Profet, A New Generation of Word Prediction: An Evaluation Study

Alice Carlberger
Johan Carlberger
Tina Magnuson
M. Sharon Hunnicutt
Speech, Music and Hearing, KTH
100 44 Stockholm, Sweden
{alice,johanc,tina,sheri}@speech.kth.se

Sira E. Palazuelos-Cagigas
Santiago Aguilera Navarro
Ingenieria Electronica
Univ. Politecnica de Madrid
Ciudad Universitaria s/n
28040 Madrid, Spain
{sira,aguilera}@die.upm.se

Abstract

Profet, a word prediction program, has been in use for the last ten years as a writing aid, and was designed to accelerate the writing process and minimize the writing effort for persons with motor dysfunction. It has also proved to be beneficial in spelling and text construction for persons with reading and writing difficulties due to linguistic impairments. With higher linguistic demands on support for individuals with severe reading and writing difficulties, including dyslexia, the need for an improved version of Profet has arisen. In this paper, the new functionality will be presented, and the possible implications for support at different linguistic levels will be discussed. Results from an evaluation study with individuals with motoric dysfunction and/or dyslexia will be presented at the workshop in Madrid.

1 Functionality of the Current Version of Profet

Word prediction systems have existed since the early 1980s and were originally intended for the motorically disabled but later also for persons with linguistic impairments. Several different word prediction methods exist: Prediction can either be based on text statistics or linguistic rules. Some prediction programs also adapt to the user's language by using subject lexicons or learning modules. Among the first to develop word prediction programs for the PC were KTH with Predict (later Profet) (Hunnicutt, 1986) and ACSD with PAL (Swiffin et al, 1987) (Arnott et al, 1993). Programs for the Macintosh include Co:Writer, which is distributed by Don Johnston, Inc.

Profet is a statistically based adaptive word prediction program and is used as an aid in writing by individuals with motoric and/or linguistic disabilities, e.g., mild aphasia and dyslexia (Hunnicutt, 1986), (Hunnicutt, 1989a). The program has undergone several stages of development at KTH since 1982 and runs on PC and Macintosh with Infovox and Monologue speech synthesis. It is used in Sweden (Profet) and Great Britain (Prophet) but is also being localized into Danish, Bokmål (Norwegian), Dutch, Spanish, French, Russian, and Finnish. Upon typing at least one letter at the beginning of a word, the user is presented with a list of up to nine suggestions of likely word candidates. A word is chosen with the function key indicated to its right. However, if the intended word is not among the choices, the user can type the next letter of the target word, at which point he or she is presented with a new list of suggestions in the prediction window. Each time another letter is typed, a new list will be displayed, provided there is a match in the lexicon. A list of word suggestions is also presented after completion of a word, if that word is the first word in a pair in the bigram lexicon. However, when the user starts to type a new word, the predictor, being restricted to one information source at a time, solicits only the main lexicon, thus ignoring any previously typed word. The negative effect of this restriction is counterbalanced to a certain degree by the recency function, which, after each space and punctuation, records the word just completed. In this manner, a recently used word is promoted in the prediction list the next time the first letter(s) is/are typed.

By selecting words in the prediction window, the motorically disabled user can economize keystrokes and physical energy. Similarly, the user who has difficulties spelling but is able to recognize the intended word in a list, is relieved of having to spell the whole word. However, the user who has problems with decoding can elect to have the prediction list spoken by the speech synthesizer, which can also speak letters, words, sentences or paragraphs written by the user.

The present version of Profet is strictly frequency-based and solicits three information sources, one at a time, namely, the main lexicon with some 10,000 word unigrams; the word bigram lexicon containing approximately 3000 reference words with up to nine frequency-ordered successors each; and the user lexicon, which adapts the main lexicon with the user's own words and words that have a rank exceeding 1000. Moreover, the user can create his own subject lexicons for classification of vocabulary according to topic, e.g., music, computers, and stamp collecting.

