A Methodological Framework for Writing Assistance Systems: Applications to Sibylle and VITIPI Systems.

Philippe Boissière (*), Igor Schadle[1](*), Jean-Yves Antoine (**)
(*) **I**nstitut de **R**echerches en **I**nformatique de **T**oulouse, Toulouse III University
118 Route de Narbonne, F- 31 062 TOULOUSE Cedex 9 France
(**) LI - François Rabelais, Tours University
Campus Universitaire de Blois, 3 pl. Jean Jaurès, F- 41.000 BLOIS, France
(*) {boissier, schadle}@irit.fr, (**) Jean-Yves.Antoine@univ-tours.fr

Abstract: This paper concerns the evaluation of Writing Assistance Systems (WAS). More precisely, it focuses on the evaluation of the linguistic process (word prediction) of these systems. We show that most of present evaluation frameworks are not really fitted to the real needs of disabled. A methodological and a common evaluation framework of WAS. is then proposed. This evaluation is based on an objective metric which is easy to compute andto interpret, as well as an "ecological" collection of test data (i.e. texts wrote by a disableb people). This framework has been applied in an evaluation campaign that involved two systems: Sibylle and VITIPI. These experiments are described into details and results are given as well. Finally, we suggest future prospects for this evaluation framework.

Keywords: Writing Assistance System, evaluation, word prediction, word completion.

I Introduction

Writing is very important for everyone, in everyday life. Unfortunately, some disabled people are unable to write with an ordinary pencil. Thanks to a computer, some of them can write easier, but special (specific or adapted) devices and/or software are needed to restore some communicative abilities of the user. This is the aim of Writing Assistance Systems (WAS) which suffer unfortunately from an extreme slowness and to be less convenient than an ordinary pencil.

The use of an ordinary or a specialized keyboard depends on the user's abilities. If the user can still control a certain extend his/her gestures, a physical keyboard can be adapted. Devices such as "finger guide" or "grid keyboard" could also be added on the keyboard in order to avoid the user to select several keys concurrently in a single keystroke. If physical keyboards cannot be used, virtual or onscreen keyboards should be well adapted. The virtual keyboard is displayed on a screen, and the user has just to point out the desired keypad with its most convenient pointing device (mouse, trackball, eye tracking ...). Various virtual keyboards can be found such has

[1] During its post Ph D at IRIT Laboratory

Clavicom[2], Wivik[3], Keystrokes [1][4]. Besides, software like Sokeyto [2] allows the user to create his/her own customized virtual keyboard. Unfortunately, some disabled subjects are unable to use a pointing device; they can only press on a contactor with a single switch. In such difficult situations, the key selection is achieved through a virtual scanning. If the system uses a linear scanning (SibyLettres [3] (Cf. § IIIa)), a cursor highlights successively each key. The user hits when reaching the desired key in only **one step**. With a row-column scanning, the selection takes **two steps**, first the row and then the key in the row. The main problem with such these systems comes from their extreme slowness. To allow faster typing, word prediction (or completion) techniques have been developed. This paper focuses on the evaluation of these techniques. It will be based on several experiments conducted on two French-speaking WAS: VITIPI and Sybille. At first, we discuss the important question of the evaluation of WAS. In a second part, we present the evaluated systems. Then, we describe the methodology of the evaluation we have followed during our experiments. Finally, we discuss our results and our perspectives.

II Problems for an evaluation of Writing Assistance Systems

Research on WAS suffers from serious methodological lacks in terms of evaluation. As far as we know, there is no evaluation campaign that could have involved several different systems. Generally speaking, previous evaluations of WAS were concerning a single system were conducted by its designers; various evaluation metrics could be used and the assessment is made on different data from one experiment to another. We will report some of those experiences in order to extract a list of features which have a strong influence on evaluation in our opinion.

The first feature concerns the test data. Very different kinds of corpus have been used by researchers to assess their system: the *BNC* corpus [4], news magazines (*"Times"* [5]), newspapers (*"Le Monde"* [6]), but also books French version of *"The Jungle Book" by* R. Kipling or even weather forecasts [8]. Considering that these corpora are corresponding to very different language registers, one can agree that the results of the corresponding experiments can hardly be balanced.

