A Cognitive Cost Model of Annotations Based on Eye-Tracking Data

Katrin Tomanek
Language & Information Engineering (JULIE) Lab
Universität Jena
Jena, Germany

Udo Hahn
Language & Information Engineering (JULIE) Lab
Universität Jena
Jena, Germany

Steffen Lohmann
Dept. of Computer Science & Applied Cognitive Science
Universität Duisburg-Essen
Duisburg, Germany

Jürgen Ziegler
Dept. of Computer Science & Applied Cognitive Science
Universität Duisburg-Essen
Duisburg, Germany

Abstract

We report on an experiment to track complex decision points in linguistic meta-data annotation where the decision behavior of annotators is observed with an eye-tracking device. As experimental conditions we investigate different forms of textual context and linguistic complexity classes relative to syntax and semantics. Our data renders evidence that annotation performance depends on the semantic and syntactic complexity of the decision points and, more interestingly, indicates that full-scale context is mostly negligible – with the exception of semantic high-complexity cases. We then induce from this observational data a cognitively grounded cost model of linguistic meta-data annotations and compare it with existing non-cognitive models. Our data reveals that the cognitively founded model explains annotation costs (expressed in annotation time) more adequately than non-cognitive ones.

1 Introduction

Today's NLP systems, in particular those relying on supervised ML approaches, are meta-data greedy. Accordingly, in the past years, we have witnessed a massive quantitative growth of annotated corpora. They differ in terms of the natural languages and domains being covered, the types of linguistic meta-data being solicited, and the text genres being served. We have seen large-scale efforts in syntactic and semantic annotations in the past related to POS tagging and parsing, on the one hand, and named entities and relations (propositions), on the other hand. More recently, we are dealing with even more challenging issues such as subjective language, a large variety of co-reference and (e.g., RST-style) text structure phenomena, Since the NLP community is further extending their work into these more and more sophisticated semantic and pragmatic analytics, there seems to be no end in sight for increasingly complex and diverse annotation tasks.

Yet, producing annotations is pretty expensive. So the question comes up, how we can rationally manage these investments so that annotation campaigns are economically doable without loss in annotation quality. The economics of annotations are at the core of *Active Learning* (AL) where those linguistic samples are focused on in the entire document collection, which are estimated as being most informative to learn an effective classification model (Cohn et al., 1996). This intentional selection bias stands in stark contrast to prevailing sampling approaches where annotation examples are randomly chosen.

When different approaches to AL are compared with each other, or with standard random sampling, in terms of annotation efficiency, up until now, the AL community assumed *uniform* annotation costs for each linguistic unit, e.g. words. This claim, however, has been shown to be invalid in several studies (Hachey et al., 2005; Settles et al., 2008; Tomanek and Hahn, 2010). If uniformity does not hold and, hence, the number of annotated units does not indicate the true annotation efforts required for a specific sample, empirically more adequate cost models are needed.

Building predictive models for annotation costs has only been addressed in few studies for now (Ringger et al., 2008; Settles et al., 2008; Arora et al., 2009). The proposed models are based on easy-to-determine, yet not so explanatory variables (such as the number of words to be annotated), indicating that accurate models of annotation costs remain a desideratum. We here, alternatively, consider different classes of syntactic and semantic complexity that might affect the cognitive load during the annotation process, with

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1158–1167,
Uppsala, Sweden, 11-16 July 2010.

the overall goal to find additional and empirically more adequate variables for cost modeling.

The complexity of linguistic utterances can be judged either by structural or by behavioral criteria. Structural complexity emerges, e.g., from the static topology of phrase structure trees and procedural graph traversals exploiting the topology of parse trees (see Szmrecsányi (2004) or Cheung and Kemper (1992) for a survey of metrics of this type). However, structural complexity criteria do not translate directly into empirically justified cost measures and thus have to be taken with care.

The behavioral approach accounts for this problem as it renders observational data of the annotators' eye movements. The technical vehicle to gather such data are eye-trackers which have already been used in psycholinguistics (Rayner, 1998). Eye-trackers were able to reveal, e.g., how subjects deal with ambiguities (Frazier and Rayner, 1987; Rayner et al., 2006; Traxler and Frazier, 2008) or with sentences which require re-analysis, so-called garden path sentences (Altmann et al., 2007; Sturt, 2007).

