ABSTRACT
An initial attempt has been made to include syntactic and semantic information in programs for word prediction. A study has begun in which 1500 words in a Swedish lexicon have been marked with semantic categories. Graphical semantic overviews and sketches of four texts have been composed using this classification. These graphical descriptions appear to be faithful to the subject matter. Small changes in prediction priority depending on semantic and syntactic information have not as yet resulted in keystroke savings. However, these are only the first of a number of planned uses of this information.

INTRODUCTION
Two word prediction programs have been developed since 1983 as speaking and writing aids [1]. One program, which is being used by persons experiencing extreme difficulties in movement, predicts a likely candidate word from the beginning letter(s) of a word. The candidate is chosen either from a variable-size frequency ranked lexicon or from a list of words previously used, thus taking both frequency and recency into account. The lexicon is updated during a session, so that it becomes individualized. Output can be presented as synthesized speech, and can be saved for later use or editing. Another output file is a work log which lists all key choices and the time elapsed between keystrokes. The program has been distributed to 14 locations in one of four languages (English, Swedish, Danish, Norwegian). It has recently been rewritten as a predicting editor which will be tested during the coming year. A new version will also be written to serve as a resident program which can be used with other programs.

A second program is now in trial form through consultation with several language therapists engaged in aphasia rehabilitation. With this program, a word is accessed through optional features. The program does not presuppose the ability to spell a word, but rather uses any (non-semantic) information that the person has about a word to "access" it. The word, or a list of word predictions, can then be synthesized for the user to hear and choose among. Lexical adaptation and output files are also included, as with the first system described.

SEMANTIC INFORMATION
A study has begun in which 1500 words in the 10,000-word Swedish lexicon have been marked with semantic categories. The words correspond to the standard international Blissymbol set and are marked with the categories assigned by the Easter Seal Communications Institute [2]. Semantic information is being used to optimize ranking in the list of predicted words. The main categories into which the words are classified (called Level 1 hereafter) are: 1) The World we Live in; 2) Living Things; 3) Being Alive; 4) How we View the World; and 5) Living Together. Several of these categories are divided into two or more subcategories (Level 2). Category 3, Being Alive, is divided into a) Things we Do and b) Things we Need. Category 4, How we View the World, is divided into a) Our Senses and b) Classification. Category 5, Living Together, is divided into four subcategories, a) Communication, b) Transportation, c) Occupations and d) Recreation. A further subcategorization divides the categories (or subcategories) into between three and eight (Level 3) subcategories. An example is given by the four subcategories of Living Together: Occupations, which is divided into (1) Primary Industries and Manufacturing, (2) Commerce, (3) Government and (4) Community Services. Words that belong to more than one category (or subcategory) are labelled as "general" words in that category (or subcategory). Using this categorization a "semantic overview" and a "semantic sketch" can be constructed. Such an overview has been computed for four short texts. Parts of these overviews will be shown below. A sketch will also be shown for one of the texts.

Semantic Content of Test Texts
Text 1 is from a Swedish story about a boy who is made small enough to ride on the back of a goose [3]. The passage taken here describes a troll he met on his journey who loved playing in the wind, and who could control the weather. The general semantic overview for this text is shown in Figure 1.

<table>
<thead>
<tr>
<th>5</th>
<th>10</th>
<th>6</th>
<th>13</th>
<th>13</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Categories:
0: General
1: The World we Live in
2: Living Things
3: Being Alive
4: How we View the World
5: Living Together

Figure 1: Semantic Overview, Text 1
Swedish story

Most semantically labelled words have occurred in the two categories Being Alive (13 words) and How we View our World (also 13 words). Level 3 categories which the most words fell into were Weather, Human Beings, Feeling, Size and Weight, and Place and Direction. Figure 2 shows a "semantic sketch" of Text 1.
A study was first done to determine the maximal possible savings in keystrokes if word class were known. In a 1331-word sample from a personal communicato text, it was found that perfect knowledge of part of speech would improve keystroke savings by 2.6 % for one prediction or by 5.1 % for six predictions (choices). In a similar test for English by Swiftin, Arnott, and Newell [4], a keystroke savings of between 0.5 % and 2.0 % was found for known word class. The texts for their study were three samples of 3,500 words each, covering three different subject areas. They used prediction lists of one, five and ten words (choices), and a lexicon containing precisely the words in the texts, marked for word class. A further test indicated that known word class would improve keystroke savings between 4.3 % and 6.4 % if frequency (rank) information were removed.