The second important feature of an evaluation session is the metric used to assess the performances of WAS. The method is to compute the ratio of keystrokes saved by the user. Various kinds of metrics could be found in the literature. Entropy was used by [9] to express the bit compression rate enabled by the system. Speed input can be estimated in words per minute [10], [11], or the actions of the user (and the system as well) can be monitored to compute a keystrokes saving rate [4]. Some of these metrics are not very easy to understand. For example, what does entropy means for the disabled subjects who are not skilled in mathematics? Moreover, it is not possible to compare results issued from these various metrics. Furthermore, the reliability of these

---

2 http://www.handicap-icom.asso.fr/adaptations/aides_techniques/clavicom.html
3 http://www.wivik.com/
4 http://www.assistiveware.com/keystrokes.php

metrics is perfectible. [13] has shown that the speed writing is not the most relevant metric for assessing the hardness of typing. Depending on the disabled ability, speed writing is widely changing. Thus, speed writing evaluations should depend more on the tested users than the intrinsic performances of the system.

At last, the usability of a WAS depends strongly on the main features of its user interface. There are various ways to interact with the system. The user can enter abbreviated forms that correspond to the desired words like in SMS language [14], [15] or can even stand for complete part of speech (ex: asap → as soon as possible) [4]. The system is in charge of completing these abbreviations. Another solution to save keystrokes is to predict the next word. There are various ways to propose these predictions. A list of words can be shown to the user who will select the desired one, if possible. This solution is adopted by Sibylle [16] or Keystrokes [1]. Another solution is to propose the most accurate word [17]. The user interface of a WAS concerns also its ability to correct (or not) automatically orthographic errors or miss typing mistakes (expecting that the use of WAS should reduce the frequency of errors). Likewise, it should be noted that the typed text could be displayed on a specific window integrated in the system, while other WAS allow a free typing on any kind of applications. Various interfaces of WAS are detailed in [8].

In conclusion, if some user interface considerations have been widely studied by previous WAS evaluation (Cf. [8]), this paper aims on the contrary at focusing on the two first criteria (test data and evaluation metric). This question will be detailed on section IV.

III Description of the involved systems

We have conducted a comparative evaluation which involved two WAS systems coming from two different laboratories: Sibylle and VITIPI.

a) Sibylle

Sibylle was first achieved in the VALORIA laboratory. New French and German versions of this system are now developed [18]. The user interface of Sibylle was defined in collaboration with KERPAPE rehabilitation centre[5]. This software is intended to cerebral palsy and physically handicapped children, but was already used by other patients. Children from the KERPAPE centre suffer for severe speech and motion impairments: they can only use a single switch with a big button which could be activated by every part of the body. The Sibylle interface that was used in our experiments was divided into three main areas: the virtual keyboard strictly speaking (letters and numbers), the list of words prediction words list, and an integrated text editor. New versions of Sibylle can now be used with any window applications and do not present this specific editor.

The main specificity of Sibylle is to display **dynamically** the list of predicted words but also the virtual keyboard. Indeed, the letter arrangement is adapted after every selection: the display is

[5] Kerpape Functional Rehabilitation Centre. http://www.kerpape.mutualite56.fr/page_uk/index_uk.htm

then refreshed in order to present first the most probable letters according to the previous context. This letter prediction is based on a statistical language model (SibyLetter). Likewise, word prediction is achieved by a structural language model (SibyWord) which combines a standard N-gram with a shallow parsing into *chunks* [6]. More details on Sibylle could be found in [16].

b) VITIPI

Generally speaking, VITIPI is a word completion system that was developed at IRIT Toulouse. When the user starts the keyboarding of the first letters of a word, the system displays either the ending of this word or a part of it as soon as there is no ambiguity. As long as the word remains incomplete, the system goes on writing as many letters as possible. Furthermore, VITIPI corrects typing errors or orthographic mistakes in real time. VITIPI does not use syntactical or semantic rules. It is based on what should call a kind of n-grams without likelihood which is implemented in a single transducer. Thanks to this transducer, parts of words are automatically displayed (for the purpose of completion) according to the lexicon. In the same way, this transducer handles word succession like standard language models. The VITIPI Knowledge Data Base (KDB) is made up with a vocabulary and a transducer. We built a minimal KDB that models general French language with the most usual words collected by N. Catach [19]. With a French lexicon made up 5,930 words, 26% of letters can be predicted. An English vocabulary is also available with a smaller size of 2,566 words. With this lexicon, 35% of letters can be predicted in average in an English text. More details on VITIPI could be found in [8], [17].

c) Systems features : comparison

A previous framework for WAS evaluation was proposed by [8]. It was based on a list of features.