The rationale behind the use of eye-tracking devices for the observation of annotation behavior is that the length of gaze durations and behavioral patterns underlying gaze movements are considered to be indicative of the hardness of the linguistic analysis and the expenditures for the search of clarifying linguistic evidence (anchor words) to resolve hard decision tasks such as phrase attachments or word sense disambiguation. Gaze duration and search time are then taken as empirical correlates of linguistic complexity and, hence, uncover the *real* costs. We therefore consider eye-tracking as a promising means to get a better understanding of the nature of the linguistic annotation processes with the ultimate goal of identifying predictive factors for annotation cost models.

In this paper, we first describe an empirical study where we observed the annotators' reading behavior while annotating a corpus. Section 2 deals with the design of the study, Section 3 discusses its results. In Section 4 we then focus on the implications this study has on building cost models and compare a simple cost model mainly relying on word and character counts and additional simple descriptive characteristics with one that can be derived from experimental data as provided from eye-tracking. We conclude with experiments which reveal that cognitively grounded models outperform simpler ones relative to cost prediction using annotation time as a cost measure. Based on this finding, we suggest that cognitive criteria are helpful for uncovering the real costs of corpus annotation.

2 Experimental Design

In our study, we applied, for the first time ever to the best of our knowledge, eye-tracking to study the cognitive processes underlying the annotation of linguistic meta-data, named entities in particular. In this task, a human annotator has to decide for each word whether or not it belongs to one of the entity types of interest.

We used the English part of the MUC7 corpus (Linguistic Data Consortium, 2001) for our study. It contains *New York Times* articles from 1996 reporting on plane crashes. These articles come already annotated with three types of named entities considered important in the newspaper domain, *viz.* "persons", "locations", and "organizations".

Annotation of these entity types in newspaper articles is admittedly fairly easy. We chose this rather simple setting because the participants in the experiment had no previous experience with document annotation and no serious linguistic background. Moreover, the limited number of entity types reduced the amount of participants' training prior to the actual experiment, and positively affected the design and handling of the experimental apparatus (see below).

We triggered the annotation processes by giving our participants specific *annotation examples*. An example consists of a text document having one single *annotation phrase* highlighted which then had to be semantically annotated with respect to named entity mentions. The annotation task was defined such that the correct entity type had to be assigned to each word in the annotation phrase. If a word belongs to none of the three entity types a fourth class called "no entity" had to be assigned.

The phrases highlighted for annotation were *complex noun phrases* (CNPs), each a sequence of words where a noun (or an equivalent nominal expression) constitutes the syntactic head and thus dominates dependent words such as determiners, adjectives, or other nouns or nominal expressions (including noun phrases and prepositional phrases). CNPs with even more elaborate internal syntactic structures, such as coordinations, appositions, or relative clauses, were isolated from

their syntactic host structure and the intervening linguistic material containing these structures was deleted to simplify overly long sentences. We also discarded all CNPs that did not contain at least one *entity-critical* word, i.e., one which might be a named entity according to its orthographic appearance (e.g., starting with an upper-case letter). It should be noted that such orthographic signals are by no means a sufficient condition for the presence of a named entity mention within a CNP.

The choice of CNPs as stimulus phrases is motivated by the fact that named entities are usually fully encoded by this kind of linguistic structure. The chosen stimulus – an annotation example with one phrase highlighted for annotation – allows for an exact localization of the cognitive processes and annotation actions performed relative to that specific phrase.

2.1 Independent Variables

We defined two measures for the complexity of the annotation examples: The *syntactic* complexity was given by the number of nodes in the constituent parse tree which are dominated by the annotation phrase (Szmrecsányi, 2004).[1] According to a threshold on the number of nodes in such a parse tree, we classified CNPs as having either high or low syntactic complexity.

The *semantic* complexity of an annotation example is based on the inverse document frequency df of the words in the annotation phrase according to a reference corpus.[2] We calculated the semantic complexity score of an annotation phrase as $max_i \frac{1}{df(w_i)}$, where w_i is the i-th word of the annotation phrase. Again, we empirically determined a threshold classifying annotation phrases as having either high or low semantic complexity. Additionally, this automatically generated classification was manually checked and, if necessary, revised by two annotation experts. For instance, if an annotation phrase contained a strong trigger (e.g., a social role or job title, as with *"spokeswoman"* in the annotation phrase *"spokeswoman Arlene"*), it was classified as a low-semantic-complexity item even though it might have been assigned a high inverse document frequency (due to the infrequent word *"Arlene"*).