2 Testing the Current Version of Profet

First of all, a study conducted by a speech pathologist with a number of subjects will be presented. Then follow two quantitative studies without subjects.

Profet, previously called Predict, has been evaluated for several years, initially together with individuals with slow and laborious writing stemming from a motoric dysfunction. As slow writing speed is often believed to be a very important issue for individuals with motoric impairments, its main purpose was to accelerate the writing process. In an effort to systematically investigate the aid provided by this program, a study was conducted in which time-saving and effort-saving were chosen as parameters. Time-saving was measured as the number of output characters produced during a given time, and efficiency as a decrease in the number of keystrokes for a given text. Eight persons with motor disabilities participated in the study, six with cerebral palsy and one with a muscular disease, two of them also evidencing writing difficulties of a linguistic nature.

A "single-case design" was used. Prior to introduction of word prediction to the writer, a baseline was established during repeated sessions with texts written without any writing support. This made it possible to compare texts written with vs without Profet. The baseline test consisted of two tasks: a) to copy from a given text and b) to write about a topic that was chosen freely before the test began. Tests of the same type were then administered at three separate sessions with two months of training between each test.

The degree of improvement relating to speed and efficiency was found to vary considerably among subjects depending on their underlying writing abilities and which strategies they employed. With subjects A and B, the number of characters in text per minute increased and the total number of keystrokes decreased, as expected. Subject C, however, was too fast a typist to benefit from the program. Subject D, who was not extremely slow, felt that the program helped her because it forced her to use a more efficient typing strategy. For subject E, who was extremely slow and very easily exhausted, the program had only begun to have an effect but was expected to continue to improve performance even after the study had ended. However, contrary to our expectations, subject F, who had a severe motor disability, showed no improvement. For subject G, the only difference was decreased writing speed. Lastly, although the improvements exhibited in subject H were small, they motivated him to increase his writing significantly.

In summary, the results of this first study indicate that a) there was most often a reduction of keystrokes, which meant less effort; b) a reduction in the number of keystrokes did not necessarily mean a savings in time; c) the writing strategy had to be changed due to a higher cognitive load on the writing process, i.e., the time-saving gained by fewer keystrokes was consumed by longer time looking for the right alternative, which involved shifting one's gaze from the keyboard to the screen and back to the keyboard, then having to make a decision and hit the right key; d) speed was not the most important aspect to the user, but the effort-saving (as typing is often very laborious for a person with a motor impairment; one comment was: "I get less exhausted when I write with Profet"), and the possibility of producing more correct texts; e) the written texts were often better spelled and, on the whole, had a better linguistic structure, which was an unexpected, positive finding; f) a typical Profet error that occurred was when the subject chose an incorrect prediction (This type of error, where the word is spelled correctly but completely unrelated to the context, gives the text a bizarre look, and the text actually ends up being more unintelligible than if the word had merely been misspelled. However, the improvement in spelling outweighs this problem); and g) the possibility of adding speech synthesis to the other functions of Profet was an important and helpful feature to severely dyslectic individuals. The implication of these findings is that the effect and efficiency of a writing aid of this type to a great extent is dependent upon the underlying writing strategy and skills of the user.

Two subjects that participated in the speed enhancement evaluation study turned out to have severe writing difficulties at different linguistic levels: the character level (spelling errors), morphological level (agreement and occasional inflection errors), and syntactic level (incorrect word order, poor grammatical variability and incorrect handling of function words).

