Getting at the Cognitive Complexity of Linguistic Metadata Annotation – A Pilot Study Using Eye-Tracking

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

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

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

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

Abstract

We report on an experiment where the decision behavior of annotators issuing linguistic metadata is observed with an eye-tracking device. As experimental conditions we consider the role of textual context and linguistic complexity classes. Still preliminary in nature, our data suggests that semantic complexity is much harder to deal with than syntactic one, and that full-scale textual context is negligible for annotation, with the exception of semantic high-complexity cases. We claim that such observational data might lay the foundation for empirically grounded annotation cost models and the design of cognitively adequate annotation user interfaces.

Keywords: Natural Language Metadata Annotation; Annotation Behavior; Eye-Tracking; Syntactic Complexity; Semantic Complexity; Cognitive Cost Modeling

Introduction

Supervised approaches to machine learning (ML) are currently very popular in the natural language processing (NLP) community. While linguistic regularities are no longer handcrafted by human experts in this paradigm, human intervention is still required to produce sufficient amounts of reliably annotated training material from which ML classifiers may learn or, considered as empirically valid ground truth, against which NLP systems can be evaluated.

The assignment of linguistic metadata (e.g., related to parts of speech, syntactic parses, or semantic interpretations) to plain natural language corpus data, a process called *annotation*, is a complex cognitive task. It requires a sound competence of the natural language in the corpus, as well as a decent level of domain and even text genre expertise.

Meanwhile lots of annotated corpora have been built which contain these precious human judgments (e.g., PennTreeBank (Marcus, Santorini, & Marcinkiewicz, 1993), PennPropBank (Palmer, Gildea, & Kingsbury, 2005) or OntoNotes (Pradhan et al., 2007)). Almost all of these annotated corpora were assembled by collecting the documents to be annotated on a random sampling basis (once the original document set had been restricted thematically or chronologically).

Only recently, more sophisticated approaches to select the annotation material are being investigated in the NLP community. One of the most promising approaches is known as *Active Learning* (AL) (Cohn, Ghahramani, & Jordan, 1996 ; Tomanek, Wermter, & Hahn, 2007) where an intentional selection bias is enforced and only those linguistic samples are selected from the entire document collection which are considered to be most informative to learn an effective classification model. When different approaches to AL are compared with each other, or with standard random sampling, in terms of annotation efficiency the AL community, up until now, assumed *uniform* annotation costs for each linguistic unit, e.g., words (Ringger et al., 2008 ; Settles, Craven, & Friedland, 2008 ; Arora, Nyberg, & Rosé, 2009). This claim, however, has been shown to be invalid in several studies (Hachey, Alex, & Becker, 2005 ; Settles et al., 2008 ; Tomanek & 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 have to be developed. Accordingly, we here consider different classes of syntactic and semantic complexity that might affect the cognitive load during the annotation process, with the overall goal to find 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 et Kemper (1992) for a survey of metrics of this type). However, structural complexity criteria do not translate directly into empirically justified cost measures.

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 & Rayner, 1987 ; Rayner, Cook, Juhas, & Frazier, 2006 ; Traxler & Frazier, 2008) or with sentences requiring re-analysis, so-called garden path sentences (Altmann, Garnham, & Dennis, 2007 ; Sturt, 2007).

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

[Federal Aviation Adminstration]$_{ORG}$ investigators were to examine the aircraft, said spokeswoman *[Arlene]*$_{PER}$. She said *[Martinair Holland]*$_{ORG}$ is certified to fly large jet aircraft into the *[US]*$_{LOC}$ as a scheduled passenger service.

When the *[Cessna]*$_{ORG}$ took off in rain and snow from the 6,900-foot runway at *[Cheyenne Municipal Airport]*$_{LOC}$ in *[Wyoming]*$_{LOC}$, *[Reid]*$_{PER}$ was seated at one control panel, *[Jessica]*$_{PER}$ was seated at another and her father was in a passenger seat in a four-seat *[Cessna]*$_{ORG}$ 177B, a 21-year-old single-engine plane owned by *[Reid]*$_{PER}$.

Figure 1: Text snippets taken from MUC7 documents annotated by *LOCation*, *PERson*, and *ORGanization* entity types.

Experimental Design

The focus of our study is on semantic annotation, the annotation of named entity mentions in particular. In this task, a human annotator has to decide for each word in a sentence whether it belongs to one of the entity types of interest or not. For the first time ever to the best of our knowledge, we applied eye-tracking to study the cognitive processes underlying the annotation of linguistic metadata.

We used the English part of the MUC7 corpus (Linguistic Data Consortium, 2001) for our study, which contains *New York Times* articles from the year 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". Figure 1 depicts typical text snippets from these articles along with the available annotations.

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 education 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 for 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, "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 given 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.

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 parse tree 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 $df(w_i)$ of each word w_i of the respective CNP according to a reference corpus.[2] We calculated the semantic complexity score as $max_i \frac{1}{df(w_i)}$, where w_i is the i-th word of the annotation phrase. Again, we determined a threshold classifying CNPs as having either high or low semantic complexity. This automatically generated classification was then 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 *"spokeswoman Arlene"*; cf. Figure 1), it was classified as a low-semantic-complexity item even though it was assigned a high inverse document frequency due to the infrequent word *"Arlene"*.