[1]Constituency parse structure was obtained from the OpenNLP parser (`http://opennlp.sourceforge.net/`) trained on PennTreeBank data.

[2]We chose the English part of the Reuters RCV2 corpus as the reference corpus for our experiments.

Two experimental groups were formed to study different contexts. In the *document context* condition the whole newspaper article was shown as annotation example, while in the *sentence context* condition only the sentence containing the annotation phrase was presented. The participants[3] were randomly assigned to one of these groups. We decided for this between-subjects design to avoid any irritation of the participants caused by constantly changing contexts. Accordingly, the participants were assigned to one of the experimental groups and corresponding context condition already in the second training phase that took place shortly before the experiment started (see below).

2.2 Hypotheses and Dependent Variables

We tested the following two hypotheses:

Hypothesis H1: *Annotators perform differently in the two context conditions.*

H1 is based on the linguistically plausible assumption that annotators are expected to make heavy use of the surrounding context because such context could be helpful for the correct disambiguation of entity classes. Accordingly, lacking context, an annotator is expected to annotate worse than under the condition of full context. However, the availability of (too much) context might overload and distract annotators, with a presumably negative effect on annotation performance.

Hypothesis H2: *The complexity of the annotation phrases determines the annotation performance.*

The assumption is that high syntactic or semantic complexity significantly lowers the annotation performance.

In order to test these hypotheses we collected data for the following dependent variables: *(a)* the annotation accuracy – we identified erroneous entities by comparison with the original gold annotations in the MUC7 corpus, *(b)* the time needed per annotation example, and *(c)* the distribution and duration of the participants' eye gazes.

[3]20 subjects (12 female) with an average age of 24 years (mean = 24, standard deviation (SD) = 2.8) and normal or corrected-to-normal vision capabilities took part in the study. All participants were students with a computing-related study background, with good to very good English language skills (mean = 7.9, SD = 1.2, on a ten-point scale with 1 = "poor" and 10 = "excellent", self-assessed), but without any prior experience in annotation and without previous exposure to linguistic training.

2.3 Stimulus Material

According to the above definition of complexity, we automatically preselected annotation examples characterized by either a low or a high degree of semantic and syntactic complexity. After manual fine-tuning of the example set assuring an even distribution of entity types and syntactic correctness of the automatically derived annotation phrases, we finally selected 80 annotation examples for the experiment. These were divided into four subsets of 20 examples each falling into one of the following complexity classes:

sem-syn: low semantic/low syntactic complexity
SEM-syn: high semantic/low syntactic complexity
sem-SYN: low semantic/high syntactic complexity
SEM-SYN: high semantic/high syntactic complexity

2.4 Experimental Apparatus and Procedure

The annotation examples were presented in a custom-built tool and its user interface was kept as simple as possible not to distract the eye movements of the participants. It merely contained one frame showing the text of the annotation example, with the annotation phrase being highlighted. A blank screen was shown after each annotation example to reset the eyes and to allow a break, if needed. The time the blank screen was shown was not counted as annotation time. The 80 annotation examples were presented to all participants in the same randomized order, with a balanced distribution of the complexity classes. A variation of the order was hardly possible for technical and analytical reasons but is not considered critical due to extensive, pre-experimental training (see below). The limitation on 80 annotation examples reduces the chances of errors due to fatigue or lack of attention that can be observed in long-lasting annotation activities.

Five introductory examples (not considered in the final evaluation) were given to get the subjects used to the experimental environment. All annotation examples were chosen in a way that they completely fitted on the screen (i.e., text length was limited) to avoid the need for scrolling (and eye distraction). The position of the CNP within the respective context was randomly distributed, excluding the first and last sentence.

The participants used a standard keyboard to assign the entity types for each word of the annotation example. All but 5 keys were removed from the keyboard to avoid extra eye movements for finger coordination (three keys for the positive entity classes, one for the negative "no entity" class, and one to confirm the annotation). Pre-tests had shown that the participants could easily issue the annotations without looking down at the keyboard.

We recorded the participant's eye movements on a Tobii T60 eye-tracking device which is invisibly embedded in a 17" TFT monitor and comparatively tolerant to head movements. The participants were seated in a comfortable position with their head in a distance of 60-70 cm from the monitor. Screen resolution was set to 1280 x 1024 px and the annotation examples were presented in the middle of the screen in a font size of 16 px and a line spacing of 5 px. The presentation area had no fixed height and varied depending on the context condition and length of the newspaper article. The text was always vertically centered on the screen.