Of the two subjects who had difficulties with spelling and text construction, one showed substantial improvement and the other showed moderate improvement but reported a significant difference in ease of writing. These results indicated the power of prediction techniques as linguistic support for writing and stimulated the interest for the present focus on use of word prediction for persons with reading and writing difficulties and/or dyslexia. In a follow-up study, the potential to use the program as a support for spelling and sentence construction was also investigated by comparing spelling and word choice as well as qualitative aspects such as intelligibility and general style. Subsequent studies have included individuals with writing difficulties due to linguistic and/or dyslectic difficulties as well. In these linguistically oriented studies, the focus has been on spelling and morphosyntactic improvement or strategy changes. Qualitative aspects of the texts, such as intelligibility and stylistics, were judged by readers uninitiated as to the purpose of the study. To summarize the findings from this follow-up study: the use of Profet resulted in considerably better spelling, not much morphological improvement, inclusion of the usually non-existent function words, and more correct word order as well as positive subjective experiences such as "Profet helps me write more independently."

Recently, two strictly quantitative comparative studies without subjects were also performed. In the first one, which was a preliminary test conducted at our laboratory, the Swedish, British English, Danish, and Norwegian versions of Profet were run automatically with a statistical evaluation program on text excerpts approximately 6000 characters in length. The results are presented in Table 1, where *Preds* is the number of suggestions presented in the prediction window, *Chars* the number of characters in the text, *Keys* the number of keystrokes required with word prediction, and *Saved* the keystroke savings expressed as a percentage of the number of keystrokes that would have been required, had word prediction not been used. As can be seen, keystroke savings range roughly from 33% to 38% for 5 predictions, and from 35% to 42% for 9 predictions. The cross-language variations in the results could stem from several factors, one undoubtedly being an unfortunate non-reversible character conversion error for "ø", which, for Danish, resulted in predictions with the letter "o" and, for Norwegian, no predictions, for words with this character. A more linguistically valid factor would be differences in morphosyntactic language typology. For instance, the lower keystroke savings in Swedish compared to English might be explained in part by the fact that compounding (the formation of a new word, i.e., string, through the concatenation of two or more words) is a highly productive word creation strategy in Swedish, but not in English. Another factor might be the difference in test text style, the Swedish consisting of adolescent literature with a sizable amount of dialogue, the English of newspaper text from the electronic version of the Daily Telegraph, and the Danish and Norwegian of articles on language teaching. Likewise, the style of the texts from which the lexica were built must be taken into consideration. The Swedish lexicon was created from a 4 million-running-word balanced corpus augmented with a 10,000 word-frequency list and a 6,500 high-school word-list. The English lexicon was also built from a balanced corpus of some 4 million words, while the Danish was derived from a conglomerate of some 132,000 running words of newspaper text, prose, research reports, and legal and IT texts. The Norwegian lexicon was created from a 4 million-word corpus with a similar composition.

The second study, conducted at the Universidad Politecnica de Madrid within the VAESS project, analyzed, on the one hand, keystroke savings obtained with different prediction systems that had been tested at various research sites, and, on the other hand, factors affecting keystroke savings (See also Boekestein, 1996). The lack of standardization of test conditions prevented any cross-linguistic or cross-product comparison of keystroke savings.

The predictors included in the study were the Dutch (Boekestein, 1996) and Spanish (VAESS version) versions of Profet, and JAL-1 and JAL-2 for Spanish. Results from a test by Higginbotham (Higginbotham, 1992) of five word prediction systems were included as well; the systems were EZ Keys (Words, Inc.), Write 100, Predictive Linguistic Program (Adaptive Peripherals), Word Strategy (Prentke Romich Company & Semantic Corporation), and GET, all of which seem to have been tested on American or British English. Keystroke savings for these systems are presented below.

Factors affecting keystroke savings are test text size, test text subject (lexicon coverage), prediction method, maximum number of prediction suggestions, method for selecting prediction suggestions, amount of time needed to write the test text,

Language	*Preds*	*Chars*	*Keys*	*Saved*
Swedish	5	6068	4057	33.1%
Swedish	9	6068	3934	35.2%
British English	5	4107	2577	37.3%
British English	9	4107	2429	40.9%
British English	5	2640	1682	36.3%
British English	9	2640	1610	39.0%
Danish	5	4853	3254	32.9%
Danish	9	4853	3112	35.9%
Danish	5	3315	2060	37.9%
Danish	9	3315	1909	42.4%
Norwegian	5	4112	2648	35.6%
Norwegian	5	2619	1720	34.3%
Norwegian	9	6731	4117	38.8%