Features	Sibylle	VITIPI
Input text in an abbreviated form?	No	No
Displays a list of words?	Yes	Yes (option)
Word completion without list presentation?	No	Yes
Can be used in every word processing or software?	No (tested version)	Yes

Table 1 : Features related to Computer Human Interfaces (CHI)

Features	Sibylle	VITIPI
Inserts automatically space after word?	Yes	Yes
Capitalize automatically letter in beginning sentences?	Yes	Yes
Accents automatically the letter?	No	No
Corrects automatically miss tapes?	No	Yes
Corrects automatically orthographic mistakes?	No	Yes (option)
Runs with isolated words	No	Yes (option)
Handles words succession?	Yes	Yes
Handles new entities?	Yes	Yes
Updates the KDB. Immediately? (dynamical adaptation)	Yes	Yes
Generates automatically a KDB. for a new language?	No (tested version)	Yes

Table 2: Features related to Natural Language Processing (NLP)

IV Methodology of evaluation

The achievement of an objective evaluation with various systems requires a clear definition of the main criteria which will be used for the comparison between (Sybille and VITIPI).. These criteria should i) be applied to each tested system, ii) provide homogenous and comparable results, iii) be easy to implement.

a) Objective metric of evaluation

The aim of Soukoreff metric's [12] is to assess the rate of typing errors. When the user types a text, he/she produces an *Input Stream*. Unfortunately, typing errors could appear, and if the user notices his/her mistake, he/she will press on the <back space> keypad to delete the unexpected character and then correct it. Thus, two additional keypads are added in the *Input Stream*. On the opposite, these two additional characters will not appear in the *Output Stream*. Soukoreff defines therefore the KeyStroke Per Character (KSPC) as an indicator of typing error rate:

$$KSPC = \frac{Input\_Stream}{Output\_Stream}$$ [Equation 1]

Beside this question of errors correction, one could consider that the aim of a WAS. is to reduce the length of the *Input Stream*. The more the *Input Stream* is shortest, the more the system is efficient and the faster is the writing. Thus, the KSPC metric presented in the [Equation 1] indicates also the percentage of letters really typed by the user. If we are interested in computing the percentage of letters automatically written by the system, we just have to compute the complement of [Equation 1]. This leads us to the following [Equation 2] that corresponds to what is classically called the Keystroke Saving Rate (KSR) metric:

$$KSR = 1 - \frac{Nb\_char\_typed + Nb\_function\_keys}{Length\_of\_text}$$ [Equation 2] where:

Nb_char_typed is the number of characters typed by the user to write his/her text, *Nb_function_key*s the number of function keys typed by the user to obtain the desired text with the WAS and *Length_of_text* the total number of characters of the intended text. It includes numbers, punctuation, and only one space between each word in order to minimize *Nb_char_typed*.

The function keys considered in [Equation 2] depend on assessed systems. For instance, the KSR includes the correction keys used by VITIPI to correct erroneous word completion, or switching keys from the letter keyboard to the area of selection of word prediction with Sibylle. If the desired word is in the k^{th} position in the prediction list, k function keystrokes are needed to select it. As a result, Nb_function_keys is incremented by the k value. Finally, the calculation of KSR takes into account all of the actions of the user for writing his/her text, but does not include text editing actions since they concern only text layout.