All participants were familiarized with the annotation task and the guidelines in a pre-experimental workshop where they practiced annotations on various exercise examples (about 60 minutes). During the next two days, one after the other participated in the actual experiment which took between 15 and 30 minutes, including calibration of the eye-tracking device. Another 20-30 minutes of training time directly preceded the experiment. After the experiment, participants were interviewed and asked to fill out a questionnaire. Overall, the experiment took about two hours for each participant for which they were financially compensated. Participants were instructed to focus more on annotation accuracy than on annotation time as we wanted to avoid random guessing. Accordingly, as an extra incentive, we rewarded the three participants with the highest annotation accuracy with cinema vouchers. None of the participants reported serious difficulties with the newspaper articles or annotation tool and all understood the annotation task very well.

3 Results

We used a mixed-design analysis of variance (ANOVA) model to test the hypotheses, with the context condition as between-subjects factor and the two complexity classes as within-subject factors.

3.1 Testing Context Conditions

To test hypothesis H1 we compared the number of annotation errors on entity-critical words made

	above	before	anno phrase	after	below
percentage of participants looking at a sub-area	35%	32%	100%	34%	16%
average number of fixations per sub-area	2.2		14.1		1.3

Table 1: Distribution of annotators' attention among sub-areas per annotation example.

by the annotators in the two contextual conditions (complete document *vs.* sentence). Surprisingly, on the total of 174 entity-critical words within the 80 annotation examples, we found exactly the same mean value of 30.8 errors per participant in both conditions. There were also no significant differences in the average time needed to annotate an example in both conditions (means of 9.2 and 8.6 seconds, respectively, with $F(1, 18) = 0.116$, $p = 0.74$).[4] These results seem to suggest that it makes no difference (neither for annotation accuracy nor for time) whether or not annotators are shown textual context beyond the sentence that contains the annotation phrase.

To further investigate this finding we analyzed eye-tracking data of the participants gathered for the document context condition. We divided the whole text area into five sub-areas as schematically shown in Figure 1. We then determined the average proportion of participants that directed their gaze at least once at these sub-areas. We considered all fixations with a minimum duration of 100 ms, using a fixation radius (i.e., the smallest distance that separates fixations) of 30 px and excluded the first second (mainly used for orientation and identification of the annotation phrase).

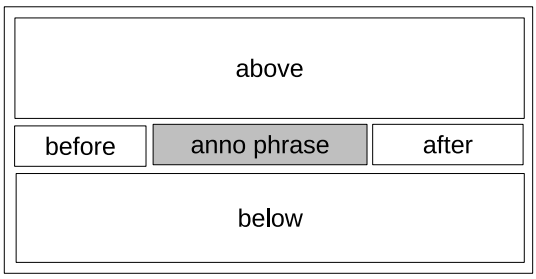

Figure 1: Schematic visualization of the sub-areas of an annotation example.

Table 1 reveals that on average only 35% of the participants looked in the textual context above the annotation phrase embedding sentence, and even less perceived the context below (16%). The sentence parts before and after the annotation phrase were, on the average, visited by one third (32% and 34%, respectively) of the participants. The uneven distribution of the annotators' attention becomes even more apparent in a comparison of the total number of fixations on the different text parts: 14 out of an average of 18 fixations per example were directed at the annotation phrase and the surrounding sentence, the text context above the annotation chunk received only 2.2 fixations on the average and the text context below only 1.3.

Thus, the eye-tracking data indicates that the textual context is not as important as might have been expected for quick and accurate annotation. This result can be explained by the fact that participants of the document-context condition used the context whenever they thought it might help, whereas participants of the sentence-context condition spent more time thinking about a correct answer, overall with the same result.

3.2 Testing Complexity Classes

To test hypothesis H2 we also compared the average annotation time and the number of errors on entity-critical words for the complexity subsets (see Table 2). The ANOVA results show highly significant differences for both annotation time and errors.[5] A pairwise comparison of all subsets in both conditions with a t-test showed non-significant results only between the SEM-syn and syn-SEM subsets.[6]

Thus, the empirical data generally supports hypothesis H2 in that the annotation performance seems to correlate with the complexity of the annotation phrase, on the average.