Legend:
Preds = maximum number of prediction suggestions
Chars = number of characters in test text
Keys = number of keystrokes required with word prediction
Saved = keystroke savings in percent of keystrokes required to write test text without word prediction

Table 1: Keystroke Savings with the Swedish, British English, Danish, and Norwegian Versions of Profet

and type of interface. An example is the difference between an interface with automatic row-and-column scanning, which requires two keystrokes to select a letter, and an interface with linear scanning and keystrokes on a keyboard, which requires only one keystroke per letter. Differences in morphosyntactic typology should logically also influence keystroke savings. Relevant examples are inflectional paradigm size and word order flexibility. Spanish, for instance, has both a significantly larger verb inflection paradigm and a freer word order than English.

Keystroke savings are here presented for the various prediction systems. First of all, with the Dutch version of Profet, they varied between 35% and 45%, depending on the setting of the test parameters. In the testing of the Spanish VAESS version of Profet, savings were 50.34% - 51.3% for texts with lengths of 2300 - 3100 characters and the number of prediction suggestions set to 5. With the number of suggestions set to 10, the savings were 53.71% - 55.14%. It should be noted that the test texts belonged to the same corpus from which the lexicon had been built, thus assuring good lexicon coverage. For perfect adaptation of lexicon to test text, maximum savings of around 70% were obtained. The input method used was linear scanning. Testing JAL-1, JAL-2 for Spanish with frequency-based prediction yielded savings of 56.55% and 60.61%, with the number of predictions set to 5 and 10, respectively. Testing the same system with syntactic prediction with automaton yielded savings of 57.83% and 61.63 % with 5 and 10 predictions, respectively. With syntactic prediction based on the char parsing method, the savings were 58.47% with 5 predictions and 61.84% with 10. Information on test text size was unavailable for this system. For the following five predictors, no information on test conditions was available: EZ Keys 45%, Write 100 45%, Predictive Linguistic System 41%, Word Strategy 36%, and GET 31%.

3 Why a New Version of Profet?

The current project started in July 1995 and originated through the search for new applications, the desire for more accurate prediction and enhancement of the pedagogical aspects of the user interface. The goal of our research is a grammatically more accurate prediction, psychological user support, and integration with spellchecking developed by HADAR in Malmö, Sweden, into a writing support tool for dyslexics. The project is funded by the National Labour Market Board (AMS), The Swedish Handicap Institute (HI), and the National Social Insurance Board (RFV).

4 Hypothesis

Our hypothesis is that certain aspects of the disabled individual's writing will improve with the appropriate use of, and training with, the new version of Profet with its augmented functionality. The purpose of this study is to find out a) if the user's spelling can be improved further by integrating Profet with a spellchecker, b) if the user's use of morphology (including the presence of required endings, the choice of endings and degree of agreement) improves with extension of scope and addition of grammatical tags, and c) if the subjects will approve of the predictions to a higher extent after incorporation of semantic tags.

Test results of a first version of the new Profet show an increase in keystroke savings compared with the current version. (See **Testing the New Version of Profet** below). However, as previously mentioned, there is also a qualitative, non-quantifiable aspect to writing that has to be evaluated.

5 Description of the New Version of Profet

To date, the modifications of the prediction system include extension of scope, addition of grammatical and semantic information as well as automatic grammatical tagging of user words. To accommodate the weighting of multiple information sources, the strictly frequency-based program has been replaced by one based on probabilities. Furthermore, an efficient lexicon development algorithm has been developed, facilitating the creation of new lexica, from either untagged or grammatically tagged text.