Two experimental groups were formed to study different kinds of textual context. 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

[1] Constituency parse trees were generated using the OpenNLP TreeBank parser (`http://opennlp.sourceforge.net/`).

[2] We chose the English part of the Reuters RCV2 corpus, a collection of over 400,000 news stories from 1996 and 1997, as the reference corpus for our experiments.

[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 practice and without previous exposure to academic linguistic education.

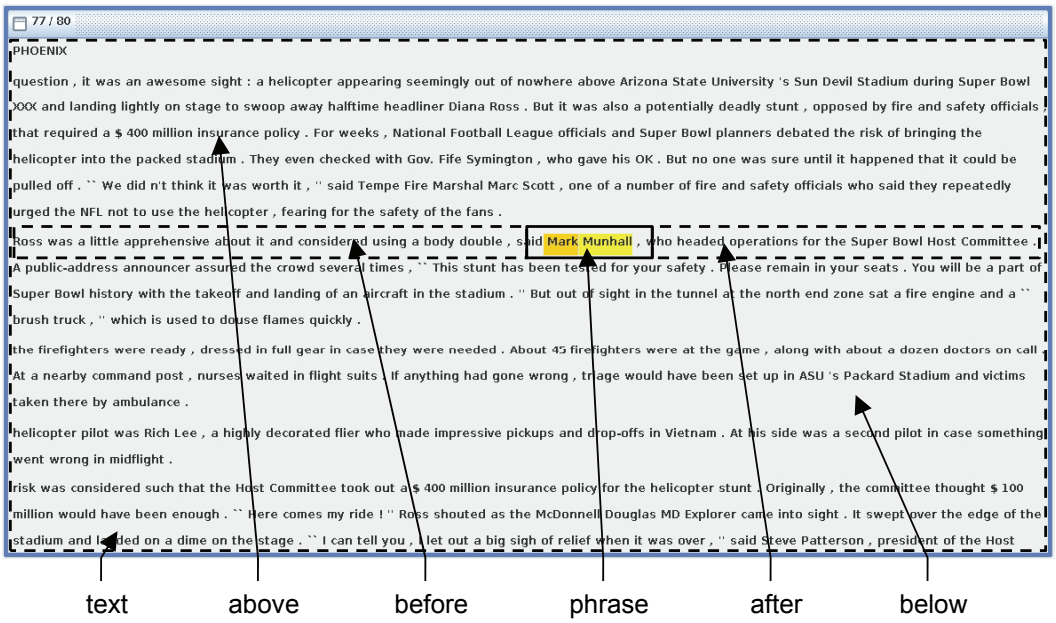

Figure 2: Subareas for the eyetracking analysis. Annotation example is of low semantic and low syntactic complexity.

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).

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 or de-anaphorization of entity classes. Accordingly, lacking context, an annotator is expected to annotate worse than under the condition of full context. As an adverse effect, the availability of (too much) context might overload and so distract annotators, with a potentially negative effect on annotation performance.

Hypothesis H2: *Annotators' performance is different for varying levels of syntactic and semantic complexity.*

The assumption is that high syntactic or semantic complexity, in contrast to low complexity, for both complexity types 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.

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

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 highlighted (as with *"Mark Munhall"* in Figure 2). A blank screen was shown after each annotation example to reset the eyes and to allow for 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 as a drawback due to extensive, pre-experimental training (see below). The limitation to 80 annotation examples reduced the chances of errors due to fatigue or lack of attention that can be observed in long-lasting annotation sessions.

subareas	above	left	*phrase*	right	below
percentage of participants looking at a subarea	35	32	100	34	16
average number of fixations in a subarea per participant	2.2		14.1		1.3

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

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 (text length was limited) to avoid the need for scrolling and thus eye distraction. The contextual position of the CNP 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 "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 each participant's eye movements on a Tobii T60 eyetracking 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 about 60 minutes were spent on annotation exercises. During the next two days, the actual experiments were conducted, each one lasting between 15 and 30 minutes, including calibration of the eye-tracking device. Another 20-30 minutes of training time directly preceded each individual experiment. After the experiment, the 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. The participants were also 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 either the newspaper articles or the annotation tool and all subjects agreed that they understood the annotation task very well.

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.

Testing Context Conditions

To test hypothesis H1 we compared the number of annotation errors on entity-critical words made 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 on 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 that contains the annotation phrase beyond the sentence.

To further investigate this finding we analyzed the eye-tracking data of the participants gathered for the document context condition. We divided the whole text area into several subareas as shown in Figure 2. We then determined the average proportion of participants that directed their gaze at least once at these subareas. 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 as it was mainly used for orientation and identification of the annotation phrase.

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 (see Table 1): 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, 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.

[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.

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 (e.-c.) 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).

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 repeated *t*-test measurements 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.

