579–584.

Maria Barrett, Ana Valeria Gonzlez-Garduo, Lea Frermann, and Anders Søgaard. 2018b. Unsupervised induction of linguistic categories with records of reading, speaking, and writing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2028–2038.

Maria Barrett and Anders Søgaard. 2015a. Reading behavior predicts syntactic categories. In *Proceedings of the nineteenth conference on computational natural language learning (CoNLL)*, pages 345–249.

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 (CogACL)*, pages 1–5.

Mattias Nilsson Benfatto, Gustaf Öqvist Seimyr, Jan Ygge, Tony Pansell, Agneta Rydberg, and Christer Jacobson. 2016. Screening for dyslexia using eye tracking during reading. *PloS one*, 11(12):e0165508.

Joachim Bingel, Maria Barrett, and Sigrid Klerke. 2018. Predicting misreadings from gaze in children with reading difficulties. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 24–34.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

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.

Agatha Christie. 1920. The mysterious affair at styles. 1920. *The Secret*.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.

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

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing nlp with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

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)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Evan Jaffe, Cory Shain, and William Schuler. 2018. Coreference and focus in reading times. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 1–9.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th European conference on eye movement*.

Alan Kennedy and Joël Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision research*, 45(2):153–168.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1528–1533.

Pascual Martínez-Gómez, Tadayoshi Hara, and Akiko Aizawa. 2012. Recognizing personal characteristics of readers using eye-movements and text features. *Proceedings of COLING 2012*, pages 1747–1762.

Franz Matthies and Anders Søgaard. 2013. With blinkers on: Robust prediction of eye movements across readers. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 803–806, Seattle, Washington, USA.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017a. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017b. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 377–387.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. In *AAAI*, pages 3747–3753.

Barbara Plank. 2016a. Keystroke dynamics as signal for shallow syntactic parsing. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 609–618.

Barbara Plank. 2016b. What to do about non-standard (or non-canonical) language in NLP. *CoRR*, abs/1608.07836.

Barbara Plank and Željko Agić. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 614–620, Brussels, Belgium. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422.

Keith Rayner, Kathryn H Chace, Timothy J Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading*, 10(3):241–255.

Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva, and Le An Ha. 2017. Using gaze data to predict multiword expressions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*, pages 601–609.

Erik F Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Titus von der Malsburg, Shravan Vasishth, and Reinhold Kliegl. 2012. Scanpaths in reading are informative about sentence processing. In *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*, pages 37–54.

Victoria Yaneva, Shiva Taslimipoor, Omid Rohanian, et al. 2017. Cognitive processing of multiword expressions in native and non-native speakers of english: Evidence from gaze data. In *International conference on computational and corpus-based phraseology*, pages 363–379. Springer.