The word lexicons (unigrams and bigrams) were created with the new lexicon creation algorithm from a union corpus of the 300,000-word subset of the Stockholm-Umeå Corpus (SUC)[1], while awaiting the forthcoming 1 million-word final version, and a 150 million-word conglomerate of electronic texts[2], including running text from newspapers, legal documents, novels, adolescent literature, and cookbooks. For comparison with the present version of Profet, the size of the new lexicons was set to 7000 words and 14,000 bigrams, respectively.

Grammatical and/or semantic knowledge has been used in advanced systems worldwide since the early 1990s (Tyvand and Demasco, 1993) (Guenthner et al, 1993) (Guenthner et al, 1993a) (Booth, Morris, Ricketts and Newell, 1992) and has proven able to increase communication rate (Arnott et al, 1993) (Tyvand and Demasco, 1993) (Le Pévédic and Maurel, 1996). The grammatical information that was added to our system consisted of a set of 146 grammatical tags based on that of SUC. The tag statistics for the database were derived from the SUC subset. Tag unigram (146), bigram (5163), and trigram (43,862) lexicons were created with the same lexicon-creating algorithm as the word lexicons. The inclusion of trigrams involved an extension of scope compared with the current version of Profet. Another new feature is the automatic grammatical classification of user words, which is based on n-gram statistics.

Thirdly, a tentative effort was made to incorporate semantic information about the noun phrase into the prediction algorithm. Four semantic categories were established for nouns and adjectives: *inanimate, animate, human,* and *inanimate behaving as human,* an example of the latter being "company" as in "The company laid off 20% of its employees." The unigram word lexicon was then hand-tagged and prediction tests run, *with* vs *without* semantic information. As stated earlier, the addition of semantic information was not motivated by a desire for further keystroke savings (Hunnicutt, 1989b). Rather, the goal was to promote coherent thinking in the writing process by demoting semantically incongruous word choices. As expected, fewer of these words appeared in the list of suggestions, and no keystroke savings were gained. In fact, the results exhibited a 1% decrease in savings, which seems to have two explanations. First of all, the addition of semantic tags increased the total number of tags from 146 to 338, resulting in sparser training data. Secondly, the semantic tagging was done statically, i.e., each word received one and only one semantic tag, independent of context. A large percentage of the words belonged to all four categories. It would therefore be useful to expand the semantic classification system.

6 Testing the New Version of Profet

Preliminary quantitative tests of the new prediction system were run with an evaluation program developed at the laboratory. This was done *without* vs *with* an increasing number of grammatical tag types: (1) unigrams, (2)unigrams and bigrams, and (3) unigrams, bigrams, and trigrams. The test texts consisted of two types: a 10,000-word section of a novel of which the rest was used in the development of the lexicon of the predictor, and a 7500-word collection of essays written by students at the Stockholm Institute of Education and not used in the lexicon development. Each of the text types was divided into a 1000-word section and a 5000- word section, each of which was contained within the larger. The

[1] Currently available on CD-ROM through the European Corpus Initiative (ECI).

[2] Sources: Språkdata 24 million words, SRF Tal & Punkt 37 million words, Göteborgsposten 5 million words, and Pressens Bild 100 million words.

test results seem to indicate that the most significant keystroke savings are furnished by the grammatical bigrams: at least 7.4% over the grammatical unigrams, whose minimum savings amount to a mere 3.1% compared to prediction without any grammatical information. The most substantial savings are scored by the grammatical bigrams in the four largest texts: 27.3% - 33.6% in the essay texts (non-lexicon-corpus) and 16% in each of the novel texts (lexicon corpus). Unexpectedly, grammatical trigrams do not appear to add more than 1% in savings, at the most, over bigrams. However, further testing is needed. They are expected to at least be of a qualitative value to the user.