It could be observed that if the user writes a text without a WAS, then he/she has to type every letter of the text. Thus, KSR is equal to 0. On the opposite, *Nb_char_typed* should be significantly lower than *Length_of_text* thanks to the performance of word (or letter) prediction. If all the text could be predicted without typing any character, then KSR will be equal to 1. Then, the KSR can vary from 0 to 1. This metric neither depends on a user's speed, nor a particular system or a specific language. It is easier to interpret than a number of characters/bit, such an entropy metric which is unfortunately commonly used on software keyboards.

KSR appears to be a useful metric for assessing WAS performances. Nevertheless, the computation of a global KSR measures on "artificial" test data (newspapers corpus for instance) remains uninformative, since they are far from real uses. Indeed, your language style varies strongly while you are writing a scientific publication, a personal or an administrative mail or an e-mail.?

b) Ecological test data: Texts collection

We wanted to precisely to achieve an objective but « ecological » evaluation. It means that we have tried to be as close as possible to a disabled people who is really writing his/her text. In this paper, we will only consider one ecological situation: email communication. We therefore collected real emails that were written by a disabled people. This person was about 44 years old, he has a university post-graduate degrees. He is a cerebral palsy disabled without orthographic disorder. This texts collection is made up with 205 e-mails (19,640 words corresponding to 106,813 characters) and the corresponding vocabulary size is 2,712 words). This corpus has been divided into 26 sub-corpora with the same size. Since Sybille and VITIPI have dynamical learning abilities (see table 2), our aim was indeed to assess their ability to adapt themselves to the user language by observing the evolution of the KSR during the running of those 26 sub-corpora. Each sub-corpus is compound of 7.8 e-mails on average (about 4,100 characters).

Note that this corpus is completely anonymized and should be widely used for further experiences. Typing errors were corrected (multiple successions of spaces, spaces before comma...). Some orthographic mistakes were also corrected. On the opposite, we keep some writing specificities of the user such as capitalisation of words to outline them, use of smileys or some typical French e-mail abbreviations like, repetition of points, exclamation or question marks……!

Our aim is to measure the system ability to avoid errors. As Schadle [3] said, the number of typing errors is reduced by the use of WAS. For instance, VITIPI system [17] corrects 72% of typing errors.

c) Adapting the WAS to the user: incremental evaluation

In order to measure the learning abilities of the systems, we followed an incremental procedure of evaluation. At first, the systems are tested on the first of the 26 sub-corpora with their initial KDB. Since our systems have learning abilities, their KDB are then updated by precisely

considering this first sub-corpus. The updated KDB are used to test the second sub-corpus. At then end of the test, the KDB are updated once again. And so on until the 26th sub-corpus is reached.

V Results and discussion

We are going to discus results according to two parameters: the size of the learning corpora and the size of the previous context considered by the prediction system (N value) of the N-gram.

a) VITIPI

The performances of VITIPI are shown on Figure 1 and Figure 2. These results are obtained from three different experimental conditions, depending on the initial state of the system:

- With isolated word and without any initial lexicon (i.e. without initial KDB). Isolated word means that the previous context is not considered for word prediction (N=1 in the N-gram framework this condition is marked "*N=1 –Lexicon*" in the chart caption.
- Taking into account two previous words and without any initial lexicon. Two previous words means that N=3 is used in the N-gram framework. This condition is marked "*N=3 –Lexicon*" in the chart caption.
- Taking into account two previous words with an initial lexicon (Catach lexicon Cf. § IIIb). This condition is marked "*N=3 +Lexicon*" in the chart caption.

At the beginning of the evaluation process, for the first sub-corpus, one can observe that lexicon is a very useful help. More the number of messages is growing, more the system learns the user's own vocabulary. At the end, for the last sub-corpus, lexicon has no part for KSR growing. On the opposite, KSR is growing from 14% by taking into account two previous words. We just have to remember that VITIPI does not display list of words contrary to Sibylle.