[4] In general, we observed a high variance in the number of errors and time values between the subjects. While, e.g., the fastest participant handled an example in 3.6 seconds on the average, the slowest one needed 18.9 seconds; concerning the annotation errors on the 174 entity-critical words, these ranged between 21 and 46 errors.

[5] Annotation time results: $F(1, 18) = 25$, $p < 0.01$ for the semantic complexity and $F(1, 18) = 76.5$, $p < 0.01$ for the syntactic complexity; Annotation complexity results: $F(1, 18) = 48.7$, $p < 0.01$ for the semantic complexity and $F(1, 18) = 184$, $p < 0.01$ for the syntactic complexity.

[6] $t(9) = 0.27$, $p = 0.79$ for the annotation time in the document context condition, and $t(9) = 1.97$, $p = 0.08$ for the annotation errors in the sentence context condition.

experimental condition	complexity class	e.-c. words	time mean	time SD	errors mean	errors SD	errors rate
document condition	sem-syn	36	4.0s	2.0	2.7	2.1	.075
	SEM-syn	25	9.2s	6.7	5.1	1.4	.204
	sem-SYN	51	9.6s	4.0	9.1	2.9	.178
	SEM-SYN	62	14.2s	9.5	13.9	4.5	.224
sentence condition	sem-syn	36	3.9s	1.3	1.1	1.4	.031
	SEM-syn	25	7.5s	2.8	6.2	1.9	.248
	sem-SYN	51	9.6s	2.8	9.0	3.9	.176
	SEM-SYN	62	13.5s	5.0	14.5	3.4	.234

Table 2: Average performance values for the 10 subjects of each experimental condition and 20 annotation examples of each complexity class: number of entity-critical words, mean annotation time and standard deviations (SD), mean annotation errors, standard deviations, and error rates (number of errors divided by number of entity-critical words).

3.3 Context and Complexity

We also examined whether the need for inspecting the context increases with the complexity of the annotation phrase. Therefore, we analyzed the eye-tracking data in terms of the average number of fixations on the annotation phrase and on its embedding contexts for each complexity class (see Table 3). The values illustrate that while the number of fixations on the annotation phrase rises generally with both the semantic and the syntactic complexity, the number of fixations on the context rises only with semantic complexity. The number of fixations on the context is nearly the same for the two subsets with low semantic complexity (sem-syn and sem-SYN, with 1.0 and 1.5), while it is significantly higher for the two subsets with high semantic complexity (5.6 and 5.0), independent of the syntactic complexity.[7]

complexity class	fix. on phrase mean	fix. on phrase SD	fix. on context mean	fix. on context SD
sem-syn	4.9	4.0	1.0	2.9
SEM-syn	8.1	5.4	5.6	5.6
sem-SYN	18.1	7.7	1.5	2.0
SEM-SYN	25.4	9.3	5.0	4.1

Table 3: Average number of fixations on the annotation phrase and context for the document condition and 20 annotation examples of each complexity class.

These results suggest that the need for context mainly depends on the semantic complexity of the annotation phrase, while it is less influenced by its syntactic complexity.

[7] ANOVA result of $F(1, 19) = 19.7$, $p < 0.01$ and significant differences also in all pairwise comparisons.

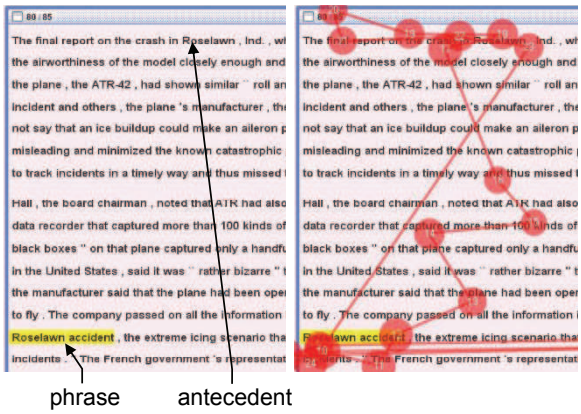

Figure 2: Annotation example with annotation phrase and the antecedent for *"Roselawn"* in the text (left), and gaze plot of one participant showing a scanning-for-coreference behavior (right).