In our present study, the aim of which is the comparison between the current and new versions of Profet, a test design similar to the one described in the two evaluation studies above will be used. A baseline based on their current method of writing will be established prior to the introduction of the new Profet version. Test tasks will include dictation and free writing. The subjects must be linguistically competent enough to benefit from the different features of the new version of Profet, i.e., able to make a choice. When the inflections of a specific word are presented visually or aurally, the subject must be able to distinguish between the forms and make the correct selection. Two subjects with motoric dysfunction and reading and writing difficulties and five persons with dyslexia will participate in the evaluation of the new version. The two subjects with motoric dysfunction have participated in the earlier studies and are well acquainted with computers and writing support. A baseline based on the current version of Profet has already been established. Our goal, then, is to compare texts written by these two individuals with the current vs new version, respectively, of Profet. The five subjects with dyslexia have reading and writing difficulties as their main problem. Therefore, speed and efficiency will not be studied. Tentative results from the Profet evaluation will be presented at the workshop in Madrid in July 1997.

References

Arnott, J., Hannan, J.M., and Woodburn, R.J. 1993. Linguistic Prediction for Disabled Users of Computer-Mediated Communication. In The Swedish Handicap Insitute, editor, *Proceedings of the ECART2 Conference*, section 11.1. Kommentus, Stockholm, Sweden.

Boekestein, M. 1996. Word Prediction. M. A. thesis, Department of Language and Speech, Katholieke Universiteit Nijmegen, the Netherlands, August 1996.

Booth, L., Morris, C., Ricketts, I.W., and Newell, A.F. 1992. Using a syntactic word predictor with language impaired young people In H.J. Murphy, editor, *Proceedings of the California State University, Northridge (CSUN), 7th Annual Conference on Technology and Persons with Disabilities, Los Angeles, California, USA*, pp. 57–61, Office of Disabled Student Services, California State University, Northridge, California, USA, 18-21 March 1992. [CPRC/Ref.JLA/MAA00068.004]

Guenthner, F., Krüger-Thielmann, K., Pasero, R., Sabatier, P. 1993. Communication Aids for Handicapped Persons In The Swedish Handicap Institute, editors, *Proceedings of the ECART2 Conference, Stockholm, 1993, sect. 1.4.*

Guenthner, F., Langer, S., Krüger-Thielmann, K., Pasero, R., Sabatier, P. 1993. KOMBE. Communication Aids for the Handicapped. CIS Report 92-55, Munich, Germany.

Higginbotham, D. J. 1992. Evaluation of keystroke savings across five assistive communication technologies. In *Augmentative and Alternative Communication*, 8, pages 258–272. At the web site of the National Center to Improve Practice (NCIP), http://www.edc.org/FSC/NCIP

Hunnicutt, S. 1986. Lexical Prediction for a Text-to-Speech System. In E. Hjelmquist and L.-G. Nilsson, editors, *Communication and Handicap: Aspects of Psychological Compensation and Technical Aids.* Elsevier Science Publishers, Amsterdam, Netherlands.

Hunnicutt, S. 1989. ACCESS: A Lexical Access Program. In *12th Annual Conference of RESNA*, pages 284–285, New Orleans, Louisiana, June 25-30. RESNA Press, Washington, D.C.

Hunnicutt, S. 1989. Using Syntactic and Semantic Information in a Word Prediction Aid. In *Eurospeech 89*, Vol. 1, pages 191–193, Paris, France.

Le Pévédic, B. and Maurel, D. 1996. La prédiction d'une catégorie grammaticale dans un système d'aide à la saisie pour handicapés. Actes TALN, Marseille, France.

Swiffin, A.L., Arnott, J.L., and Newell, A.F. 1987. The use of syntax in a predictive communication aid for the physically impaired. In Steele/Garrey, editors, *Proceedings of the Tenth Annual Conference on Rehabilitation Technology*, pages 124 - 126. RESNA Press, Washington, D.C.

Tyvand, S. and Demasco, P. 1993. Syntax statistics in word prediction. In The Swedish Handicap Insitute, editor, *Proceedings of the ECART2 Conference*, section 11.1. Kommentus, Stockholm, Sweden.