Lund University, Dept. of Linguistics
Working Papers 49 (2001), 66–69

Linguistic structures for email and echat

Sheri Hunnicutt and Tina Magnuson
Dept. of Speech, Music and Hearing
sheri@speech.kth.se, tina@speech.kth.se

Abstract

Persons who use a symbol system to communicate have not been able to take advantage of Internet-based services to the same extent as persons who write using ordinary alphabetic text. However, such persons could be among those who would have most use of Internet services due to physical or intellectual handicaps and the resulting social and practical isolation. This contribution considers some known differences in communication for symbol system users compared to their counterparts whose written communication consists of ordinary text, discusses some differences between ordinary text and email or echat texts and argues the need for special linguistic structures to support Internet use for symbol system users.

1 Introduction

In order to serve the needs of a wide range of persons with communication, language and/or cognitive impairments in accessing the electronic highway, a European Union (IST) project has been defined. This project has recently begun in January 2001 and is called WWAAC, an acronym for World Wide AAC. The initials AAC have been used for many years to describe alternative and augmentative communication, that is, communication using methods other than those which are ordinary, either as a substitute for the ordinary methods (alternative) or as a supplement for the ordinary methods (augmentative). Many persons who employ AAC are users of symbol systems in which a single symbol, icon or picture is substituted for one or more words, reducing the amount of access time for persons with motor disabilities and providing a graphical method of expression for some persons with intellectual or language difficulties.

The tasks of the WWAAC project may be grouped into four key components: (1) task support research and development for overall ease of use of Internet services, (2) Internet applications in the form of a browser, an email program and an echat program, (3) a web-authoring tool and guidelines to assist web masters in ensuring that web sites are accessible for symbol system users, and (4) language support for communication. It is this latter area of language support in which our work will be carried out. Knowledge of the communication of symbol system users will form the base for the use of a concept set to enable communication between persons who use different symbol systems and between persons who use a symbol system but want to communicate with a person employing ordinary text. Mechanisms will be built to support the generation of messages.

Because the project has only begun, this paper will discuss the foundation for this work.

2 Communication by symbol users

Graphic symbols can be divided into two categories: a) pictorial symbols which have a high degree of iconicity with an intrinsic meaning, often constituting a close-ended system

or, b) generic, modular symbols where the meaning is given by the user or receiver, constituting an open-ended, linguistic-like generative system. The choice is dependent on the individual's cognitive and linguistic status. When constructing a system for a specific target group, the special circumstances affecting this group naturally have to be taken into account. Apart from the more obvious physical and/or speech impairment, these circumstances include additional factors intrinsic to the symbol user, such as linguistic and/or cognitive development delay, literacy difficulties and less experience in communication. Extrinsic factors affecting the individual may be restricted access to a functional AAC system or inability to operate the equipment independently. The individual may also have a language impairment affecting linguistic units at phonological, morphological, syntactical, lexical and pragmatic levels.

Furthermore, the individual must learn the specific language code of a graphic symbol system. This is complicated by the asymmetry between the input, which is auditory, and the output, which is graphic, causing the symbol user to employ an output modality different from the modality in which the language is learned. This situation is unique to symbol users, compared to a manual sign user, for example, who is most often taught language through manual signs.

Graphic symbols are not only an alternative to writing but also to speech, which means that speech output has to be introduced at a preliterate phase. On the other hand, aided communication intervention most often does not take place until the child is two years old or older, thus delaying language development considerably.

Graphic symbol message structures seem to be governed by two factors, namely that they are remarkably similar across subjects and significantly different from spoken output regardless of age and disability (Soto, 1998). Examples of differences are predominance of single symbol messages and word order differences. Furthermore, messages are often co-constructed with a partner.

3 Email and echat text

It is well known that email texts are often more like spoken language than like written language, resembling a spoken conversation. The parts of email texts that are not copied from other texts are usually written quickly. In spontaneous texts of these sorts, as is the case with spontaneous speech, writers do not produce perfect utterances. They often write partial sentences or only phrases or words in response to questions asked or opinions offered by a correspondent. The language used in echat discussions is even more atypical of ordinary text in written language. Echat texts may contain many abbreviations and are punctuated by symbol sequences that represent facial characteristics such as the smiles shown in Table 1 (Sjöberg, 2001).

Table 1. Some symbol-meaning correspondences from typical echat text (after Sjöberg, 2001).

Symbol	Meaning
:) or ☺	smile
;)	flirt
: –)	teasing smile
&-)	Glasses

4 Linguistic structures to support email and echat use

4.1 Lexicons

In order to be able to provide support for symbol system users to communicate with their choice of conversation partner, vocabulary for messaging must include symbols from each of the symbol systems to be supported as well as text counterparts in each of the natural language systems to be represented. An existing database, previously developed by one of the project partners, will be further developed for this use. The symbols in this database represent a number of symbol systems including Blissymbols, PCS, Beta, Pictograms, Beeldlezen, Rebus and ordinary photographs.

