Symbol supported News text on the Internet – A corpus-based approach

Katarina Mühlenbock
Dept of Swedish Language
University of Gothenburg

Johan Roxendal
Dept of Linguistics
University of Gothenburg

Janaina Rudberg
Dept of Linguistics
University of Gothenburg

Mats Lundälv
DART
Sahlgrenska University Hospital
Gothenburg

Abstract

This paper describes a students' project within the Educational Programme in Computational Linguistics, supervised by one of the authors. An editing tool adding symbol support to text adapted for poor readers was developed and evaluated. The results show that in a restricted domain and at a basic language level, lemma information solely seems to suffice for a reasonably correct linking of words to their corresponding representations as signs, symbols or speech.

1 Background

1.1 Impaired reading skill and AAC

A considerable proportion of the population, among 1.3% of all individuals (Beukelman and Mirenda, 2005), are affected by severe communication disorders, making them unable to use written or spoken language. Different language supportive aids have evolved over the years for this group, mainly as graphical systems containing pictures and symbols, nonverbal systems as sign language, or a combination of these, possibly also comprising speech synthesis and speech recognition. Technically, all these supportive measures and methods are referred to as Augmentative and Alternative Communication (AAC).

1.2 Web accessibility

Individuals with functional reading skill can benefit from written information presented as simplified "easy-to-read" text, adapted lexica and word lists. Recently, studies in the field of intellectual disability research have proved that adding AAC-symbols to adapted text can increase the understanding of written information substantially for these individuals (Jones et al, 2007; Poncelas and Murphy, 2007). Internet-based systems aiming to provide an alternative way to access web content are also rapidly developing, among these *NavigAbile* <www.navigabile.eu>, an EC-funded project in which information and news material are rendered accessible for a large group of people earlier excluded from this platform (Mühlenbock et al, 2008; Lundälv and Mühlenbock, 2008).

2 Project description

2.1 The editing tool

Among the services offered by the *NavigAbile* system, the News section has turned out to be one of the most frequently visited by the AAC users. Weekly news from the *8 PAGES* internet site <www.8sidor.se>, containing headlines from Sweden and other countries in an easy-to-read manner, have been linked to the system. The editing tool presented in this paper brings this approach a bit further by adding AAC symbol support to the content.

2.2 Basic assumptions

We postulate that the vocabulary drawn from an easy-to-read corpus by definition is compact, since there is a great difference in word-meaning redundancy between large corpora and a small specialized corpus as this one. Another assumption is that the corpus vocabulary largely matches the ideographic representations of concepts in the symbolic bliss language, or at least, that there is an acceptable coverage. We also assume that for this purpose, manual semantic disambiguation can outdo more sophisticated methods such as grammar-based translation (Lidskog, 2007) or semantic concept-coding of the lemma forms (Derbring, 2008).

3 Material

- News material, around 170,000 tokens, from the LäSBarT corpus (Mühlenbock, 2008), each token provided with either lemma form or proper name attribute, was used for extraction of the reference wordlist. Compound words in the list were assigned information about the linking of various elements. The wordlist contained ~ 12,000 lexems pertaining to ~ 5,200 lemmas.
- A library with jpg-pictures covering about 3,000 high-frequency concepts pertaining to the vocabularies in bliss was used for symbol representation.
- A library with pictures of persons, country flags and organization logos was used for representing common proper names.

4 Method

The human editor submits chunks of text from the News domain to the editing tool. In the uncomplicated case, a simple 1-1 mapping is performed, assigning the adequate graphical symbol to each token. This is simply achieved by looking up the lemma form in the lexicon. When ambiguities appear, a frequency check is made in order to establish the most likely form. Unknown compounds are handled by an analyzer trying to decompose the complex word by means of the possible compound elements in the lexicon, and adding two or more different symbols to one token. Frequent names are provided with a personal photo, a flag or a trademark. The result is presented to the human editor, who is able to manually correct the output before publishing it on the Internet (Fig 1).

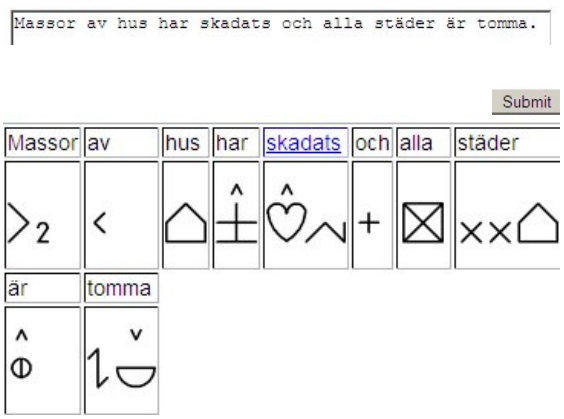

Figure 1. Web page output

5 Results and future work

Despite this apparently simplistic approach, the tool performs remarkably well. A sample of text drawn from a current issue of *8 PAGES* produced an output with a precision of 0.96, recall of 0.75 and F-score of 0.84. The performance might be improved by a range of measures, including access to larger symbol libraries and further context processing.

References

Beukelman, D. and P. Mirenda (2005). *Augmentative and alternative communication: Supporting children and adults with complex communication needs.* Baltimore, Paul H. Brookes.

Derbring, S. (2008). *Development of StoCC – linking AAC symbols to accurate concepts*. Master's Thesis at the Computatinal Linguistics Programme, Uppsala universitet, submitted.

Jones, F. W., K. Long, et al. (2007). Symbols can improve the reading comprehension of adults with learning disabilities. *Journal of Intellectual Disabilitiy Research* **51**: 545-550.

Lidskog, J. (2007). *Swedish Bliss – grammar based translation from Swedish into Bliss*. Master's Thesis at the Computational Linguistics Programme, Göteborgs universitet.

Lundälv, M. and K. Mühlenbock (2008). SYMBERED and NavigAbile - two means to achieve cross-language and symbol supported web accessibility. *13th Biennial Conference - Int Society for Augmentative and Alternative Communication*, Montreal, Canada.

Mühlenbock, K., M. Lundälv, et al. (2008). NavigAbile - Evaluation of services and methodologies for Internet exploitation by AAC users.*13th Biennial Conference - Inl Society for Augmentative and Alternative Communication*, Montreal, Canada.

Mühlenbock, K. (2008). Legible, readable or plain words - presentation of an easy-to-read Swedish corpus. *Readability and Multilingualism, workshop at 23rd Scandinavian Conference of Linguistics*. Uppsala, Sweden

Poncelas, A. and G. Murphy (2007). Accessible Information for People with Intellectual Disabilities: Do Symbols Really Help? *Journal of Applied Research in Intellectual Disabilities* **20**: 466-474.