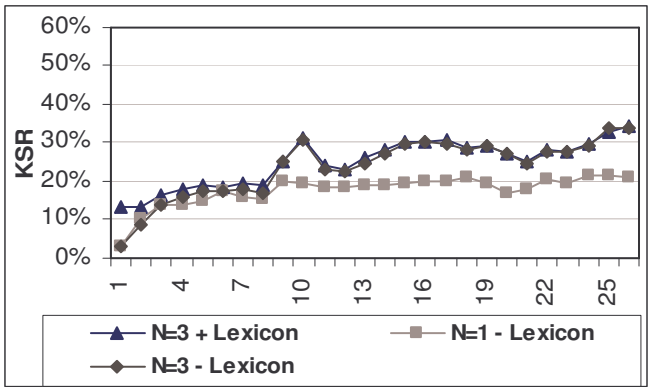

Figure 1: KSR VITIPI results

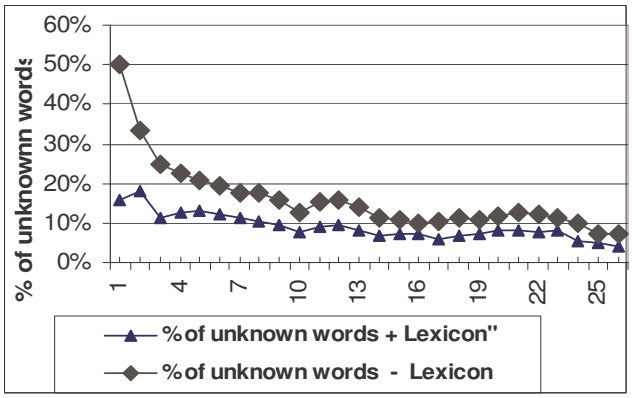

Figure 2: Percentage of unknown words

The effect of the learning abilities of VITIPI could be observed also on the Figure 2. On the abscissa are indicated the new words percentage for each sub-corpus. The number of new words decreases down to 3% for the experiments carried out with an initial lexicon (6% without lexicon).

b) Sibylle

Here, our aim was to compare the performances of Sibylle with three baselines that are well known in statistical language modelling the tri-gram, the bi-gram and the unigram models. Figure 3

shows the evolution of the KSR of Sibylle by comparison to the baselines. One should note that the baseline models were evaluated without initial learning, contrary to Sibylle which owns at least some grammatical knowledge for segmenting sentences into chunks. All the running tests (*baseline* and Sibylle) were carried out with a selection list of five predicted words.

The observed KSR are very closer for bi-gram and tri-gram models. This result was easily foreseeable since a tri-gram model needs a learning corpus with a significantly larger size than ours (at least hundreds of thousands words). Whatever the *baseline* is, one could see however the positive influence of learning. For example, the KSR of Sibylle starts with a initial value nearly 50% and then regularly grows to finish up to 60%. At the beginning, the bi-gram and tri-gram models (without initial knowledge) seem to catch up the KSR of Sibylle, but one can see that the difference finally stays around a constant 7%, what's shows the benefit of a structural language modelling in comparison with standard statistical baselines.

