Workshop *Predicting and Improving Readability*

**Time and place**
Thursday, April 5th 2018
13:15 - 17:00
Drift 21, 3512 BR Utrecht
Sweelinckroom (ground floor)

**Programme**
13:15 Walk in
13:30 Ted Sanders: Welcome
13:35 Scott Crossley
14.05 Orphée de Clercq
14.35 Carel Jansen
15.05 tea break
15.30 Suzanne Kleijn
16.00 Henk Pander Maat
16.30 General discussion
17.00 Drinks

**Abstracts**

**Scott Crossley**

*Crowdsourcing text complexity models*

The current study builds on work by De Clercq et al. (2014) and Crossley et al. (2017) by using crowdsourcing techniques to collect human ratings of text comprehension, processing, and familiarity across a large corpus comprising a diverse variety of topic domains (science, technology, and history). Pairwise comparisons among the ratings were calculated using a Bradley-Terry model to estimate the difficulty of each text in comparison to all the other texts. A number of linguistic features taken from state-of-the-art NLP tools were then used to develop models for text comprehension, text processing, and text familiarity. The accuracy of these models were then compared to classic readability formulas. A subset of the corpus was used in a follow up behavioral in which reading time and text comprehension were measured. We then modeled the reading times and comprehension scores collected from this study using the derived reading, comprehension, and familiarity formulas and classic readability formulas along with a number of other fixed effects including reading proficiency scores. Results indicated that formulas developed for text comprehension, processing, and familiarity explained greater variance than classic readability formulas. In addition, reading times and text comprehension scores in the behavioral study were predicted by newer models of text comprehension and processing but not by classic readability formulas.
Readability research has progressed significantly in recent years by incorporating natural language processing (NLP) and machine learning techniques (Collins-Thompson, 2014). We have explored in close detail the feasibility of constructing a readability prediction system for English and Dutch generic text using supervised machine learning (De Clercq and Hoste, 2016). Based on readability assessments by both experts and crowdsourcing (De Clercq et al. 2014), we have implemented different types of text characteristics ranging from easy-to-compute superficial text characteristics (average word length, average sentence length, ...), to features requiring deep linguistic processing (depth of syntactic tree, length of referential chain, semantic roles, ...). We will present the experiments we conducted using a wrapper-based genetic algorithm approach to perform readability prediction. We will show that this is a promising approach that provides considerable insights in which feature combinations contribute most to the overall prediction. We will also report on ongoing domain-adaptation experiments.


Carel Jansen

How to collect readability data: Advantages and disadvantages of different versions of the Cloze Test

My presentation will go into the question what kind of readability data we should or should not use. I will plea for using an Extended Version of the Classical Cloze Test (EVCCT) as the base for assessing the validity of existing reading assessment tools and for developing new, more advanced tools in this field. According to the EVCCT approach, always a series of N cloze tests derived from the same text should be used; in each of these tests every Nth word (either starting with the first, the second, the third, or the Nth word in the first part of the text) should be replaced by a line of the same length; and for measuring the outcomes of the test the Exact Scoring method should be applied. I hope to be able show the theoretical and practical advantages of using this version of the cloze test compared to, for instance, the Rational Deletion Cloze Test (with Exact or Semantic Scoring), the Multiple Choice Cloze Test, the C-Test, and the Cloze-Elide Test.
Suzanne Kleijn

**Generalizability of readability factors across Dutch speaking populations**

In my dissertation (Kleijn, 2018) I studied the effects of different linguistic features on the readability of texts for Dutch adolescents. Sixty texts were turned into cloze tests using the newly developed HyTeC-cloze procedure and all texts were carefully manipulated on one stylistic linguistic feature to create an ‘easy’ and ‘difficult’ version of the same text. As a result, causal effects of these linguistic features on readability could be separated from correlational relationships. Or in other words: we know how well these factors predict readability versus how much they can actually improve the readability of texts for Dutch adolescents. In the current study we look at the generalizability of these results with regard to other Dutch speaking populations. In two replication studies we collected comprehension data from Dutch and Flemish adults as well as data from Flemish adolescents. I will present the results of these studies and compare them to the earlier findings for Dutch adolescents.

Henk Pander Maat

**Comparing readability assessments based on comprehension data and comprehensibility ratings**

Readability diagnosis crucially depends on data. But what kind of data? Whereas classic readability prediction used cloze comprehension data, recent readability work often relies on comprehensibility rating. These ratings may have either been collected from experts or from target group readers. We may conceptually assess the construct validity of these data types as operational approximations of readability, but we may also ask to what extent these data sources yield different readability assessments. In this presentation, we will take the second approach by comparing two kinds of data for twenty texts. These texts have been cloze-tested among a sample of adult readers, and rated for comprehensibility among a different sample of adult evaluators. To what extent do these two data types yield the same readability rankings? And to what extent are comprehension performance scores and ratings predicted by the same text features?