Symbols with similar concept correspondences have been linked in this database so that suitable substitutions can be made. This is not a simple matter, however, since symbols do not exist in each symbol set for each concept. Of a total of over 5000 concepts with symbolic representations in the database, 63% are covered by symbols in the PCS system, but only 17% in the Pictogram system. When users of different symbol systems desire to correspond with one another, the situation is further complicated. As can be seen in Figure 1, the best common coverage for any two of the three symbol systems shown is 26% for Blissymbols and Pictograms. For all three symbol systems, the intersection, or overlap, reduces to less than 11%.

Total number of concepts	**5218**	
	Number	Percentage
Concepts covered: Bliss&PCS&Pictograms	566	**10,85**
Concepts covered in Bliss & PCS	610	**11,69**
Concepts covered in Bliss & Pictograms	1359	**26,04**
Concepts covered in PCS & Pictograms	783	**15,01**
Concepts covered in Bliss	**2271**	**43,52**
Overlap with PCS	1359	59,84
Overlap with Pictograms	610	26,86
Concepts covered in PCS	**3287**	**62,99**
Overlap with Bliss	1359	41,34
Overlap with Pictograms	783	23,82
Concepts covered in Pictogram	**916**	**17,55**
Overlap with Bliss	610	66,59
Overlap with PCS	783	85,48

Bold percentages are relative to the total number of concepts.
Non-bold percentages are relative to the number of concepts in one database.

Figure 1. Concept coverage among symbol systems in a 5,000-concept database

Further ambiguities can be encountered when choosing glosses in each of the natural languages to correspond to each concept since one must choose among a variety of synonyms that can differ from one another, and from the corresponding symbols in each of the symbol systems, to various degrees.

4.2 Database of grammatical structures

Because email and echat messages differ from ordinary text, as described above, a collection of email and echat messages is now underway in the Swedish language. These messages will be tagged using the TnT tagger (Brants, 2000) and parsed with the SPARK parser (Aycock, 1998) complemented with a Swedish grammar. The frequency of the grammatical structures will then be compared with the grammatical structures found in a similar parsing of a sample of texts from the SUC Corpus (Ejerhed, Källgren, Wennstedt & Åström, 1992).

4.3 Phrase and sentence generation

Several methods have been sketched out for using the vocabulary and grammatical structures as described above in order to form phrases and sentences for symbolic communication. One of these methods, to be used for symbol-to-text communication, is already being implemented. It involves the adaptation and extension of grammatical rules previously developed and used with the communication aid Blisstalk that combines an overlay of Blissymbols with speech synthesis output (Hunnicutt, 1986).

5 Summary

There are a number of considerations that must be addressed in order to provide access to Internet services for persons who use symbol systems for their communication method. In a European project with this goal, a number of tasks are being undertaken to provide linguistic support for these persons. Included in these tasks are the further development of a database of concepts and their correspondences in a number of symbol systems, the collection, tagging and parsing of texts from email and echat environments, and the frequency determination of grammatical structures used in these texts. Several methods, grounded on a knowledge of symbol system use, are being considered for the development of functions to aid symbol users in the development of messages for these environments.

Acknowledgement

We would like to express our appreciation to Beata Megyesi for collaboration regarding tagging and parsing of Swedish texts and for the use of her grammar for the SPARK parser.

References

Aycock, John. 1998. *SPARK, Release 0.6*, Department of Computer Science, University of Victoria, Canada.

Brants, Thorsten. 2000. *TnT – A Statistical Part-of-Speech Tagger*. Department of Computational Linguistics, Saarland University. http://www.coli.uni-sb.de/~thorsten/tnt

Ejerhed, E., Källgren, G., Wennstedt, O. & Åström, M. (1992). *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project*, DGL-UUM-R-33, Department of General Linguistics, University of Umeå, REPORT NO. 33.

Hunnicutt, Sheri. (1986) 'Bliss Symbol-to-Speech Conversion: "Blisstalk",' *Journal of the American Voice I/O Society*, Vol. 3, June.

Sjöberg, Patrik. 2001. *Word prediction in an Internet chat*. Master's thesis, Department of Linguistics, Uppsala University, Sweden.

Soto, Gloria. in press. 'Understanding the impact of graphical symbol use on the message construction characteristics of individuals with severe speech impairments'. *Proceedings of The Fifth ISAAC Symposium on Research in Augmentative and Alternative Communication*, Dublin, Ireland, August 28-29, 1998.