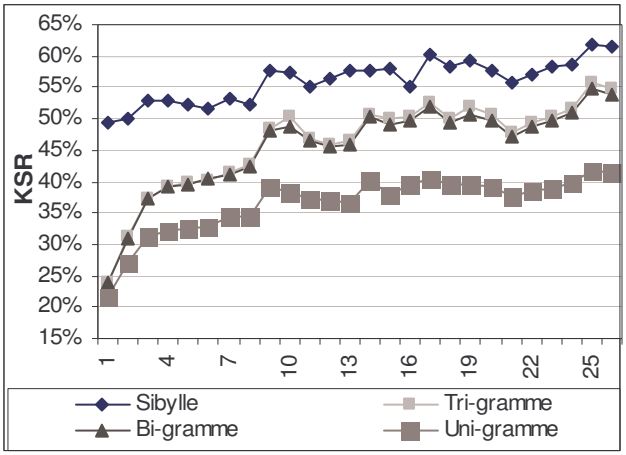

Figure 3: KSR Sibylle results

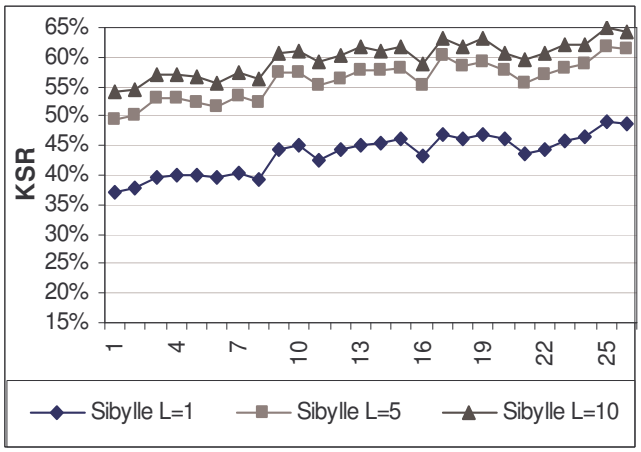

Figure 4: Influence of the words list size on the KSR

We have also measured the influence of the size of the list of prediction words. Sibylle has been evaluated with three different sizes: lists of 1, 5 and 10 words. The results (See Figure 4) show that the gap between a 1 word and a 5 words list is about 12% or 13%. However, there is lower a benefit of using a 5 words list rather than a 10 words one (about 4%). Those results should be analysed more precisely thanks to the implementation of a protocol focus on CHI evaluation. Even if moving from a 1 word to a 5 words list increases the cognitive load, the display of a 5 words list could be conceivable if you consider that the increase of the KSR is sufficiently noticeable. On the contrary, moving from a 5 word to a 10 word list leads to a really marginal increase that can not be balanced with the increase of the cognitive load.

VI Conclusion

This article pursues the work initiated by [8] of an evaluation framework for WAS. WAS are based on two distinct processes of equal importance: (1) word prediction, and eventually sentence prediction. This process is concerned with NLP propositions layout. (2) This process deals with the interface between the user and the system and concerns directly CHI.

Nevertheless, we think that there is a third level of importance which concerns the computer access (mouse, joystick, contactors, “fingers guide” or “grid keyboard”). Here stop the skills of a computer scientist, there begin those of an occupational therapist who has to deal on the one hand with computer tools (more or less adapted), and on the other hand with the physical and cognitive abilities of the user. The conception of a software keyboard is strongly linked to the computer access; two systems cannot be compared without taking this side into account.

This paper focused on the evaluation of the linguistic process (prediction). We have proposed an objective metric whose aim is to be i) generic (it can apply to any WAS), ii) independent of the interface, iii) easy to interpret. This metric was applied to two WAS following an incremental (learning) paradigm of evaluation. This general framework has required an evolution for the language model of Sibylle in order to allow a dynamical learning on the user’s data (until yet, this system had only been evaluated on a newspaper corpus, like ordinary statistical language models). By the same way, this evaluation campaign has incited the designer of VITIPI to add high level knowledge into their system. This work shows the benefits that could be obtained from comparative evaluation campaigns of WAS. Furthermore, we have built a test corpus which is really representative of a situation of communication that encounters disabled people.

We want to continue this work on the definition of a methodological framework for the evaluation of WAS. We invite all designers of such systems to meet us on this job. We are also thinking of collecting corpora in other situations of communication (writing communication, free texts, …). For the moment being, our corpus does not present strong linguistic mistakes. With regard to the NLP process, we wood like on the contrary to test in the future prediction systems with ungrammatical writings coming from dyslexic or aphasic people. Besides, from a CHI point of view, it is very difficult to evaluate the interfaces of WAS because of the strong differences of disabilities that can be found from one patient to another. Now, two people with different disabilities are not faced with the same problem; and even if they have the same disability they will deal with various difficulties. The observation has to be made “in vivo”, in the ordinary living place of the disabled people, with the essential devices that he/she needs. Only on those conditions, the adaptability of an interface for a particular disabled people could be validated or not. The E-Assist platform conceived at the IRIT Institute [20] seems to be a very efficient tool for this kind of experimentation. Furthermore, we have designed a test corpus which is really representative of a situation of communication that encounters disabled people.