This finding is also qualitatively supported by the gaze plots we generated from the eye-tracking data. Figure 2 shows a gaze plot for one participant that illustrates a scanning-for-coreference behavior we observed for several annotation phrases with high semantic complexity. In this case, words were searched in the upper context, which according to their orthographic signals might refer to a named entity but which could not completely be resolved only relying on the information given by the annotation phrase itself and its embedding sentence. This is the case for *"Roselawn"* in the annotation phrase *"Roselawn accident"*. The context reveals that Roselawn, which also occurs in the first sentence, is a location. A similar procedure is performed for acronyms and abbreviations which cannot be resolved from the immediate local context – searches mainly visit the upper context. As indicated by the gaze movements, it also became apparent that texts were rather scanned for hints instead of being deeply read.

4 Cognitively Grounded Cost Modeling

We now discuss whether the findings on dependent variables from our eye-tracking study are fruitful for actually modeling annotation costs. Therefore, we learn a linear regression model with time (an operationalization of annotation costs) as the dependent variable. We compare our 'cognitive' model against a baseline model which relies on some simple formal text features only, and test whether the newly introduced features help predict annotation costs more accurately.

4.1 Features

The features for the baseline model, character- and word-based, are similar to the ones used by Ringger et al. (2008) and Settles et al. (2008).[8] Our cognitive model, however, makes additional use of features based on linguistic complexity, and includes syntactic and semantic criteria related to the annotation phrases. These features were inspired by the insights provided by our eye-tracking experiments. All features are designed such that they can automatically be derived from *unlabeled* data, a necessary condition for such features to be practically applicable.

To account for our findings that syntactic and semantic complexity correlates with annotation performance, we added three features based on syntactic, and two based on semantic complexity measures. We decided for the use of multiple measures because there is no single agreed-upon metric for either syntactic or semantic complexity. This decision is further motivated by findings which reveal that different measures are often complementary to each other so that their combination better approximates the inherent degrees of complexity (Roark et al., 2007).

As for syntactic complexity, we use two measures based on structural complexity including *(a)* the number of nodes of a constituency parse tree which are dominated by the annotation phrase (cf. Section 2.1), and *(b)* given the dependency graph of the sentence embedding the annotation phrase, we consider the distance between words for each dependency link within the annotation phrase and consider the maximum over such distance values as another metric for syntactic complexity. Lin (1996) has already shown that human performance on sentence processing tasks can be predicted using such a measure. Our third syntactic complexity measure is based on the probability of part-of-speech (POS) 2-grams. Given a POS 2-gram model, which we learned from the automatically POS-tagged MUC7 corpus, the complexity of an annotation phrase is defined by $\sum_{i=2}^{n} P(\text{POS}_i|\text{POS}_{i-1})$ where POS_i refers to the POS-tag of the i-th word of the annotation phrase. A similar measure has been used by Roark et al. (2007) who claim that complex syntactic structures correlate with infrequent or surprising combinations of POS tags.

As far as the quantification of semantic complexity is concerned, we use *(a)* the inverse document frequency $df(w_i)$ of each word w_i (cf. Section 2.1), and a measure based on the semantic ambiguity of each word, i.e., the number of meanings contained in WORDNET,[9] within an annotation phrase. We consider the maximum ambiguity of the words within the annotation phrase as the overall ambiguity of the respective annotation phrase. This measure is based on the assumption that annotation phrases with higher semantic ambiguity are harder to annotate than low-ambiguity ones. Finally, we add the Flesch-Kincaid Readability Score (Klare, 1963), a well-known metric for estimating the comprehensibility and reading complexity of texts.

As already indicated, some of the hardness of annotations is due to tracking co-references and abbreviations. Both often cannot be resolved locally so that annotators need to consult the context of an annotation chunk (cf. Section 3.3). Thus, we also added features providing information whether the annotation phrases contain entity-critical words which may denote the referent of an antecedent of an anaphoric relation. In the same vein, we checked whether an annotation phrase contains expressions which can function as an abbreviation by virtue of their orthographical appearance, e.g., consist of at least two upper-case letters.

Since our participants were sometimes scanning for entity-critical words, we also added features providing information on the number of entity-critical words within the annotation phrase. Table 4 enumerates all feature classes and single features used for determining our cost model.

[8] In preliminary experiments our set of basic features comprised additional features providing information on the usage of stop words in the annotation phrase and on the number of paragraphs, sentences, and words in the respective annotation example. However, since we found these features did not have any significant impact on the model, we removed them.