The fourth text was a report of the grounding of a Russian submarine off the southern coast of Sweden several years ago. Very few semantically marked content words occurred in this text. Those that did occur, however, were in the auspicious main category Living Together with several instances of words concerning Language (what had been said about the situation) and Government.

**Tests**

A study was first done to determine the maximal possible savings in keystrokes if word class were known. In a 1331-word sample from a personal communicato text, it was found that perfect knowledge of part of speech would improve keystroke savings by 2.6 % for one prediction or by 5.1 % for six predictions (choices). In a similar test for English by Swiftin, Arnott, and Newell [4], a keystroke savings of between 0.5 % and 2.0 % was found for known word class. The texts for their study were three samples of 3,500 words each, covering three different subject areas. They used prediction lists of one, five and ten words (choices), and a lexicon containing precisely the words in the texts, marked for word class. A further test indicated that known word class would improve keystroke savings between 4.3 % and 6.4 % if frequency (rank) information were removed.

The fourth text was a report of the grounding of a Russian submarine off the southern coast of Sweden several years ago. Very few semantically marked content words occurred in this text. Those that did occur, however, were in the auspicious main category Living Together with several instances of words concerning Language (what had been said about the situation) and Government.

**Tests**

A study was first done to determine the maximal possible savings in keystrokes if word class were known. In a 1331-word sample from a personal communicato text, it was found that perfect knowledge of part of speech would improve keystroke savings by 2.6 % for one prediction or by 5.1 % for six predictions (choices). In a similar test for English by Swiftin, Arnott, and Newell [4], a keystroke savings of between 0.5 % and 2.0 % was found for known word class. The texts for their study were three samples of 3,500 words each, covering three different subject areas. They used prediction lists of one, five and ten words (choices), and a lexicon containing precisely the words in the texts, marked for word class. A further test indicated that known word class would improve keystroke savings between 4.3 % and 6.4 % if frequency (rank) information were removed.

**SYNTACTIC INFORMATION**

In order to include syntactic information in prediction, our 10,000-word lexicons needed to be marked with word class, i.e., part of speech. The Swedish lexicon had already been marked for purposes of synthetic speech production. The English lexicon is currently being complemente by inclusion of word class for each of the 10,000 words.

A study was first done to determine the maximal possible savings in keystrokes if word class were known. In a 1331-word sample from a personal communicato text, it was found that perfect knowledge of part of speech would improve keystroke savings by 2.6 % for one prediction or by 5.1 % for six predictions (choices). In a similar test for English by Swiftin, Arnott, and Newell [4], a keystroke savings of between 0.5 % and 2.0 % was found for known word class. The texts for their study were three samples of 3,500 words each, covering three different subject areas. They used prediction lists of one, five and ten words (choices), and a lexicon containing precisely the words in the texts, marked for word class. A further test indicated that known word class would improve keystroke savings between 4.3 % and 6.4 % if frequency (rank) information were removed.
It seemed clear, however, in both the Swedish and English studies, that removing words with an inappropriate word class from a prediction list would improve a user's degree of comfort and pleasure in using a system. This aspect should certainly not be overlooked even if the percent savings in keystrokes appears small.

The first attempt at a grammar was not completely successful. It was a precedence type grammar, and was designed to accommodate grammatical information by simply ruling out any word with a word class considered unlikely by the algorithm. That is to say, the grammar had veto power. It was found that this veto power was too strong; correct words were vetoed almost as often as incorrect words were prohibited (and the correct word chosen) in another portion of the communicator text mentioned above. This text was quite irregular grammatically, but was probably exemplary of texts one could expect from a conversation. Thus, overall performance in keystrokes saved was not especially improved by supplying this grammatical information, even though constraints for standard noun and verb phrases were useful. It is to be assumed, however, that this problem is tightly coupled to single predictions. As soon as multiple predictions are allowed, veto power is no longer necessary as a strategy. It can be replaced by priority decisions. A new grammar will soon be added which takes advantage of multiple predictions.

CONCLUSIONS

Studies to determine the usefulness of semantic and syntactic information in word prediction have been initiated. The tests completed so far show only small savings in keystrokes for syntactic information, and no savings for low-level semantic information. It is planned to change the grammar to take advantage of present successful strategies for noun and verb phrases while not constraining choices as much as at present in other contexts. Semantic information will be further exploited by increasing the semantic context, allowing several possible predictions and allowing for semantic matches at more general semantic levels.

This work has been supported by a grant from the Swedish National Board for Technical Development.

REFERENCES


EUROSPEECH '89, Paris, France, September 1989