VII Acknowledgments

Thanks a lot to Nadine Vigouroux and Frederic Vella for important feedback on this work and the careful reading

VIII Bibliography

[1] Berard C., Neimeijer D. Evaluating effort reduction through different word prediction systems. *IEEESMC, The Hague (Netherlands), Oct 2004, Ref ISBN : 0-7803-8566-7.*

[2] Vella F. Vigouroux N., Truillet Ph. SOKEYTO: a design and simulation environment of software keyboards. *8th proceedings of (AAATE 2005), Lille, France. Sept. 2005. pp. 723-727.*

[3] Schadle, I., Antoine J.-Y., Le Pévédic B., Poirier F. SibyLettre : Prédiction de lettre pour la communication assistée. *Revue RIHM, 3(2), 2003, pp.115-133.*

[4] Väyrynen P. Perspectives on the Utility of Linguistic Knowledge in English Word Prediction. *Ph.D of the OULU University. November 19th 2005.*

[5] Lesher, G. W., Moulton, B. J., Higginbotham, D. J. Effects of n-gram order and training text size on word prediction. *Proc. of the RESNA '99 Annual Conference, Arlington 1999, pp.52-54.*

[6] Schadle I., Antoine J.-Y., Le Pévédic B., Poirier F. SibyMot - Modélisation stochastique du langage intégrant la notion de chunks. *Proceedings of TALN 2004, Fès.*

[7] Menier, G., Poirier, F. Système adaptatif de prédiction de texte. *Atelier Thématique TALN 2001, Tours, 2-5 Juillet 2001, pp. 213-222.*

[8] Boissière Ph., Dours D. An evaluation framework for writing assistance systems: Application to VITIPI. *Modelling, Mesurement & Control, Série C, (bioengineering). (AMSE 2002) pp.119-128.*

[9] Ward D. J. Adaptative Computer Interfaces, in *PhD. Thesis Cambridge University, Nov. 2001.*

[10] Mackenzie I., Zhang S. The Design and Evaluation of a High-Performance Soft Keyboard In *Proc. of CHI'99, Computer Human Interfaces, Pitsburg (USA) 15-20 May 1999, pp.25-31.*

[11] Smith B.A., Zhaï S. Optimized Virtual Keyboards with and without alphabetical ordering. In *Proc. of Interact'2001, IFIP TCI3, Tokyo, Japan, pp.92-99.*

[12] Soukoreff, R. W., & MacKenzie, I. S. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. *Proceedings of the ACM Conference on Human Factors in Computing Systems – CHI 2003, New York: ACM (2003), pp.113-120.*

[13]Vigouroux N., Vella F., Raynal M., Boissière Ph. Solutions et défis pour une meilleure accessibilité et utilisabilité des communicateurs - Optimisation de la saisie de texte. In *Handicap et Environnement. Entretiens de l'Institut de Garches, Nanterre, 24-25 novembre 2005. pp.209-222.*

[14] Willis T., Pain H., Trewin S., Clark S. Informing Flexible Abbreviation Expression for User with Motor Disabilities In *Proceedings of the 8th ICCHP'2002. LNCS 2398, pp.251-266.*

[15] Shieber S.M., Baker E., Abbreviated Text Input In IUI'03, *Proceedings of International Conference on Intelligent User Interface, Miami, Florida (USA), January 12-15, 2003.*

[16] Schadle, I. Sibylle : Système linguistique d'aide à la communication pour les personnes handicapées. *Thèse de doctorat, Université de Bretagne Sud, Décembre 2003.*

[17] Boissière, P., Dours, D. VITIPI: Versatile Interpretation of Text Input by Persons with Impairments. *5th Proceedings of ICCHP, Linz July 1996, pp.165-172.*

[18] Wandmacher T. Antoine J.-Y., Schadle I., Krueger-Thielmann K. (2006) Sibylle AAC system: exploiting syntax and semantics for word prediction. *Proc. ISAAC'2006. Duesseldorf, Germany.*

[19] Catach N., Les listes orthographiques de base du Français (LOB). *Nathan 1984.*

[20] Raynal, R. Maubert, S. Vigouroux, N. Vella, F. Magnien, L.. E-Assiste: A platform Allowing Evaluation of Text Input System. In *3rd Int. Conf. on Universal Access in Human-Computer Interaction (UAHCI 2005), Lawrence Erlabaum Associates (LEA), ISBN 0-8058-5807-5, Las Vegas, USA, 22 - 27 July 2005.*