[9] `http://wordnet.princeton.edu/`

Feature Group	# Features	Feature Description
characters (basic)	6	number of characters and words per annotation phrase; test whether words in a phrase start with capital letters, consist of capital letters only, have alphanumeric characters, or are punctuation symbols
words	2	number of entity-critical words and percentage of entity-critical words in the annotation phrase
complexity	6	syntactic complexity: number of dominated nodes, POS n-gram probability, maximum dependency distance; semantic complexity: inverse document frequency, max. ambiguity; general linguistic complexity: Flesch-Kincaid Readability Score
semantics	3	test whether entity-critical word in annotation phrase is used in document (preceding or following current phrase); test whether phrase contains an abbreviation

Table 4: Features for cost modeling.

4.2 Evaluation

To test how well annotation costs can be modeled by the features described above, we used the MUC7$_\mathcal{T}$ corpus, a re-annotation of the MUC7 corpus (Tomanek and Hahn, 2010). MUC7$_\mathcal{T}$ has time tags attached to the sentences and CNPs. These time tags indicate the time it took to annotate the respective phrase for named entity mentions of the types *person*, *location*, and *organization*. We here made use of the time tags of the 15,203 CNPs in MUC7$_\mathcal{T}$. MUC7$_\mathcal{T}$ has been annotated by two annotators (henceforth called *A* and *B*) and so we evaluated the cost models for both annotators. We learned a simple linear regression model with the annotation time as dependent variable and the features described above as independent variables. The baseline model only includes the basic feature set, whereas the 'cognitive' model incorporates all features described above.

Table 5 depicts the performance of both models induced from the data of annotator A and B. The coefficient of determination (R^2) describes the proportion of the variance of the dependent variable that can be described by the given model. We report adjusted R^2 to account for the different numbers of features used in both models.

model	R^2 on A's data	R^2 on B's data
baseline	0.4695	0.4640
cognitive	0.6263	0.6185

Table 5: Adjusted R^2 values on both models and for annotators A and B.

For both annotators, the baseline model is significantly outperformed in terms of R^2 by our 'cognitive' model ($p < 0.05$). Considering the features that were inspired from the eye-tracking study, R^2 is increased from 0.4695 to 0.6263 on the timing data of annotator A, and from 0.464 to 0.6185 on the data of annotator B. These numbers clearly demonstrate that annotation costs are more adequately modelled by the additional features we identified through our eye-tracking study.

Our 'cognitive' model now consists of 21 coefficients. We tested for the significance of this model's regression terms. For annotator A we found all coefficients to be significant with respect to the model ($p < 0.05$), for annotator B all coefficients except one were significant. Figure 6 shows the coefficients of annotator A's 'cognitive' model along with the standard errors and t-values.

5 Summary and Conclusions

In this paper, we explored the use of eye-tracking technology to investigate the behavior of human annotators during the assignment of three types of named entities – persons, organizations and locations – based on the eye-mind assumption. We tested two main hypotheses – one relating to the amount of contextual information being used for annotation decisions, the other relating to different degrees of syntactic and semantic complexity of expressions that had to be annotated. We found experimental evidence that the textual context is searched for decision making on assigning semantic meta-data at a surprisingly low rate (with

Feature Group	Feature Name/Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	855.0817	33.3614	25.63	0.0000
characters (basic)	token_number	-304.3241	29.6378	-10.27	0.0000
	char_number	7.1365	2.2622	3.15	0.0016
	has_token_initcaps	244.4335	36.1489	6.76	0.0000
	has_token_allcaps	-342.0463	62.3226	-5.49	0.0000
	has_token_alphanumeric	-197.7383	39.0354	-5.07	0.0000
	has_token_punctuation	-303.7960	50.3570	-6.03	0.0000
words	number_tokens_entity_like	934.3953	13.3058	70.22	0.0000
	percentage_tokens_entity_like	-729.3439	43.7252	-16.68	0.0000
complexity	sem_compl_inverse_document_freq	392.8855	35.7576	10.99	0.0000
	sem_compl_maximum_ambiguity	-13.1344	1.8352	-7.16	0.0000
	synt_compl_number_dominated_nodes	87.8573	7.9094	11.11	0.0000
	synt_compl_pos_ngram_probability	287.8137	28.2793	10.18	0.0000
	syn_complexity_max_dependency_distance	28.7994	9.2174	3.12	0.0018
	flesch_kincaid_readability	-0.4117	0.1577	-2.61	0.0090
semantics	has_entity_critical_token_used_above	73.5095	24.1225	3.05	0.0023
	has_entity_critical_token_used_below	-178.0314	24.3139	-7.32	0.0000
	has_abbreviation	763.8605	73.5328	10.39	0.0000

Table 6: 'Cognitive' model of annotator A.

the exception of tackling high-complexity semantic cases and resolving co-references) and that annotation performance correlates with semantic and syntactic complexity.