Context and Complexity

We also examined whether the need for inspecting the context increases with the complexity of the annotation phrase. So 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] 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.

This finding is qualitatively supported by gaze plots we generated from the eye-tracking data. Figure 3 shows such a plot for one participant which illustrates a scanning-for-coreference behavior we observed for many annotation phrases with high semantic complexity. Words were searched in the upper context, which according to their orthographic appearance might refer to a named entity, but which could not fully 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 also performed for acronyms and abbreviations which cannot be resolved from the immediate local context. As indicated by the gaze movements, it also became apparent that texts were rather scanned for hints instead of being deeply read.

complexity class	fixation on phrase mean	fixation on phrase SD	fixation on context mean	fixation 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.

Summary and Conclusions

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

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

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

[7] ANOVA result of $F(1,19) = 19.7$, $p < 0,01$ and significant differences also in all pairwise comparisons.

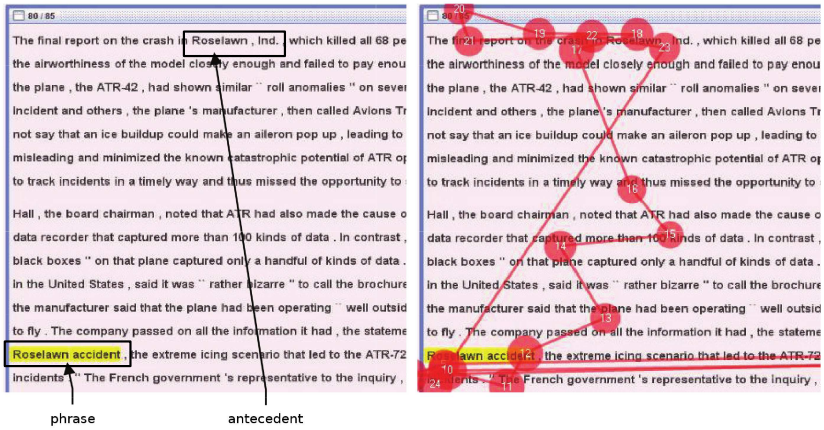

Figure 3: Annotation example with annotation phrase and the antecedent for *"Roselawn"* in the text (left), and gaze plot of one participant showing a scanning-for-coreference behavior (right).

The results of these experiments can be taken as a heuristic clue to focus on cognitively plausible features of learning empirically rooted cost models for annotation (see Tomanek, Lohmann, Ziegler, et Hahn (2010) for more details).

Références

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

Arora, S., Nyberg, E., & Rosé, C. (2009). Estimating annotation cost for active learning in a multi-annotator environment. In *NAACL HLT Workshop on Active Learning for Natural Language Processing* (pp. 18–26).

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, *13*, 53–76.

Cohn, D., Ghahramani, Z., & Jordan, M. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, *26*, 505–526.

Hachey, B., Alex, B., & Becker, M. (2005). Investigating the effects of selective sampling on the annotation task. In *CoNLL 2005 – 9th Conference on Computational Natural Language Learning* (pp. 144–151).

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

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: PENN TREEBANK. *Computational Linguistics*, *19*, 313–330.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106.

Pradhan, S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2007). Ontonotes: A unified relational semantic representation. In *ICSC 2007 – International Conf. on Semantic Computing* (pp. 517–526).

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

Rayner, K., Cook, A., Juhas, z. B., & Frazier, L. (2006). Immediate disambiguation of lexically ambiguous words during reading: Evidence from eye movements. *British Journal of Psychology*, *97*, 467–482.

Ringger, E., Carmen, M., Haertel, R., Seppi, K., Lonsdale, D., McClanahan, P., et al. (2008). Assessing the costs of machine-assisted corpus annotation through a user study. In *LREC 2008 – 6th International Conference on Language Resources and Evaluation* (pp. 3318–3324).

Settles, B., Craven, M., & Friedland, L. (2008). Active learning with real annotation costs. In *NIPS 2008 Workshop on Cost-Sensitive Machine Learning* (pp. 1–10).

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition*, *105*, 477–488.

Szmrecsányi, B. M. (2004). On operationalizing syntactic complexity. In *JADT 2004 – 7th International Conf. on Textual Data Statistical Analysis* (pp. 1032–1039).

Tomanek, K., & Hahn, U. (2010). Annotation time stamps: Temporal metadata from the linguistic annotation process. In *LREC 2010 – 7th International Conference on Language Resources and Evaluation.*

Tomanek, K., Lohmann, S., Ziegler, J., & Hahn, U. (2010). A cognitive cost model of annotations based on eye-tracking data. In *ACL 2010 – 48th Annual Meeting of the Association for Computational Linguistics.*

Tomanek, K., Wermter, J., & Hahn, U. (2007). An Approach to Text Corpus Construction which Cuts Annotation Costs and Maintains Reusability of Annotated Data. In *EMNLP/CoNLL 2007 – Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 486–495).

Traxler, M., & Frazier, L. (2008). The role of pragmatic principles in resolving attachment ambiguities: Evidence from eye movements. *Memory & Cognition*, *36*, 314–328.