The results of these experiments were taken as a heuristic clue to focus on cognitively plausible features of learning empirically rooted cost models for annotation. We compared a simple cost model (basically taking the number of words and characters into account) with a cognitively grounded model and got a much higher fit for the cognitive model when we compared cost predictions of both model classes on the recently released time-stamped version of the MUC7 corpus.

We here want to stress the role of cognitive evidence from eye-tracking to determine *empirically relevant* features for the cost model. The alternative, more or less mechanical feature engineering, suffers from the shortcoming that is has to deal with large amounts of (mostly irrelevant) features – a procedure which not only requires increased amounts of training data but also is often computationally very expensive.

Instead, our approach introduces empirical, theory-driven relevance criteria into the feature selection process. Trying to relate observables of complex cognitive tasks (such as gaze duration and gaze movements for named entity annotation) to explanatory models (in our case, a time-based cost model for annotation) follows a much warranted avenue in research in NLP where feature farming becomes a theory-driven, explanatory process rather than a much deplored theory-blind engineering activity (cf. ACL-WS-2005 (2005)).

In this spirit, our focus has not been on fine-tuning this cognitive cost model to achieve even higher fits with the time data. Instead, we aimed at testing whether the findings from our eye-tracking study can be exploited to model annotation costs more accurately.

Still, future work will be required to optimize a cost model for eventual application where even more accurate cost models may be required. This optimization may include both exploration of additional features (such as domain-specific ones) as well as experimentation with other, presumably non-linear, regression models. Moreover, the impact of improved cost models on the efficiency of (cost-sensitive) selective sampling approaches, such as Active Learning (Tomanek and Hahn, 2009), should be studied.

References

ACL-WS-2005. 2005. *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*. accessible via `http://www.aclweb.org/anthology/W/W05/W05-0400.pdf`.

Gerry Altmann, Alan Garnham, and Yvette Dennis. 2007. Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, 31(2):685–712.

Shilpa Arora, Eric Nyberg, and Carolyn Rosé. 2009. Estimating annotation cost for active learning in a multi-annotator environment. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 18–26.

Hintat Cheung and Susan Kemper. 1992. Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, 13:53–76.

David Cohn, Zoubin Ghahramani, and Michael Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

Lyn Frazier and Keith Rayner. 1987. Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, 26:505–526.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *CoNLL 2005 – Proceedings of the 9th Conference on Computational Natural Language Learning*, pages 144–151.

George Klare. 1963. *The Measurement of Readability*. Ames: Iowa State University Press.

Dekang Lin. 1996. On the structural complexity of natural language sentences. In *COLING 1996 – Proceedings of the 16th International Conference on Computational Linguistics*, pages 729–733.

Linguistic Data Consortium. 2001. Message Understanding Conference (MUC) 7. Philadelphia: Linguistic Data Consortium.

Keith Rayner, Anne Cook, Barbara Juhasz, and Lyn Frazier. 2006. Immediate disambiguation of lexically ambiguous words during reading: Evidence from eye movements. *British Journal of Psychology*, 97:467–482.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 126:372–422.

Eric Ringger, Marc Carmen, Robbie Haertel, Kevin Seppi, Deryle Lonsdale, Peter McClanahan, James Carroll, and Noel Ellison. 2008. Assessing the costs of machine-assisted corpus annotation through a user study. In *LREC 2008 – Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 3318–3324.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 1–8.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS 2008 Workshop on Cost-Sensitive Machine Learning*, pages 1–10.

Patrick Sturt. 2007. Semantic re-interpretation and garden path recovery. *Cognition*, 105:477–488.

Benedikt M. Szmrecsányi. 2004. On operationalizing syntactic complexity. In *Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Vol. II*, pages 1032–1039.

Katrin Tomanek and Udo Hahn. 2009. Semi-supervised active learning for sequence labeling. In *ACL 2009 – Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 1039–1047.

Katrin Tomanek and Udo Hahn. 2010. Annotation time stamps: Temporal metadata from the linguistic annotation process. In *LREC 2010 – Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Matthew Traxler and Lyn Frazier. 2008. The role of pragmatic principles in resolving attachment ambiguities: Evidence from eye movements. *Memory & Cognition*, 36:314–328.